NATHALIE CRUZ SENA

ESTRATÉGIAS METODOLÓGICAS PARA MAPEAMENTO DE SOLOS

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: Márcio Rocha Francelino

Carlos Ernesto G. R. Schaefer

VIÇOSA - MINAS GERAIS

2020

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

T
S474e
2020

Sena, Nathalie Cruz, 1992-
Estratégias metodológicas para mapeamento de solos / Nathalie Cruz Sena. – Viçosa, MG, 2020.
96f.: il. (algumas color.).

Orientador: Elpídio Inácio Fernandes Filho.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Solos, 2020.
Inclui bibliografia.

1. Mapeamento digital de solos. 2. Amostragem. 3. Balanceamento. 4. Aprendizado de máquina. 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. 631.47

Bibliotecário(a) responsável: Renata de Fátima Alves CRB-6/2875

NATHALIE CRUZ SENA

ESTRATÉGIAS METODOLÓGICAS PARA MAPEAMENTO DE SOLOS

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: 19 de fevereiro de 2020.

Assentimento:

Nathalie Cruz Sena
Autor

Elpídio Inácio Fernandes Filho
Orientador

“Entrega o teu caminho ao Senhor, confia Nele, e Ele o fará”.
(Salmos 37:5)

A pequena Marina,
por ser luz na minha vida.
DEDICO.

AGRADECIMENTOS

A Deus, sobre todas as coisas, que és minha força e meu guia.

A minha família, por sonharem comigo este sonho.

Ao Departamento de Solos e ao Programa de Pós Graduação em Solos e Nutrição de Plantas (UFV) pela oportunidade de aprendizagem e crescimento.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 pelo apoio para realização do presente trabalho.

A CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico pela bolsa de estudo concedida e à FAPEMIG por disponibilizar outros apoios financeiros.

Ao professor Elpídio Inácio Fernandes pela oportunidade da pesquisa, pelos ensinamentos e orientação.

Ao professor Márcio Francelino, por ser meu norte, por toda atenção e apoio.

Ao professor João Carlos Ker, pelo cuidado, abraços e sorrisos compartilhados.

Ao professor Carlos Ernesto por todos os ensinamentos e por ser uma inspiração para todos nós.

Aos colegas do Labgeo – Laboratório de Geoprocessamento (UFV), em especial ao Gustavo Veloso por toda paciência, prestatividade e orientações fundamentais para o desenvolvimento da pesquisa, a quem serei eternamente grata.

A equipe responsável pelos trabalhos de campo: Prof Márcio, Alisson, Luiz, David Lukas, Eduardo Senra, Viviane, Felipe Santana e Raphael Wakin... Sem vocês nada disso seria possível.

A todos os funcionários e servidores do Departamento de Solos, em especial, a Zélia, a Claudinha e a Carol, pelo amor e carinho.

Aos meus amigos do DPS - UFV, em especial ao Francis, Eliana, Rayanne, Marcel e Alisson por serem abrigo, por fazerem a caminhada ser mais leve e por compartilharem momentos incríveis comigo.

As minhas amigas que mesmo longe, estiveram presentes, Leiliane Bozzi, Thaís Pessoa e Gabriela Botelho.

Por cada um que cruzou meu caminho, que ficou, que seguiu...

Gratidão!

RESUMO

SENA, Nathalie Cruz, D.Sc., Universidade Federal de Viçosa, fevereiro de 2020. **Estratégias metodológicas para mapeamento de solos.** Orientador: Elpídio Inácio Fernandes Filho. Coorientadores: Márcio Rocha Francelino e Carlos Ernesto G. R. Schaefer.

Atualmente, há uma crescente necessidade por informações pedológicas em escalas mais detalhadas, precisas, de fácil acesso e com menor custo para direcionar o planejamento e tomada de decisões no que diz respeito às práticas de uso, manejo e conservação dos solos. Assim, o mapeamento digital do solo (MDS) pode ser utilizado para auxiliar na geração de informações pedológicas, integrando técnicas convencionais de levantamento e mapeamento de solos com conhecimentos de diversas áreas, como técnicas de Sensoriamento Remoto, Geoestatística e Aprendizado de Máquinas. É nesta perspectiva que o presente trabalho visa desenvolver estratégias para o mapeamento de solos mais eficientes, a fim de minimizar as limitações atuais dos métodos convencionais. O artigo 1 *Análise das abordagens metodológicas utilizadas nos trabalhos de mapeamento digital de solos* abrange uma revisão de literatura acerca das abordagens metodológicas empregadas nos trabalhos de MDS no período de 2008 a 2018, a fim de avaliar as lacunas, perspectivas e desafios futuros. O artigo 2 *Análise dos atributos do terreno em diferentes resoluções espaciais para aplicação em mapeamento digital de solos no sudeste do Brasil* analisa o efeito das diferentes resoluções espaciais dos MDEs e atributos derivados do terreno e suas implicações para aplicação em modelos preditivos do solo. O artigo 3 *Avaliação do método hipercubo latino condicionado (cLHS) para seleção de locais de amostragem* avalia a utilização do método cLHS para seleção de locais de amostragem a serem utilizados no mapeamento digital de solos e analisa o desempenho operacional com base nas suas potencialidades e restrições. O artigo 4 "*Técnicas para mapeamento digital de solos utilizando conjunto de dados desbalanceado*" avalia como o desbalanceamento de classes afeta o desempenho dos modelos preditos, a fim de auxiliar no desenvolvimento de técnicas para os desafios futuros de mapeamento de solo no País.

Palavras-chave: Mapeamento Digital de Solos. Amostragem. Balanceamento. Aprendizado de Máquinas.

ABSTRACT

SENA, Nathalie Cruz, D.Sc., Universidade Federal de Viçosa, February, 2020. **Methodological strategies for soil mapping.** Adviser: Elpídio Inácio Fernandes Filho. Co-advisers: Márcio Rocha Francelino and Carlos Ernesto G. R. Schaefer.

Currently, there is an increasing need for pedological information on more detailed, accurate, easily accessible and less costly scales to guide planning and decision making with regard to soil use, management and conservation practices. Thus, digital soil mapping (DSM) can be used to assist in the generation of pedological information, integrating conventional survey and soil mapping techniques with knowledge from different areas, such as Remote Sensing, Geostatistics and Machine Learning techniques. It is in this perspective that the present work aims to develop strategies for more efficient soil mapping in order to minimize the current limitations of conventional methods. The article 1 *Analysis of the methodological approaches used in the digital soil mapping* works covers a literature review about the methodological approaches used in the DSM studies in the period from 2008 to 2018, in order to assess the gaps, perspectives and future challenges. Article 2 *Analysis of the attributes of the terrain in different spatial resolutions for application in digital soil mapping in southeastern Brazil* analyzes the effect of different spatial measures of the DEMs and the attributes of terrain and their implications for application in predictive soil models. Article 3 *Evaluation of the conditioned Latin hypercube method (cLHS) for the selection of sampling sites* assesses the use of the cLHS method for selecting sampling sites to be used in digital soil mapping and analyzes operational performance based on its potential and restrictions. Article 4 *"Techniques for digital soil mapping using an imbalanced dataset"* assesses how class imbalance affects the performance of predicted models, in order to assist in the development of techniques for future soil mapping challenges in the country.

Keywords: Digital Soil Mapping. Sampling. Balancing. Machine Learning.

SUMÁRIO

INTRODUÇÃO GERAL

A demanda por informações pedológicas cada vez mais precisas, atualizadas, de fácil acesso e com baixo custo é alta por muitos setores da sociedade brasileira e são de suma importância para direcionar o planejamento e tomada de decisões no que diz respeito ao uso, manejo, conservação e aproveitamento das potencialidades dos solos.

Atualmente, o Brasil possui a maior parte do seu território coberto por levantamentos exploratórios de solos na escala 1.000.000, realizados na década de 70 pelo Projeto RadamBrasil (BRASIL, 1978). Estes auxiliaram no reconhecimento e mapeamento da distribuição das classes de solos de forma generalizada. Contudo, estes mapeamentos já não atendem as demandas atuais dos diversos setores da sociedade.

Em 2015, foi estabelecida a criação do Programa Nacional de Solos do Brasil (PRONASOLOS), que prevê, dentre vários objetivos, a retomada da realização dos levantamentos pedológicos sistemáticos em caráter multiescalar no País. Entretanto, as técnicas e protocolos que serão aplicados ainda não foram definidos, mas acredita-se que a associação de métodos convencionais e digitais será necessária para a execução das atividades propostas. Nesta perspectiva, ressalta-se a busca pelo desenvolvimento e aplicação de novos métodos e tecnologias para a execução desse desafio.

É neste sentido que o mapeamento digital de solos (MDS) surge como alternativa para auxiliar na geração de informações pedológicas mais detalhadas e ampliar as aplicações destes tipos de dados. Para isso, são utilizados métodos quantitativos que proporcionam uma nova perspectiva de mapeamento e uma melhor representação espacial dos solos. Vale destacar que esse método não descarta a necessidade do pedólogo, mas associa a ele o uso de ferramentas computacionais e estatísticas, que permitem a espacialização de classes e propriedades dos solos relacionando-os com variáveis ambientais.

Assim, faz-se necessário conhecer o que tem sido empregado de métodos e técnicas para o levantamento e mapeamento de solos nos últimos anos, analisar e propor novas metodologias que possibilitem a geração de informações numa perspectiva de melhor custo-benefício, a fim de produzir mapas de solos com acurácia conhecida e reproduzível para diferentes aplicações em um ambiente SIG.

Para tanto, o objetivo do presente trabalho consiste em aplicar em uma área piloto no estado de Minas Gerais, técnicas de mapeamento convencional e digital de solos, verificar as potencialidades e limitações dos métodos; e subsidiar futuros trabalhos de mapeamento de solos no País e planejamentos ambientais ou agrícolas.

Artigo 1

ANÁLISE DAS ABORDAGENS METODOLÓGICAS UTILIZADAS NOS TRABALHOS DE MAPEAMENTO DIGITAL DE SOLOS

1. Introdução

A Ciência do Solo alcançou diversos avanços a partir da década de 1980 em decorrência da crescente necessidade de obter conhecimentos sobre solos, seja para uso ambiental ou agrícola, seu manejo e conservação, em escalas mais detalhadas. O desenvolvimento de novas técnicas para a modelagem de solos associado à tecnologia da informação, dada pela maior capacidade de processamento dos computadores, incorporação de produtos do sensoriamento remoto e maior acessibilidade aos dados de alta resolução espacial, facilitaram a geração de informações espaciais para auxiliar e direcionar o planejamento do uso racional e sustentável dos solos.

No mapeamento digital de solos (MDS), o solo passa a ser estudado e descrito como uma entidade dinâmica em um contexto de paisagem interconectada, cujas variações espaciais dos tipos de solos e/ou propriedades podem ser inferidas a partir de modelos numéricos, observações em campo e do conhecimento dos efeitos das variáveis ambientais sobre a gênese e morfologia dos solos (McBratney et al., 2003; Lagacherie e McBratney, 2007). O MDS possui capacidade de predizer e espacializar informações pedológicas e atualizar os mapeamentos de solos convencionais, com custos menos onerosos e menor demanda de tempo. Wang et al. (2018) apontam que é provável que o MDS desempenhe um papel cada vez mais importante no monitoramento futuro de mudanças nas propriedades e características do solo.

Contudo, ainda que diversas pesquisas já tenham sido realizadas, não existem procedimentos, padrões ou protocolos estabelecidos no processo de mapeamento digital de solos. Diversas abordagens têm sido aplicadas ao longo dos anos, como as regressões logísticas (Kempen et al., 2009; Jeune et al., 2018), suporte vetor de máquinas – SVM (Ballabio, 2009; Guevara et al., 2018), random forest (Grimm et al., 2008; Vagen et al., 2016), lógica fuzzy (Menezes et al., 2018), árvores de decisão (Giasson et al., 2011; Adhikari et al., 2014), redes neurais artificiais (Chagas et al., 2013; Arruda et al., 2016), dentre outras.

Embora outros trabalhos de revisão de literatura tenham sido realizados (Bui, 2006; Minasny e McBratney, 2016; Arrouays et al., 2017), estes comumente abordam o conceito e histórico do MDS, fatores que levaram o seu desenvolvimento e avanços alcançados. Não existem estudos recentes que enfatizem as abordagens metodológicas utilizadas para o mapeamento digital de propriedades e classes de solos. Nesse sentido, o objetivo do trabalho foi realizar uma revisão de literatura das abordagens metodológicas empregadas no MDS e avaliar as lacunas, perspectivas e desafios futuros.

2. Metodologia

Foi realizado o levantamento das informações acerca das abordagens metodológicas empregadas no MDS nos últimos anos, por meio das plataformas globais digitais ScienceDirect e Scopus, utilizando como palavras-chave "*digital soil mapping*". Os dados foram filtrados de acordo com os anos de publicação (2008 a 2018), tipo de documento (artigo científico) e tipo de acesso (acesso livre).

Durante a análise individual de cada artigo, foram considerados para o presente estudo somente trabalhos que aplicaram o MDS *stricto sensu*, proposto por ten Caten et al. (2012). Neste tipo de trabalho são utilizados algoritmos e modelos numéricos no qual se empregam covariáveis ambientais para predizer propriedades e classes de solos. No MDS *lato sensu*, não ocorre uma classificação numérica de solos, mas apenas uma delimitação de classes de solos apoiada por sistemas informatizados (ten Caten et al., 2012), sendo assim, esta abordagem não foi considerada neste estudo.

Não houve distinção entre periódicos científicos nacionais e internacionais. Portanto, foram revisadas 104 publicações nas quais foram avaliados nove critérios, os quais estão descritos abaixo (Adaptado de Grunwald, 2009):

a. Região geográfica;
b. Predição de propriedades e classes de solo;
c. Plano de amostragem (densidade e distribuição amostral);
d. Covariáveis utilizadas no estudo com base nos fatores ambientais do modelo *scorpan*: solos (s), clima (c), organismos (o), relevo (r); material de origem (p); tempo (a); e coordenadas geográficas (n);

e. Incorporação de dados legados de solos;
f. Modelo/Algoritmo de predição do solo;
g. Avaliação do desempenho dos modelos (validação e métricas)
h. Custo
i. Tempo

3. Resultados e Discussão

3.1 Região geográfica

Os resultados apontam que a maior parte dos trabalhos de MDS foi realizada em países da América do Sul (n: 37) e da Europa (n: 30). Por outro lado, as pesquisas ainda são limitadas em países da Ásia (n: 8) e Oceania (n: 6) (Figura 1). Foram constatados também outros três trabalhos que englobaram mais de um local/país específico, sendo eles: (1) escala global, (2) em alguns países da América Latina e (3) realizado em conjunto com países da Europa e América do Sul.

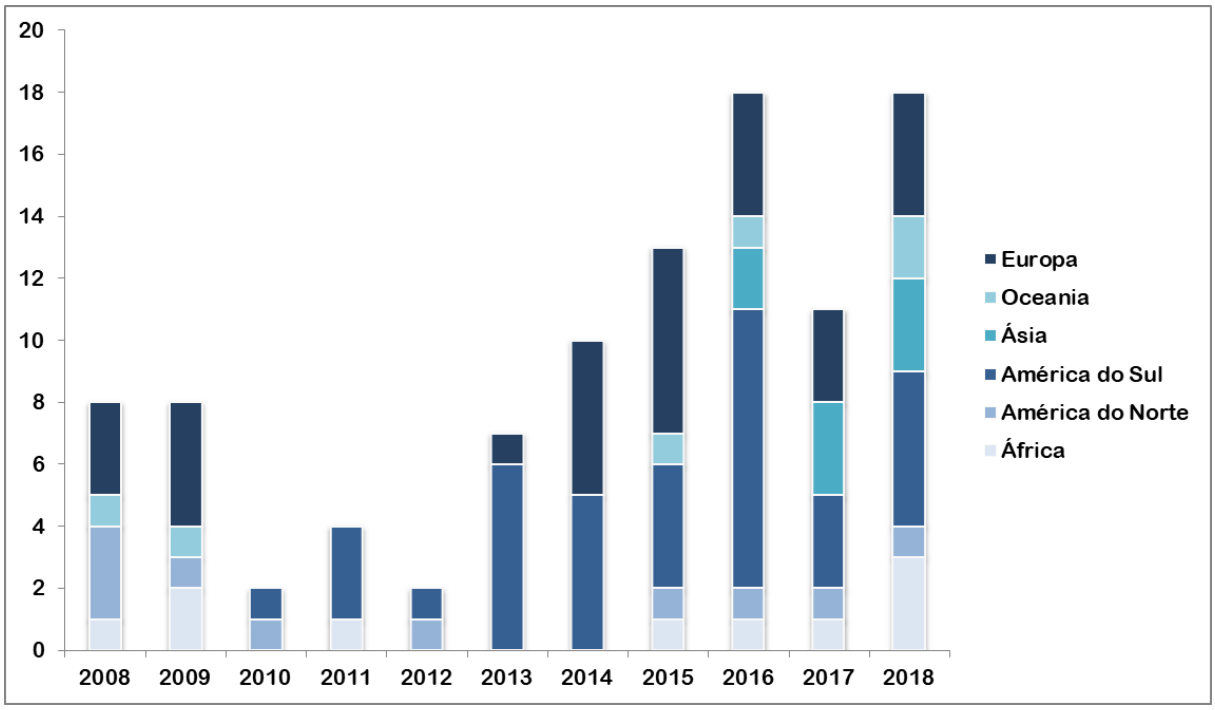


Figura 1. Número de trabalhos de MDS por região geográfica entre os anos de 2008 a 2018.

No entanto, há uma tendência de que novas pesquisas se desenvolvam nessas regiões que apresentam estudos limitados. Iniciativas internacionais como o *SoilGrids*, que consiste em um sistema global de mapeamento digital de solos, o qual utiliza métodos de aprendizado de

máquina para mapear a distribuição espacial das propriedades do solo em todo o mundo, possui como um dos objetivos, apoiar pesquisas em países da África e grande parte da Ásia e da América Latina, os quais geralmente possuem infraestruturas limitadas para produzir informações sobre o solo em resolução fina (Hengl et al., 2014).

Grunwald (2009) realizou uma revisão abrangente sobre mapeamento e modelagem digital de solos entre anos de 2007 e 2008, na qual constatou que a maior parte dos estudos foi desenvolvida em países da Europa e nos Estados Unidos. Alguns autores consideram que o desenvolvimento de trabalhos de MDS nestes países, por exemplo, deve-se a disponibilidade de levantamentos de solos em escalas mais detalhadas e a aplicação de recursos continuados na caracterização e cartografia dos solos com crescente grau de detalhamento (Mora-Vallejo et al., 2008; Embrapa, 2016; Curi et al., 2017). Na Europa, por exemplo, existe o Banco de Dados Europeu do Solo (ESDB) que consiste em uma fonte importante de informações de solos, com dados analíticos de perfis dos solos e propriedades hidráulicas dos solos da Europa, cujos arquivos possuem tamanhos de célula de 1 km x 1 km e 10 km x 10 km; além de mapas, documentos, serviços e aplicativos que favorecem o desenvolvimento das pesquisas em MDS, como os estudos de Ballabio et al. (2016) e Schmidt et al. (2018). Contudo, desde o estudo realizado por Grunwald (2009) até os dias atuais, observa-se que o desenvolvimento das pesquisas em MDS tem ganhado destaque nos países da América do Sul.

Contribuições do Brasil para o MDS

Quase a totalidade de trabalhos avaliados na América do Sul foi realizada no Brasil (n: 36), principalmente nos estados do Rio de Janeiro, Rio Grande do Sul, São Paulo e Minas Gerais. Historicamente, nestes estados foram concentrados os principais centros de pesquisas e onde foram realizados os primeiros levantamentos de solos do País. Carvalho et al. (2013) verificaram que entre 1949 e 1960, do total de 14 levantamentos de solos realizados no Brasil, 11 foram na região Sudeste. Lima (2013) aponta que o MDS no Brasil tem se afirmado com um número cada vez maior de artigos publicados em revistas científicas especializadas, bem como com a participação de pesquisadores brasileiros em publicações internacionais e ressalta a importância de expandir as técnicas de MDS para outras regiões do país, visto que os vazios cartográficos em solos concentram-se principalmente nas regiões Norte e Nordeste do Brasil (Mendonça Santos e Santos, 2007).

ten Caten et al. (2012) relatam que o desenvolvimento do MDS é recente no Brasil, datando do início dos anos 2000. Em 2006, a Embrapa Solos organizou no Rio de Janeiro, com o apoio da International Union of Soil Sciences e da Sociedade Brasileira de Ciência do Solo, o 2nd Global Workshop on Digital Soil Mapping, que reuniu 75 pesquisadores de 17 países, para apresentar e discutir os avanços em mapeamento digital de solos. Uma seleção de artigos foi publicada como livro, intitulado Digital Soil Mapping With Limited Data (Hartemink et al., 2008).

Nos últimos anos, iniciativas como a criação do comitê de Pedometria na Divisão Solo no Espaço e no Tempo da Sociedade Brasileira de Ciência do Solo em 2010; a Rede Brasileira de Pesquisa em Mapeamento Digital de Solos (RedeMDS) em 2011 e o Programa Nacional de Solos do Brasil (PronaSolos) em 2015; auxiliaram no desenvolvimento e aplicação de novas tecnologias para o mapeamento digital de solos no Brasil e, consequentemente, contribuíram para o avanço das pesquisas no País.

Dalmolin et al. (2017) apontam que o Brasil, principalmente nos últimos anos, tem acompanhado a proporção de publicação dos pesquisadores internacionais, com merecido destaque nessa área de estudo. Os autores complementam que trabalhos publicados por brasileiros são ainda pouco citados quando comparados aos de pesquisadores estrangeiros, devido às limitações advindas do uso do idioma português para a redação dos manuscritos. Além disso, os trabalhos são desenvolvidos essencialmente em escalas locais, não despertando o interesse para a citação em estudos de escalas nacionais ou continentais de outras regiões do globo.

Embora alguns avanços já tenham sido alcançados, existem ainda muitas lacunas para a consolidação do MDS no Brasil, como a grande extensão territorial do País e a abrangência de estudos restrita nas regiões sul-sudeste. Destaca-se ainda a carência de informações cartográficas em escala adequada, falta de profissionais qualificados para o emprego da tecnologia da informação, crises financeiras e a resistência dos pedólogos convencionais em adotar os novos métodos pautados em sistemas automatizados.

3.2 Predição de propriedades e classes de solo

A principal abordagem do MDS consiste na predição das propriedades e classes de solos por meio de modelos matemáticos e seu mapeamento digital de forma contínua e espacial (McBratney et al., 2003). No presente estudo foi constatado que do total dos trabalhos avaliados, 71 % são focados na predição de propriedades do solo, tais como: teor de carbono orgânico (COS), textura e pH do solo e; 27 % são estudos de predição de classes de solos. Apenas 2% dos trabalhos realizaram a predição conjunta de classes e propriedades de solo. Samuel-Rosa (2012) ratifica que está ocorrendo uma mudança de paradigma no mapeamento de solo, cuja maior parte dos trabalhos desenvolvidos nos últimos anos foca, principalmente, no mapeamento digital de propriedades do solo (McBratney et al., 2003; Bishop e Minasny, 2006; Grunwald, 2009).

Os resultados apontam também que 39 % dos trabalhos realizaram predição de uma propriedade específica do solo, enquanto os demais realizaram a predição de duas ou mais propriedades. Desta forma, a maioria dos estudos de predição de propriedades do solo concentrou-se na predição do teor de carbono orgânico do solo (17 %). A crescente preocupação com as mudanças climáticas, o aquecimento global e seus efeitos podem justificar o maior número de estudos associado à quantificação e predição de COS em todo mundo.

O solo configura-se como maior reservatório de carbono orgânico nos ecossistemas terrestres e armazena de duas a três vezes mais carbono orgânico do que a atmosfera ou a vegetação terrestre (Lal, 2004; Schmidt et al., 2011). Bellamy et al. (2005) apontam que pequenas mudanças na quantidade de COS poderiam afetar muito as concentrações atmosféricas de CO_2 devido à sua sensibilidade às mudanças climáticas e às atividades humanas, como as práticas de manejo e mudanças no uso da terra. O COS também tem sua importância reconhecida por estar intimamente relacionado com a fertilidade e produtividade do solo, processos biológicos, estrutura e propriedades hidráulicas do solo (Tiessen et al., 1994).

Para tanto, o desenvolvimento dos estudos de predição do COS tem sido impulsionado pela busca de métodos confiáveis, robustos e de custo eficiente para monitoramento e verificação do sequestro de carbono no solo e biomassa (Bou et al., 2010), bem como para servir como indicador da fertilidade do solo. Wang et al. (2018) justificam que os avanços nos estudos do

COS resultam principalmente do desenvolvimento de técnicas de aprendizado de máquina e da disponibilidade de covariáveis de alta qualidade. No trabalho de Yigini e Panagos (2016), por exemplo, foi realizada a previsão dos atuais estoques de carbono orgânico do solo e a projeção para o futuro próximo (ano 2050), utilizando a técnica de regressão-krigagem e um conjunto de preditores ambientais.

3.3 Plano de amostragem de solos

A amostragem é uma questão essencial nos trabalhos de levantamento de solos, a qual deve ser representativa da variabilidade das amostras no espaço geográfico. Nesse sentido, tem sido cada vez mais abordado nos trabalhos de MDS sobre a necessidade do desenvolvimento de métodos eficientes para identificação de locais de amostragem, com melhorias na exatidão dos mapas gerados e menor custo (Ließ, 2015; Pahlavan-Rad et al., 2016; An et al., 2018).

Diferentes métodos de amostragem foram aplicados nos trabalhos avaliados. Foi observado que a maioria dos trabalhos (22 %) utilizou combinações de duas ou mais técnicas de amostragem (Teske et al., 2015; Vasques et al., 2016; Guo et al., 2018). Os métodos menos representativos nos estudos foi a amostragem aleatória ponderada e a amostragem em cluster. Além disso, 21 % dos trabalhos não informaram como foi realizado o desenho amostral. (Figura 2).

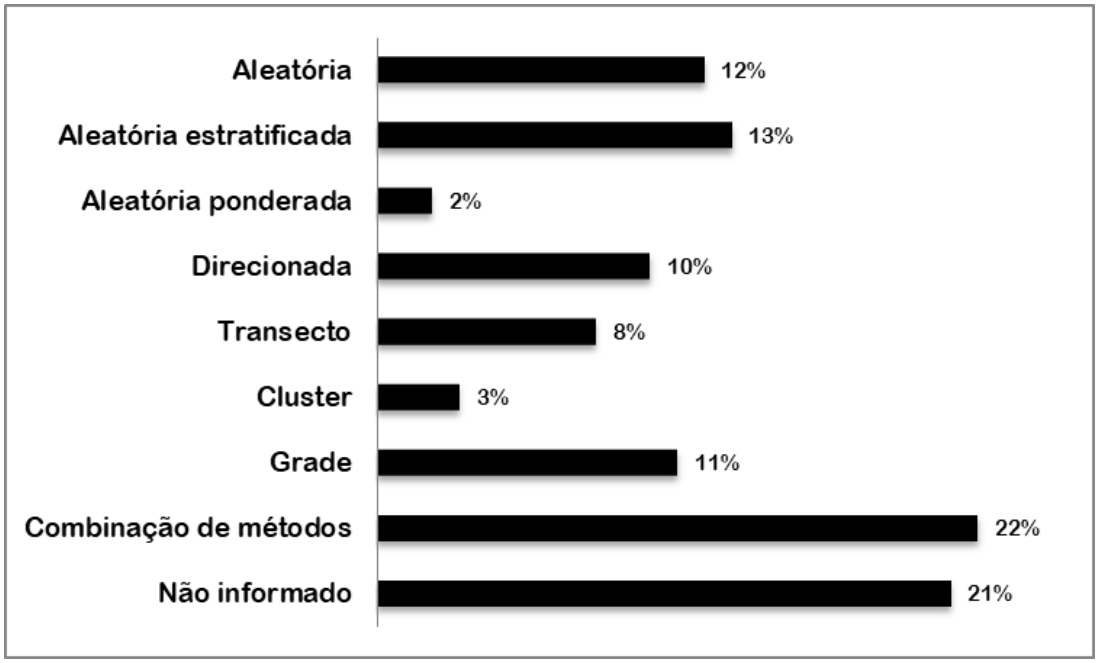


Figura 2. Métodos de amostragem utilizados nos trabalhos de MDS.

A amostragem aleatória foi empregada, por exemplo, nos trabalhos de Giasson et al., 2011; Bagatini et al., 2015 e Silva et al., 2017. Segundo Hengl et al. (2003), este método elimina a subjetividade e permite a reprodutibilidade simples. De acordo com essa estratégia, cada amostra tem a mesma probabilidade de ser selecionada e as amostras selecionadas devem representar bem a população considerada. Entretanto, Giasson et al. (2011) retratam que classes e unidades de mapeamento de solos pouco extensas e pouco representativas de uma área, podem não ser preditas pelos modelos quando utilizada a amostragem aleatória.

A amostragem aleatória estratificada é um subtipo da amostragem aleatória e foi constatada em 13 % dos trabalhos de MDS; como nos estudos de Lamsal et al. (2009) e Schmidt et al. (2018). Esta estratégia de amostragem é recomendada para amostrar as unidades de mapeamento pouco representativas (Hengl et al., 2003). Nesta abordagem, a área é subdividida em um número de estratos ou subáreas e amostras aleatórias são tomadas de cada um desses estratos, dando probabilidades iguais ou diferentes de seleção aos estratos (Ließ, 2015).

A amostragem em grade foi utilizada unicamente em 11 % dos trabalhos e em 9 % dos trabalhos esta amostragem aparece associada a outras técnicas de amostragem, tais como a amostragem em transecto, como observado nos estudos de Adhikari et al. (2014) e Menezes et al. (2018). Na grade amostral são estipuladas distâncias entre os pontos amostrados, cada ponto é georreferenciado e representa uma área determinada pela distância entre os pontos vizinhos. Guo et al. (2018) observaram que o plano de amostragem em grade teve um desempenho melhor que o método de amostragem aleatória, por exemplo.

Outra técnica que vale destacar é a amostragem direcionada ou intencional, a qual considera que a variação espacial do solo de uma área pode ser capturada pelos valores da propriedade em alguns locais típicos (chave), baseado no conceito clássico da relação solo-ambiente, cujas propriedades únicas do solo podem ser associadas a combinações únicas (ou configurações) de fatores ambientais (Zhu et al., 2010; Qin et al., 2012). Contudo, a amostragem direcionada não é aleatória e pode fornecer estimativas estatísticas não representativas da totalidade da área, consistindo mais em um método subjetivo por parte dos pedólogos, para sustentar os modelos mentais de distribuição dos solos na paisagem.

Quanto à densidade de amostragem, não existe uma padronização em relação ao número de amostras ou observações a serem utilizados nos modelos preditivos, o mesmo fato foi observado por ten Caten et al. (2012) em seu trabalho de revisão de metodologias do MDS no Brasil. Trabalhos relataram número total de amostras de solo (Gomez et al., 2008a; Baltensweiler et al., 2017); perfis de solos (Malone et al., 2009; Hitziger e Ließ, 2014; Penížek et al., 2016); número de locais de observação (Grinand et al., 2008; Chagas et al., 2010); número de camadas/horizontes de solo (Carvalho Junior et al., 2014); frequência/km² (Grinand et al., 2008; Pásztor et al., 2015).

3.4 Covariáveis preditoras

Um dos primeiros requisitos no MDS consiste na seleção e definição do conjunto de covariáveis ambientais para a predição de classes e propriedades do solo. A abordagem do MDS segue, na maioria das vezes, a função de previsão espacial denominada *scorpan* (McBratney et al., 2003), a qual é uma generalização da equação de Jenny (1941), que formaliza as relações entre as classes, propriedades do solo e fatores ambientais:

$$S = f\,(s,c,o,r,p,a,n) \qquad \text{(Eq. 1)}$$

onde s: solo, previamente medido ou outras propriedades do solo; c: clima, propriedades climáticas do ambiente; o: organismos, vegetação ou fauna ou atividade humana; r: topografia, atributos morfométricos; p: material parental, litologia; a: idade, o fator tempo; e n: espaço, posição espacial.

As informações de solos previamente mensuradas ou suas propriedades foram utilizadas em 92 % do total de trabalhos avaliados. As informações dos atributos morfométricos – atributos primários e secundários (r) em 76 % dos trabalhos. Em seguida, o fator organismos (o) foi utilizado em 63 %; o material parental ou litologia (p) em 28 % e as propriedades climáticas (c) em 23 %. A posição espacial (n) – 12 % e o fator tempo (a) – 3 % foram as covariáveis menos utilizadas. Nenhum trabalho adotou a estrutura completa do modelo *scorpan*, no qual utiliza-se todos os fatores ambientais para predizer classes e/ou propriedades do solo. Alguns trabalhos utilizaram a combinação de poucas covariáveis para integrar os modelos de predição, na Tabela 1 são apresentadas as combinações mais frequentes dentre os trabalhos avaliados.

Tabela 1. Combinações das covariáveis ambientais mais utilizadas nos trabalhos de MDS

Covariáveis ambientais	Número de trabalhos	Autores
S, R	18	Debella-Gilo et al. (2009); Carvalho Junior et al. (2011); Giasson et al. (2011); Giasson et al. (2013); Hitziger e Ließ (2014); Menezes et al. (2014); Teske et al. (2014); Bagatini et al. (2015); Giasson et al. (2015); Teske et al. (2015a); Teske et al. (2015b); Bagatini et al. (2016); Penížek et al. (2016); Silva et al. (2016); Baltensweiler et al. (2017); Wolski et al. (2017); Machado et al. (2018); Malone et al. (2018)
S, O, R	15	Liu et al. (2008); Lamsal et al. (2009); Malone et al. (2009); Stoorvogel et al. (2009); Abdel-Kader (2011); ten Caten et al. (2012); ten Caten et al. (2013); Carvalho Junior et al. (2014); Holleran et al. (2015); Szatmári et al. (2015); Söderström et al. (2016a); Demattê et al. (2016); Peng et al. (2016); Wang et al. (2016); Menezes et al. (2018)
S, O	10	Gomez et al. (2008a); Gomez et al. (2008b); Vasques et al. (2008); Nawar et al. (2015); Kumar et al. (2016); Söderström et al. (2016b); Vagen et al. (2016); Diek et al. (2017); Guo et al. (2018); Winowiecki et al. (2018)
S, O, R, P	9	Grimm et al. (2008); Grinand et al. (2008); Rodríguez-Lado et al. (2008); Chagas et al. (2010); Chagas et al. (2013); Peng et al. (2015); Chaney et al. (2016); Pinheiro et al. (2018); Vasques et al. (2016)
S, C, O, R	9	Yigini e Panagos (2014); Brogniez et al. (2015); Ballabio et al. (2016); Yigini e Panagos (2016); Forkuor et al. (2017); Nussbaum et al. (2017); Guevara et al. (2018); He et al. (2018)

Os atributos morfométricos são os preditores mais amplamente empregados no MDS devido ao seu papel importante na distribuição dos solos na paisagem e a disponibilidade de MDEs, que são fontes de dados comuns para previsão do solo, com uma variedade de resoluções espaciais (Behrens et al., 2010). A partir destes, podem ser derivados diferentes atributos topográficos como altitude, declividade, aspecto, curvaturas, índice topográfico composto, dentre outros. Vale destacar que ainda é um desafio para o MDS a espacialização de classes e propriedades de solos em áreas de relevo plano, onde as variáveis morfométricas possuem menor capacidade de predição e espacialização, necessitando o uso de outras variáveis auxiliares.

Por outro lado, ainda que a topografia seja considerada como um importante fator que influencia na distribuição espacial de classes e propriedades do solo, no estudo de Mora-Vallejo et al. (2008) embora a altitude e um número limitado de classes de declive e relevo tenham sido apontadas como variáveis explicativas nos modelos finais, os parâmetros de relevo não explicaram fortemente a variação espacial das propriedades do solo. Uma possível explicação para isso é quando se utiliza um MDE com menor resolução espacial, como o SRTM (Shuttle Radar Topography Mission - com 90 m), os principais recursos, como os terraços, não são bem representados.

Alguns trabalhos relataram as variáveis derivadas do fator organismo como as mais importantes nos modelos de predição. Grimm et al. (2008) destacam a necessidade de representações com alta resolução espacial, como as fornecidas por dados de sensoriamento remoto multi ou hiperespectral, para obter uma melhor predição de propriedades do solo, como o COS. Peng et al. (2015) observaram que no processo de modelagem, os índices de vegetação, como NDVI, EVI e NDVIgreen, foram os preditores muito importantes para a modelagem espacial do COS. Forkuor et al. (2017) apontam que o uso de dados de sensoriamento remoto pode reduzir os esforços de amostragem do solo e, portanto, reduzir os custos de mapeamento do solo.

Estudos como o de Kassai et al. (2018) avaliaram o papel da geologia na modelagem e mapeamento de solos, no qual constataram que a geologia está entre os principais preditores das propriedades do solo. Os dados de geologia foram utilizados também no estudo de Hofig et al. (2014), contudo, os autores observaram que a escala do mapa geológico (1:250.000) usado para gerar o modelo pode não ter sido a mais adequada, o que fez com que as informações geológicas contribuíssem menos do que poderiam para a predição da ocorrência dos tipos de solos. Ainda assim, a variável geológica esteve entre as mais importantes nos modelos de árvore de decisão. Silva et al. (2016) complementam que em países tropicais em desenvolvimento, como o Brasil, a maioria dos mapas geológicos limitam o mapeamento detalhado do solo, uma vez que estes mapas se encontram majoritariamente em pequena escala.

O estudo de Gray e Bishop (2016) aplicaram as covariáveis climáticas: precipitação média anual e temperatura máxima anual; associadas a modelos climáticos para mapear o carbono

orgânico do solo na Austrália e sua mudança em relação às próximas décadas. Este mesmo estudo foi um dos únicos que utilizaram o fator tempo, no qual o índice de intemperismo foi incluído como um índice para representar o grau de intemperismo do material parental com base em dados radiométricos. Já no estudo de Arruda et al. (2016) o fator tempo foi considerado indiretamente, juntamente com os fatores relevo e material parental.

Bishop et al. (2015) apontam que em muitos casos, as covariáveis podem não representar todas as condições e gradientes ambientais possíveis que afetam a variação espacial do solo. Para os autores, uma maneira de lidar com isso é incluir as coordenadas espaciais dos locais de amostra do solo como covariáveis.

3.5 Incorporação de dados legados de solos no MDS

Um total de 67 % dos trabalhos investigados utilizaram dados legados de solos. Estes dados contêm informações significativas sobre a distribuição espacial das classes e propriedades de solo e podem ser utilizados para calibrar modelos; uma vez que levantamentos de campo e análises laboratoriais são dispendiosos (Pahlavan-Rad et al., 2016).

Os trabalhos avaliados utilizaram diferentes tipos de dados legados de solos para treinar e testar os modelos de MDS, como: mapas de solos (Grimm et al., 2008; Abdel-Kader, 2011; Teske et al., 2014); levantamentos de solos (ten Caten et al., 2011; Giasson et al., 2013; Söderström et al., 2016b); banco de dados (Ballabio et al., 2016; Pinheiro et al., 2018); dados de perfis de solos (Chagas et al., 2010; Wang et al., 2016); dentre outros. Um dos exemplos de trabalho que utilizaram dados legados foi o de Ballabio et al. (2016) que mapearam propriedades físicas do solo de acordo com as especificações do *GlobalSoilMap* para a Europa, com base em 20.000 observações provenientes do banco de dados do solo *Land Use/Cover Area frame statistical Survey* - LUCAS, que consiste em um projeto destinado a coletar dados harmonizados sobre o estado do uso e cobertura da terra em toda a extensão da União Europeia.

Porém, cabe destacar alguns impasses na utilização de dados legados de solos, sobretudo, a necessidade de um mecanismo viável de compartilhamento/acessibilidade dos dados; a precisão posicional, uma vez que erros nas coordenadas geográficas dos perfis de solo, por exemplo, podem introduzir grandes erros no processo do MDS. Além disso, existe

inconsistência nos métodos de medição das propriedades do solo e sistemas de classificação entre diferentes conjuntos de dados, ressaltando a necessidade de desenvolvimento de técnicas de harmonização dos dados herdados para incorporar diferentes fontes de dados dos solos (Lagacherie, 2008; Zhang et al., 2017).

3.6 Modelos de predição de classes e propriedades de solos

Os avanços no MDS resultaram no desenvolvimento de técnicas de aprendizado de máquina, as quais baseiam-se em modelos estatísticos computacionais para encontrar padrões no conjunto de dados e a partir destes padrões realizar predições. Com base neste processo é possível então descobrir as relações entre variáveis preditoras e respostas, cujas relações aprendidas podem ser aplicadas a locais onde os dados do solo não estão disponíveis (Heung et al., 2016).

De acordo com Wang et al. (2018), o sucesso do aprendizado de máquina no MDS está relacionado a várias vantagens em relação ao levantamento tradicional do solo. Essas vantagens foram resumidas como: (1) o MDS é fácil de atualizar porque os modelos de previsão podem ser armazenados e executados novamente quando novos dados se tornam disponíveis; (2) diferentes modelos de variação espacial podem ser escolhidos devido à disponibilidade de poder computacional para processar grandes conjuntos de dados; (3) o uso adequado de ferramentas de mineração de dados e o progresso em sistemas de informações geográficas resultam em previsões com incerteza quantificada (Minasny e McBratney, 2016).

Do total de trabalhos avaliados 61 % utilizaram apenas um modelo ou algoritmo para predição de propriedades e classes de solos, 39 % utilizaram dois ou mais métodos. Heung et al. (2016) apontam que as pesquisas futuras no MDS não devem ser restritas a um único aprendiz ou a uma pequena seleção deles. Nos trabalhos que utilizaram apenas um aprendiz, foi observada a utilização de quatro principais tipos de modelos de aprendizado de máquinas. Os modelos mais aplicados foram os modelos baseados em regressão (n: 29) e modelos baseados em árvores (n: 27), enquanto os modelos baseados em redes neurais (n: 5) e modelos com aplicações em lógica fuzzy (n: 3) foram menos utilizados.

Dentre os modelos baseados em regressão, o Cubist destaca-se como uma poderosa ferramenta de mineração de dados, o qual foi empregado em cinco dos trabalhos avaliados,

como nos estudos de Panagos et al. (2014) e Gray e Bishop (2016). Este algoritmo foi desenvolvido a partir de uma versão anterior do C4.5 e da árvore de modelos M5, cujas principais vantagens estão relacionadas com a capacidade de quantificar as relações lineares e não lineares entre as variáveis e quantificar a importância das variáveis ambientais usadas na previsão (Peng et al., 2016). A regressão-krigagem, também foi empregada em cinco trabalhos, consiste em um método híbrido que combina uma regressão da variável dependente nos preditores com a krigagem simples dos resíduos da regressão (Yigini e Panagos, 2016).

Nos modelos baseados em árvores, a árvores de decisão (AD) representa o modelo popularmente utilizado (n: 15), o que pode estar relacionado com a facilidade de representação e interpretação bem como a capacidade de processar grandes volumes de dados. Os trabalhos que utilizaram as árvores de decisão - AD aplicaram, principalmente, o algoritmo “SimpleCart”, que consiste em uma implementação do algoritmo de Árvores de Classificação e Regressão (CART) proposto por (Breiman et al., 1984) e o algoritmo J48 que é uma implementação de código aberto em Java do algoritmo C4.5 proposto por Quinlan (1993). Machado et al. (2018) ao desenvolverem uma metodologia para aumentar os detalhes de um mapa de solos em Porto Alegre – Brasil, utilizaram o algoritmo J48 e relataram que a sua escolha foi baseada nos resultados satisfatórios apresentados por esse algoritmo em outros estudos do MDS, como em Giasson et al. (2011).

O random forest (RF) consiste em uma extensão das árvores de decisão, o qual foi empregado em quatro trabalhos. As principais vantagens de sua utilização é a capacidade de modelar relações dimensionais não lineares; uso de variáveis categóricas e contínuas; resistência a "overfitting"; robustez em relação à presença de ruído nos dados; estabelecimento de uma medida imparcial da taxa de erro; capacidade de determinar a relevância das variáveis utilizadas. Por outro lado, Grimm et al. (2008) destaca como principal desvantagem, a interpretação restrita dos resultados já que a relação entre preditores e respostas não pode ser analisada pormenorizadamente para cada árvore que compõe a “floresta”. No estudo de Rodríguez-Lado et al. (2015), os autores constataram o desempenho preditivo mais elevado do Random Forest em relação a Regressão Múltipla e as Redes Neurais.

Por outro lado, a menor utilização de modelos baseados em redes neurais pode estar associada, por exemplo, com a necessidade de tratamento prévio dos dados de entrada, o

grande volume de dados necessários para o processo de aprendizado e o alto custo computacional. Assim como nos modelos baseados na lógica fuzzy, que possuem como uma das principais desvantagens a dificuldade em estabelecer as regras corretamente, contudo, estudos como o de Menezes et al. (2014; 2018) empregaram este método.

3.7 Avaliação do desempenho dos modelos preditivos

As propriedades do solo ou as previsões de classes de solo fornecem avaliações quantitativas de precisão e incerteza. Estas estimativas, juntamente com as previsões espaciais do solo, são parte integrante de qualquer programa de mapeamento digital do solo (USDA, 2017). Lagacherie (2008) afirma que a sistematização e aplicação de uma metodologia de avaliação de precisão, é provavelmente o maior desafio do MDS.

Por conseguinte, Aitkenhead e Coull (2016) consideram que a validação cruzada de k-fold é uma das abordagens de melhores práticas para treinamento e validação de modelos. Este tipo de validação foi utilizada em 47 % dos estudos investigados para determinar a capacidade de generalização do modelo, sua precisão e incerteza, como nos estudos de Panagos et al. (2014) e Diek et al. (2017). A validação cruzada divide aleatoriamente os dados observados em *k* grupos e, em seguida, treina *k* modelos usando todos, exceto um dos subconjuntos onde este subconjunto de cada modelo foi usado para testes.

Aproximadamente 51 % dos estudos investigados utilizaram a validação externa ou independente para avaliar a qualidade da predição (Silva et al., 2017 e Malone et al., 2018). Neste método, o conjunto de amostras utilizado para validação é diferente do conjunto de calibração. Teske et al. (2015a) ressaltam que a avaliação de modelos preditores com dados independentes garante obter estimativas da acurácia semelhantes, indiferentemente do esquema de amostragem e do tamanho do conjunto de dados independentes.

Dentre as métricas utilizadas para avaliar o desempenho dos modelos, o erro quadrático médio da raiz (RMSE) foi o mais aplicado (n: 54) e é calculado de acordo com a equação 2:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n}} \qquad \text{(Eq. 2)}$$

em que: d é a diferença entre os valores observados e os valores preditos; e n é o número total de amostras consideradas. Assim, quanto maiores são os valores da RMSE, maiores são as discrepâncias entre os conjuntos de dados comparados (Bhering et al., 2016).

Em seguida, o coeficiente de determinação R^2 foi utilizado em 42 trabalhos, seguindo pleo o índice Kappa (n: 24) e acurácia geral (n: 21). A acurácia geral (AG) e o índice de Kappa (K) são derivados por meio das matrizes de erros. A acurácia geral expressa a proporção de classificações corretas em relação ao número total de pixels (Bagatini et al., 2016) e o índice Kappa, proposto por Congalton e Green (2008), compara a concordância entre os mapas de solo original e previsto (Giasson et al., 2011).

3.8 Custo e tempo do mapeamento de solos

Atualmente, o custo é um dos principais fatores que limitam o mapeamento de solos. No MDS o custo está diretamente relacionado com a escala de trabalho, a amostragem de campo, os métodos de processamento adotados para derivar as propriedades e classes de solo, as análises laboratoriais e o treinamento de pessoal.

Os esforços atuais para manter programas de pesquisa de solo também são caros. Por exemplo, nos Estados Unidos, as dotações de pesquisa de solo foram de US$ 43,46 milhões em 1980 e aumentaram para US$ 93,939 milhões em 2010. Em 2011 o custo para um levantamento do solo nos Estados Unidos equivalia a aproximadamente US$ 10,30 ha^{-1} (Grunwald et al., 2011). O projeto GlobalSoilMap.net tem como objetivo prever 10 propriedades do solo em seis intervalos de profundidade específicos em uma resolução de 90 m em todo o mundo (MacMillan et al., 2010), com custos estimados em US$ 0,20 ha^{-1}.

No Brasil, mais recentemente, o Programa Nacional de Solos do Brasil (PronaSolos) propôs a realização de levantamentos e mapeamentos de solos no País em caráter multiescalar, cuja estimativa orçamentária total é na ordem de 5,5 bilhões em 30 anos. Entretanto, não foi apresentado um valor aproximado por área, mas estima-se que na primeira fase do projeto (0 a 4 anos) sejam concluídos os levantamentos de solos de cerca de 430 mil km². Na segunda fase (4 a 10 anos), a meta é estender o mapeamento de solos a mais 1,3 milhões de km² de terras agricultáveis; na terceira fase (10 a 30 anos), a previsão é de se alcançar 1 milhão de km²

mapeados em escala 1:50.000, 250 mil km² em escala 1:25.000 e 6,9 milhões de km² em escala 1:100.000 (Embrapa, 2016).

No mapeamento digital de solos, a avaliação dos custos necessários e/ou envolvidos ainda é uma lacuna nos estudos. Grunwald (2009) avaliou diversos estudos de mapeamento e modelagem de solos e constatou que os trabalhos não fornecem qualquer informação sobre os custos. ten Caten et al. (2012) ao avaliarem os aspectos metodológicos de 11 trabalhos de MDS no Brasil, abordaram também que os custos envolvidos no processo não foram mencionados em nenhum dos estudos. No presente trabalho, o relato de custo foi constatado apenas em Nijbroek et al. (2018), enquanto que a demanda de tempo não foi considerada em nenhum estudo.

Nijbroek et al. (2018) estimaram o custo do trabalho de campo, transporte das amostras de solos, análises de laboratório, preparação e análise dos dados, softwares e covariáveis preditoras. Os custos para análise de dados (independentemente do número da amostra) foram de US $ 1800 e US $ 2400 para os métodos ordinary kriging e regression kriging, respectivamente, e muito mais altos (US $ 4000) para o método random forest kriging, devido ao tempo necessário para preparar os conjuntos de dados. Os autores apontam que embora todos os custos sejam estimados, os custos do trabalho de campo e de laboratório são mais fáceis de estimar do que os da preparação e análise de dados, uma vez que depende muito do nível de capacidade técnica existente. Ressaltando, dessa forma, a necessidade da capacitação técnica quando utilizados métodos que exigem conhecimento em GIS, estatística e programação.

4. Considerações finais

A maioria das informações de solos é antiga, desatualizada e em escalas generalizadas, evidenciando a necessidade de investimentos em pesquisas, aperfeiçoamento dos profissionais da área e a busca por metodologias que permitam captar com rapidez e eficiência informações sobre a variabilidade espacial dos solos. O Brasil tem contribuído de forma efetiva para as pesquisas em MDS em um panorama global, apesar de existirem diversas lacunas a serem preenchidas. Ainda que exista um número crescente de algoritmos de aprendizado de máquina, poucos estudos forneceram uma comparação do desempenho de múltiplos modelos. Ressalta-se, a necessidade do desenvolvimento e adoção de tecnologias mais sofisticadas para

medição de propriedades do solo com resolução fina e alta precisão; os quais podem operacionalizar o MDS, tornando-o mais preciso e confiável. A amostragem é uma etapa de custo elevado, e desta forma, os estudos devem elaborar alternativas de planos de amostragem que sejam mais eficientes e propiciem melhorias na exatidão dos mapas gerados com menor custo. Por fim, os dados referentes aos custos e tempo investidos durante o processo de levantamento e mapeamento de solos ainda são negligenciados nas pesquisas em MDS, e devem, portanto, ser considerados nos trabalhos futuros com o objetivo de aperfeiçoar o processo, produzir mapas pedológicos precisos e com melhor custo-benefício.

5. Bibliografia

Abdel-Kader FH. Digital soil mapping at pilot sites in the northwest coast of Egypt: A multinomial logistic regression approach. The Egyptian Journal of Remote Sensing and Space Science. 2011; 14:29-40. https://doi.org/10.1016/j.ejrs.2011.04.001

Adhikari K, Minasny B, Greve MB, Greve MH. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. Geoderma. 2014; 214–215:101–113. https://doi.org/10.1016/j.geoderma.2013.09.023

Aitkenhead MJ, Coull MC. Mapping soil carbon stocks across Scotland using a neural network model. Geoderma. 2016; 262:187–198. https://doi.org/10.1016/j.geoderma.2015.08.034

An Y, Yang L, Zhu A-X, Qin C, Shi J. Identification of representative samples from existing samples for digital soil mapping. Geoderma. 2018; 311:109–119. https://doi.org/10.1016/j.geoderma.2017.03.014

Arruda GP, Demattê JAM, Chagas CS, Fiorio PR, Souza AB, Fongaro CT. Digital soil mapping using reference area and artificial neural networks. Sci Agric. 2016; 73:266-273. http://dx.doi.org/10.1590/0103-9016-2015-0131

Arrouays D, Leenaars JGB, Richer-de-Forges AC, Adhikari K, Ballabio C, Greve M, Grundy M, Guerrero E, Hempel J, Hengl T, Heuvelink G, Batjes N, Carvalho E, Hartemink A, Hewitt A, Hong SY, Krasilnikov P, Lagacherie P, Lelyk G, Libohova Z, Lilly A, McBratney A, McKenzie N, Vasquez GM, Mulder VL, Minasny B, Montanarella L, Odeh I, Padarian J, Poggio L, Roudier P, Saby N, Savin I, Searle R, Solbovoy V, Thompson J, Smith S, Sulaeman Y, Vintila R, Rossel RV, Wilson P, Zhang GL, Swerts M, Oorts K, Karklins A, Feng L, Navarro ARI, Levin A, Laktionova T, Dell'Acqua M, Suvannang N, Ruam W, Prasad J, Patil N, Husnjak S, Pásztor L, Okx J, Hallett S, Keay C, Farewell T, Lilja H, Juilleret J, Marx S, Takata Y, Kazuyuki Y, Mansuy N, Panagos P, Liedekerke MV, Skalsky R, Sobocka J, Kobz J, Eftekhari K, Alavipanah SK, Moussadek R, Badraoui M, Silva M, Paterson G, Gonçalves MG, Theocharopoulos S, Yemefack M, Tedou S, Vrscaj B, Grob U, Kozák J, Boruvka L, Dobos E, Taboada M, Moretti L, Rodriguez D. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ. 2017; 14:1-19. https://doi.org/10.1016/j.grj.2017.06.001

Bagatini T, Giasson E, Teske R. Selection of sampling density based on data from areas already mapped for training decision tree models in digital soil mapping. R. Bras. Ci. Solo. 2015; 39:960-967. https://10.1590/01000683rbcs20140289

Bagatini T, Giasson E, Teske R. Expansão de mapas pedológicos para áreas fisiograficamente semelhantes por meio de mapeamento digital de solos. Pesq. agropec. bras. 2016; 51:1317-1325. http://dx.doi.org/10.1590/s0100-204x2016000900031

Ballabio C. Spatial prediction of soil properties in temperate mountain regions using support vector regression. Geoderma. 2009; 151:338–350. https://doi.org/10.1016/j.geoderma.2009.04.022

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

Baltensweiler A, Walthert L, Ginzler C, Sutter F, Purves RS, Hanewinkel M. Terrestrial laser scanning improves digital elevation models and topsoil pH modelling in regions with complex topography and dense vegetation. Environmental Modelling & Software. 2017; 95:13-21. https://doi.org/10.1016/j.envsoft.2017.05.009

Behrens T, Zhu A-X, Schmidt K, Scholten T. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma. 2010; 155:175–185. https://doi.org/10.1016/j.geoderma.2009.07.010

Bellamy PH, Loveland PJ, Bradley RI, Lark RM, Guy JD. Carbon losses from all soils across England and Wales 1978-2003. Nature. 2005; 437:245–248. https://doi.org/10.1038/nature04038

Bhering SB, Chagas CS, Carvalho Junior W, Pereira NR, Caldenaro Filho B, Pinheiro HSK. Mapeamento digital de areia, argila e carbono orgânico por modelos Random Forest sob diferentes resoluções espaciais. Pesq. agropec. bras. 2016; 51:1359-1370. http://dx.doi.org/10.1590/s0100-204x2016000900035

Bishop TFA, Horta A, Karunaratne SB. Validation of digital soil maps at different spatial supports. Geoderma. 2015; 241–242: 238–249. https://doi.org/10.1016/j.geoderma.2014.11.026

Bou KR, Greve MH, Bøcher PK, Greve MB, Larsen R, McCloy K. Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: The case study of Denmark. J Environ Manage. 2010; 91:1150–1160. https://doi.org/10.1016/j.jenvman.2010.01.001

Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. California: Wadsworth International Group, 1984. 358p.

Brogniez D, Ballabio C, Stevens A, Jones RJA, Montanarella L, van Wesemael B. A map of the topsoil organic carbon content of Europegenerated by a generalized additive model. European Journal of Soil Science. 2015; 66:121–134. https://doi.org/10.1111/ejss.12193

Bui E. A Review of Digital Soil Mapping in Australia. In: Lagacherie P, McBratney AB, Voltz M , editors. Developments in Soil Science. Amsterdam: Elsevier; 2006. p. 25-37. https://doi.org/10.1016/S0166-2481(06)31002-1

Carvalho CCN, Nunes FC, Antunes MAH. Histórico do levantamento de solos no Brasil: Da industrialização brasileira à era da informação. Rev. Bras. Cartogr. 2013; 65:997-1013. http://www.seer.ufu.br/index.php/revistabrasileiracartografia/article/view/43876

Carvalho Junior W, Chagas CS, Fernandes Filho EI, Vieira CAO, Schaefer CEGR, Bhering SV, Francelino M. Digital soilscape mapping of tropical hillslope areas by neural networks. Sci. agric. 2011; 68: 691-696. http://dx.doi.org/10.1590/S0103-90162011000600014

Carvalho Junior W, Chagas CS, Lagacherie P, Calderano Filho B, Bhering SB. Evaluation of statistical and geostatistical models of digital soil properties mapping in tropical mountain regions. R. Bras. Ci. Solo. 2014; 38:706-717. http://dx.doi.org/10.1590/S0100-06832014000300003

Chagas CS, Fernandes Filho EI, Vieira CAO, Schaefer CEGR, Carvalho Júnior W. Atributos topográficos e dados do Landsat7 no mapeamento digital de solos com uso de redes neurais. Pesq. agropec. bras. 2010; 45:497-507. http://dx.doi.org/10.1590/S0100-204X2010000500009

Chagas CS, Vieira CAO, Fernandes Filho EI. Comparison between artificial neural networks and maximum likelihood classification in digital soil mapping. R. Bras. Ci. Solo. 2013; 37:339-351. http://dx.doi.org/10.1590/S0100-06832013000200005

Chaney NW, Wood EF, McBratney AB, Hempel JW, Nauman TW, Brungard CW, Odgers NP. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. Geoderma. Elsevier. 2016; 274:54-67. https://doi.org/10.1016/j.geoderma.2016.03.025

Congalton R, Green K. Assessing the Accuracy of Remotely Sensed Data. Principles and Practices. 2ª ed. New York: CRC Press, 2008. 200p.

Curi NC, Ker JC, Novais RF, Vidal-Torrado P, Schaefer CEGR. Pedologia: Solos dos biomas brasileiros. 2017.

Dalmolin RSD, ten Caten A, Dotto AC. Pedometria: uma breve contextualização nacional e mundial. Boletim Informativo Soc. Bras. Ci. Solo. 2017; 43:18-21. https://www.sbcs.org.br/wp-content/uploads/2018/01/boletimsbcs32017ebook_03_01_2018_10_45_30_id_36404.pdf

Debella-Gilo M, Etzelmüller B. Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modeling integrated in GIS: Examples from Vestfold County, Norway. CATENA. 2009; 77:8-18. https://doi.org/10.1016/j.catena.2008.12.001

Demattê JAM, Alves MR, Terra FS, Bosquilia RWD, Fongaro CT, Barros PPS. Is It Possible to Classify Topsoil Texture Using a Sensor Located 800 km Away from the Surface? R. Bras. Ci. Solo. 2016; 40:1-13. http://dx.doi.org/10.1590/18069657rbcs20150335

Diek S, Fornallaz F, Schaepman ME, Jong R. Barest Pixel Composite for Agricultural Areas Using Landsat Time Series. Remote Sens. 2017; 9:1-31. https://doi.org/10.3390/rs9121245

Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA. Programa Nacional de Solos do Brasil (PronaSolos). Rio de Janeiro: Embrapa Solos, 2016. 53p.

Forkuor G, Hounkpatin OKL,Welp G, Thiel M. High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. PLOS ONE. 2017; 12:1-21. https://doi.org/10.1371/journal.pone.0170478

Giasson E, Sarmento EC, Weber E, Flores CA, Hasenack H. Decision trees for digital soil mapping on subtropical basaltic steeplands. Sci Agric. 2011; 68:167–174. http://dx.doi.org/10.1590/S0103-90162011000200006

Giasson E, Hartemink AE, Tornquist CG, Teske R, Bagatini T. Avaliação de cinco algoritmos de árvores de decisão e três tipos de modelos digitais de elevação para mapeamento digital de solos a nível semidetalhado na Bacia do Lageado Grande, RS, Brasil Cienc. Rural. 2013; 43:1967-1973. http://dx.doi.org/10.1590/S0103-84782013001100008.

Giasson E, ten Caten A, Bagatini T, Bonfatti B. Instance selection in digital soil mapping: a study case in Rio Grande do Sul, Brazil. Cienc. Rural. 2015; 45:1592-1598. http://dx.doi.org/10.1590/0103-8478cr20140694

Gomez C, Rossel RAV, McBratney AB. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma. 2008a; 146:403-411. https://doi.org/10.1016/j.geoderma.2008.06.011

Gomez C, Lagacherie P, Coulouma G. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. Geoderma. 2008b; 148:141-148. https://doi.org/10.1016/j.geoderma.2008.09.016

Gray JM, Bishop TFA. Change in Soil Organic Carbon Stocks under 12 Climate Change Projections over New South Wales, Australia. Soil Science Society of America Journal. 2016; 80:1296–1307. https://doi.org/10.2136/sssaj2016.02.0038

Grimm R, Behrens T, Märker M, Elsenbeer H. Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis. Geoderma. 2008; 146:102–113. https://doi.org/10.1016/j.geoderma.2008.05.008

Grinand C, Arrouays D, Laroche B, Martin MP. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. Geoderma. 2008; 143:180–190. https://doi.org/10.1016/j.geoderma.2007.11.004

Grunwald S. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma. 2009; 152:195–207. https://doi.org/10.1016/j.geoderma.2009.06.003

Grunwald S, Thompson JA, Boettinger JL. Digital Soil Mapping and Modeling at Continental Scales: Finding Solutions for Global Issues. Soil Sci Soc Am J. 75:1201, 2011.

Guevara M, Olmedo GF, Stell E, Yigini Y, Duarte YA, Hernández CA, Arévalo GE, Arroyo-Cruz CE, Bolivar A, Bunning S, Cañas NB, Cruz-Gaistardo CO, Davila F, Acqua MD, Encina A, Tacona HF, Fontes F, Herrera JAH, Navarro ARI, Loayza V, Manueles AM, Jara FM, Olivera C, Hermosilla RO, Pereira G, Prieto P, Ramos IA, Brina JCR, Rivera R, Rodríguez-Rodríguez J, Roopnarine R, Ibarra AR, Riveiro KAR, Schulz GA, Spence A, Vasques GM, Vargas RR, Vargas, R. No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America. SOIL. 2018; 4:173–193. https://doi.org/10.5194/soil-4-173-2018

Guo L, Linderman M, Shi T, Chen Y, Duan L, Zhang H. Exploring the Sensitivity of Sampling Density in Digital Mapping of Soil Organic Carbon and Its Application in Soil Sampling. Remote Sens. 2018; 10:1-27. https://doi.org/10.3390/rs10060888

Hartemink AE, McBratney A, Mendonça-Santos MDL. Digital Soil Mapping with Limited Data. Springer; 2008.

He S, Zhu H, Shahtahmassebi AR, Qiu L, Wu C, Shen Z, Wang K. Spatiotemporal Variability of Soil Nitrogen in Relation to Environmental Factors in a Low Hilly Region of Southeastern China. Int J Environ Res Public Health. 2018; 15:1-19. https://doi.org/10.3390/ijerph15102113

Hengl T, Rossiter DG, Stein A. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Aust J Soil Res. 2003; 41:1403–1422. https://doi.org/10.1071/SR03005

Hengl T, Jesus JM, MacMillan RA, Batjes NH, Heuvelink GBM, Ribeiro E, Samuel-Rosa A, Kempen B, Leenaars JGB, Walsh MG, Gonzalez MR. SoilGrids1km — Global Soil Information Based on Automated Mapping. PLOS ONE. 2014; 9: e114788. https://doi.org/10.1371/journal.pone.0114788

Heung B, Ho HC, Zhang J, Knudby A, Bulmer CE, Schmidt MG. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma. 2016; 265:62–77. https://doi.org/10.1016/j.geoderma.2015.11.014

Hitziger M, Ließ M. Comparison of Three Supervised Learning Methods for Digital Soil Mapping: Application to a Complex Terrain in the Ecuadorian Andes. Applied and Environmental Soil Science. 2014; 2014:1-12. http://dx.doi.org/10.1155/2014/809495

Hofig P, Giasson E, Vendrame PRS. Mapeamento digital de solos com base na extrapolação de mapas entre áreas fisiograficamente semelhantes. Pesq. agropec. bras. 2014; 49:958-966. http://dx.doi.org/10.1590/S0100-204X2014001200006

Holleran M, Levi M, Rasmussen C. Quantifying soil and critical zone variability in a forested catchment through digital soil mapping. SOIL. 2015; 1:47–64. https://doi.org/10.5194/soil-1-47-2015

Jenny H. Factors of soil formation. A system of quantitative pedology [Internet]. Dover Publ. 1941.

Jeune W, Francelino MR, Souza E de, Fernandes Filho EI, Rocha GC. Multinomial Logistic Regression and Random Forest Classifiers in Digital Mapping of Soil Classes in Western Haiti. R. Bras. Ci. Solo. 2018; 42:1-20. http://dx.doi.org/10.1590/18069657rbcs20170133

Kassai P, Sisák I. The role of geology in the spatial prediction of soil properties in the watershed of Lake Balaton, Hungary. Geologia Croatica. 2018; 71:29-39. https://doi.org/10.4154/gc.2018.04

Kempen B, Brus DJ, Heuvelink GBM, Stoorvogel JJ. Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma. 2009; 151:311-326. https://doi.org/10.1016/j.geoderma.2009.04.023

Kumar P, Pandey PC, Singh BK, Katiyar S, Mandal VP, Rani M, Tomar V, Patairiya S. Estimation of accumulated soil organic carbon stock in tropical forest using geospatial strategy. The Egyptian Journal of Remote Sensing and Space Science. 2016; 19: 109-123. https://doi.org/10.1016/j.ejrs.2015.12.003

Lagacherie P, McBratney AB. Chapter 1 Spatial Soil Information Systems and Spatial Soil Inference Systems: Perspectives for Digital Soil Mapping. Dev Soil Sci. 2007; 31:3–22. https://doi.org/10.1016/S0166-2481(06)31001-X

Lagacherie P. Digital soil mapping: a state of the art. In: Hertmink AE, McBratney A, Mendonça-Santos ML. Digital soil mapping with limited data. Springer; 2008. p. 3-14.

Lal R. Soil carbon sequestration to mitigate climate change. Geoderma. 2004; 123:1–22. https://doi.org/10.1016/j.geoderma.2004.01.032

Lamsal S, Bliss CM, Graetz DA. Geospatial Mapping of Soil Nitrate-Nitrogen Distribution Under a Mixed-Land Use System. Pedosphere. 2009; 19:434-445. https://doi.org/10.1016/S1002-0160(09)60136-3

Ließ M. Sampling for regression-based digital soil mapping: Closing the gap between statistical desires and operational applicability. Spat Stat. 2015; 13:106–122. https://doi.org/10.1016/j.spasta.2015.06.002

Lima ASL. Aplicação dos métodos semi-automático e lógica fuzzy para o mapeamento de solos da Bacia do Sarandi [dissertation]. Brasília: Universidade de Brasília; 2013.

Liu J, Pattey E, Nolin MC, Miller JR, Ka O. Mapping within-field soil drainage using remote sensing, DEM and apparent soil electrical conductivity. Geoderma. 2008; 143:261-272. https://doi.org/10.1016/j.geoderma.2007.11.011

Machado IR, Giasson E, Campos AR, Costa JJF, Silva EB da, Bonfatti BR. Spatial Disaggregation of Multi-Component Soil Map Units Using Legacy Data and a Tree-Based Algorithm in Southern Brazil. R. Bras. Ci. Solo. 2018; 42:1-14. http://dx.doi.org/10.1590/18069657rbcs20170193

Malone BP, McBratney AB, Minasny B, Laslett GM. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma. 2009; 154:138-152. https://doi.org/10.1016/j.geoderma.2009.10.007

Malone BP, McBratney AB, Minasny B. Description and spatial inference of soil drainage using matrix soil colours in the Lower Hunter Valley, New South Wales, Australia. PeerJ. 2018; 6:1-20. https://doi.org/10.7717/peerj.4659

McBratney A., Mendonça Santos M., Minasny B. On digital soil mapping. Geoderma. 2003; 117:3–52. https://doi.org/10.1016/S0016-7061(03)00223-4

MacMillan RA, Hartemink AE, McBratney AB. GlobalSoilMap.net: From planning, development and proof of concept to fullscale production mapping. 2010. In Gilkes RJ, Prakongkep N (ed.). Soil solutions for a changing world: Proc. World Congr. of Soil Sci., 19th, Brisbane, QLD, Australia. 1–6 Aug. 2010. Available at www.iuss.org/19th%20WCSS/symposium/pdf/1589.pdf. Int. Union of Soil Sci.,Wageningen, the Netherlands.

Mendonça-Santos ML, Santos HG. The state of the art of brazilian soil mapping and prospects for digital soil mapping. In: Lagacherie P, McBratney AB, Voltz M, editors. Digital soil mapping: an introductory perspective. Amsterdam: Elsevier; 2007. p.39-54.

Menezes MD, Silva SHG, Mello CR, Owens PR, Curi N. Solum depth spatial prediction comparing conventional with knowledge-based digital soil mapping approaches. Sci. Agric. 2014; 71:316-323. http://dx.doi.org/10.1590/0103-9016-2013-0416

Menezes MD, Silva SHG, Mello CR, Owens PR, Curi N. Knowledge-based digital soil mapping for predicting soil properties in two representative watersheds. Sci Agric. 2018; 75:144–153. http://dx.doi.org/10.1590/1678-992x-2016-0097

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

Mora-Vallejo A, Claessens L, Stoorvogel J, Heuvelink GBM. Small scale digital soil mapping in Southeastern Kenya. CATENA. 2008; 76:44–53. https://doi.org/10.1016/j.catena.2008.09.008

Nawar S, Buddenbaum H, Hill J. Digital Mapping of Soil Properties Using Multivariate Statistical Analysis and ASTER Data in an Arid Region. Remote Sens. 2015; 7:1181-1205. https://doi.org/10.3390/rs70201181

Nijbroek R, Piikki K, Söderström M, Kempen B, Turner KG, Hengari S, Mutua J. Soil Organic Carbon Baselines for Land Degradation Neutrality: Map Accuracy and Cost Tradeoffs with Respect to

Complexity in Otjozondjupa, Namibia. Sustainability. 2018; 10:1-20. https://doi.org/10.3390/su10051610

Nussbaum M, Walthert L, Fraefel M, Greiner L, Papritz A. Mapping of soil properties at high resolution in Switzerland using boosted geoadditive models. SOIL. 2017; 3:191–210. https://doi.org/10.5194/soil-3-191-2017

Pahlavan-Rad MR, Khormali F, Toomanian N, Brungard CW, Kiani F, Komaki CB, Bogaert P. Legacy soil maps as a covariate in digital soil mapping: A case study from Northern Iran. Geoderma. 2016; 279:141–148. https://doi.org/10.1016/j.geoderma.2016.05.014

Panagos P, Meusburger K, Ballabio C, Borrelli P, Alewell C. Soil erodibility in Europe: A high-resolution dataset based on LUCAS. Science of the Total Environment. 2014; 479–480:189–200. https://doi.org/10.1016/j.scitotenv.2014.02.010

Pásztor L, Laborczi A, Takács K, Szatmárig G, Dobos E, Illés G, Bakacsi Z, Szabó J. Compilation of novel and renewed, goal oriented digital soil maps using geostatistical and data mining tools. Hungarian Geographical Bulletin. 2015; 1:49-64. https://doi.org/10.15201/hungeobull.64.1.5

Peng Y, Xiong X, Adhikari K, Knadel M, Grunwald S, Greve MH. Modeling Soil Organic Carbon at Regional Scale by Combining Multi-Spectral Images with Laboratory Spectra. PLOS ONE. 2015; 10:1-22. https://doi.org/10.1371/journal.pone.0142295

Peng Y, Kheir RB, Adhikari K, Malinowski R, Greve MB, Knadel M, Greve MH. Digital Mapping of Toxic Metals in Qatari Soils Using Remote Sensing and Ancillary Data. Remote Sens. 2016; 8:1-19. https://doi.org/10.3390/rs8121003

Penížek V, Zádorová T, Kodešová R, Vaněk A. Influence of Elevation Data Resolution on Spatial Prediction of Colluvial Soils in a Luvisol Region. PLOS ONE. 2016; 10:1-18. https://doi.org/10.1371/journal.pone.0165699

Pinheiro HSK, Carvalho Junior W de, Chagas C da S, Anjos LHC dos, Owens PR. Prediction of Topsoil Texture Through Regression Trees and Multiple Linear Regressions. R. Bras. Ci. Solo. 2018; 42:1-21. https://doi.org/10.1590/18069657rbcs20170167

Qin C-Z, Qiu W-L, Lu Y-J, Li B-L, Pei T. Mapping soil organic matter in small low-relief catchments using fuzzy slope position information. Geoderma. 2012; 171:64–74. https://doi.org/10.1016/j.geoderma.2011.06.006

Quinlan JR. C4.5: Programs for machine learning. Morgan Kaufmann PublishersInc., San Francisco, CA, USA, 1993.

Rodríguez-Lado L, Hengl T, Reuter HI. Heavy metals in European soils: A geostatistical analysis of the FOREGS Geochemical database. Geoderma. 2008; 148:189-199. https://doi.org/10.1016/j.geoderma.2008.09.020

Rodríguez-Lado L, Martínez-Cortizas A. Modelling and mapping organic carbon content of topsoils in an Atlantic area of southwestern Europe (Galicia, NW-Spain). Geoderma. 2015; 245–246:65–73. https://doi.org/10.1016/j.geoderma.2015.01.015

Samuel-Rosa A. Funções de predição espacial de propriedades do solo [dissertation]. Santa Maria: Universidade Federal de Santa Maria; 2012. p.201.

Silva SHG, Poggere GC, Menezes MD de, Carvalho GS, Guilherme LRG, Curi N. Proximal Sensing and Digital Terrain Models Applied to Digital Soil Mapping and Modelingof Brazilian Latosols (Oxisols). Remote Sens. 2016; 614:1-22. https://doi.org/10.3390/rs8080614

Silva SHG, Teixeira AF dos S, Menezes MD de, Silva SHG, Moreira FM de S, Curi N. Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence spectrometer (pXRF). Ciênc. agrotec. 2017; 41:648-664. http://dx.doi.org/10.1590/1413-70542017416010317

Schmidt MWI, Torn MS, Abiven S. Persistence of soil organic matter as an ecosystem property. Nature. 2011; 478:49-56. https://doi.org/10.1038/nature10386

Schmidt S, Ballabio C, Alewell C, Panagos P, Meusburger K. Filling the European blank spot—Swiss soil erodibility assessment with topsoil samples. Journal of Plant Nutrition and Soil Science. 2018; 181:737–748. https://doi.org/10.1002/jpln.201800128

Söderström M, Sohlenius G, Rodhe L, Piikki K. Adaptation of regional digital soil mapping for precision agriculture. Precision Agriculture. 2016a; 17:588–607. https://doi.org/10.1007/s11119-016-9439-8

Söderström M, Eriksson J, Isendahl C, Schaan DP, Stenborg P, Rebellato L, Piikki K. Sensor mapping of Amazonian Dark Earths in deforested croplands. Geoderma. 2016b; 281:58-68. https://doi.org/10.1016/j.geoderma.2016.06.024

Stoorvogel JJ, Kempen B, Heuvelink GBM, de Bruin S. Implementation and evaluation of existing knowledge for digital soil mapping in Senegal. Geoderma. 2009; 149:161–170. https://doi.org/10.1016/j.geoderma.2008.11.039

Szatmári G, Barta K, Pásztor L. An application of a spatial simulated annealing sampling optimization algorithm to support digital soil mapping. Hungarian Geographical Bulletin. 2015; 64:35-48. https://doi.org/10.15201/hungeobull.64.1.4

ten Caten A, Dalmolin RSD, Pedron FA, Mendonça-Santos ML. Multivariate analysis applied to reduce the number of predictors in digital soil mapping. Pesq. agropec. bras. 2011; 46:553-561. http://dx.doi.org/10.1590/S0100-204X2011000500014.

ten Caten A, Dalmolin RSD, Mendonça-Santos ML, Giasson E. Mapeamento digital de classes de solos: características da abordagem brasileira. Cienc. Rural. 2012; 42:1989–1997. http://dx.doi.org/10.1590/S0103-84782012001100013

ten Caten A, Dalmolin RSD, Pedron FA, Ruiz LFC, Silva CA. An appropriate data set size for digital soil mapping in Erechim, Rio Grande do Sul, Brazil. R. Bras. Ci. Solo. 2013; 37:359-366. http://dx.doi.org/10.1590/S0100-06832013000200007

Teske R, Giasson E, Bagatini T. Comparação do uso de modelos digitais de elevação em mapeamento digital de solos em Dois Irmãos, RS, Brasil. R. Bras. Ci. Solo. 2014; 38:1367-1376. http://dx.doi.org/10.1590/S0100-06832014000500002

Teske R, Giasson E, Bagatini T. Comparação de esquemas de amostragem para treinamento de modelos preditores no mapeamento digital de classes de solos R. Bras. Ci. Solo. 2015a; 39:14–20. http://dx.doi.org/10.1590/S0100-06832014000500002

Teske R, Giasson E, Bagatini T. Produção de um mapa pedológico associando técnicas comuns aos mapeamentos digitais de solos com delineamento manual de unidades de mapeamento. R. Bras. Ci. Solo. 2015b; 39:950-959. http://dx.doi.org/10.1590/01000683rbcs20140285

Tiessen H, Cuevas E, Chacon P. The role of soil organic matter stability in soil fertility and agricultural potential. Nature. 1994; 371:783-785. https://doi.org/10.1038 / 371783a0
USDA. Soil Survey Manual. Handbook n.18, 2017.

Vågen TG, Winowiecki LA, Tondoh JE, Desta LT, Gumbricht T. Mapping of soil properties and land degradation risk in Africa using MODIS reflectance. Geoderma. 2016; 263: 216-225. https://doi.org/10.1016/j.geoderma.2015.06.023

Vasques GM, Grunwald S, Sickman JO. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. Geoderma. 2008; 146:14-25. https://doi.org/10.1016/j.geoderma.2008.04.007

Vasques GM, Coelho MR, Dart RO, Oliveira RP, Teixeira WG. Mapping soil carbon, particle-size fractions, and water retention in tropical dry forest in Brazil. Pesq. agropec. bras. 2016; 51:1371-1385. http://dx.doi.org/10.1590/s0100-204x2016000900036

Wang B, Waters C, Orgill S, Gray J, Cowie A, Clark A, Liu DL. High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia. Sci Total Environ. 2018; 630:367–378. https://doi.org/10.1016/j.scitotenv.2018.02.204

Wang S, Wang Q, Adhikari K, Jia S, Jin X, Liu H. Spatial-Temporal Changes of Soil Organic Carbon Content in Wafangdian, China. Sustainability. 2016; 8:1-16. https://doi.org/10.3390/su8111154

Winowiecki LA, Vågen TG, Kinnaird MF, O'Brien TG. Application of systematic monitoring and mapping techniques: Assessing land restoration potential in semi-arid lands of Kenya. Geoderma. 2018; 327:107-118. https://doi.org/10.1016/j.geoderma.2018.04.017

Wolski MS, Dalmolin RSD, Flores CA, Moura-Bueno JM, ten Caten A, Kaiser DR. Digital soil mapping and its implications in the extrapolation of soil-landscape relationships in detailed scale. Pesq. agropec. bras. 2017; 52:633-642. http://dx.doi.org/10.1590/s0100-204x2017000800009

Yigini Y, Panagos P. Reference Area Method for Mapping Soil Organic Carbon Content at Regional Scale. Procedia Earth and Planetary Science. 2014; 10:330–338. https://doi.org/10.1016/j.proeps.2014.08.028

Yigini Y, Panagos P. Assessment of soil organic carbon stocks under future climate and land cover changes in Europe. Science of The Total Environment. 2016; 557–558:838-850. https://doi.org/10.1016/j.scitotenv.2016.03.085

Zhang G, Liu F, Song X. Recent progress and future prospect of digital soil mapping: A review. Journal of Integrative Agriculture. 2017; 16:2871–2885. https://doi.org/10.1016/S2095-3119(17)61762-3

Zhu A-X, Yang L, Li B, Qin C, Pei T, Liu B. Construction of membership functions for predictive soil mapping under fuzzy logic. Geoderma. Elsevier; 155:164–174, 2010.

Artigo 2

ANÁLISE DOS ATRIBUTOS DO TERRENO EM DIFERENTES RESOLUÇÕES ESPACIAIS PARA APLICAÇÃO EM MAPEAMENTO DIGITAL DE SOLOS NO SUDESTE DO BRASIL

1. Introdução

Os modelos digitais de elevação (MDEs) fornecem informações importantes sobre as variações morfométricas da superfície terrestre. Por meio dos MDEs é possível extrair os atributos do terreno, que podem ser divididos em primários e secundários, os quais representam mensurações quantitativas da superfície terrestre (Mattivi et al., 2019). Os atributos primários podem ser obtidos diretamente a partir do DEM e incluem variáveis como: elevação, declividade, plano e perfil de curvatura, enquanto os atributos secundários envolvem combinações dos atributos primários, como por exemplo, a radiação solar, o índice de umidade, índice de transporte de sedimentos e outros (Oksanen e Sarjakoski, 2005); que podem ser usados para caracterizar a variabilidade espacial de processos específicos que ocorrem na paisagem (Moore et al., 1993; Sirtoli et al., 2008).

Os atributos do terreno gerados a partir do MDE integram o fator de formação relevo (r), descrito na função scorpan, a qual foi proposta por McBratney et al. (2003). Estes atributos são amplamente utilizados no mapeamento digital de solos (MDS) como variáveis auxiliares na predição espacial de classes e propriedades de solos, como composição granulométrica, carbono orgânico do solo, cor, umidade, entre outras (Mendonça-Santos et al., 2007; Kempen et al., 2011; Ballabio et al., 2012; Samuel-Rosa et al., 2013; Teske et al., 2014), devido ao seu importante papel no processo pedogenético e à crescente disponibilidade de modelos digitais de elevação em diferentes resoluções.

As resoluções dos modelos podem variar de < 1 m para dados gerados a partir de levantamento com tecnologia LIDAR (Light Detection And Ranging) até 1 km para conjuntos de dados com cobertura global (McBratney et al., 2003; Behrens et al., 2010). Com o advento dos MDEs em diferentes resoluções, têm-se discutido cada vez mais a influência da resolução desses modelos para análise multiescalar na modelagem ambiental (Smith et al., 2006; Zhu, 2008; Behrens et al., 2010; Drăguţ et al., 2011).

Behrens et al., (2010) ressalta a importância da análise digital do terreno em múltiplas escalas na previsão das classes de solo, uma vez a pedogênese é influenciada por interações entre processos paisagísticos e ambientais em diferentes escalas (Kerry e Oliver, 2011; Viscarra e Rossel, 2011; Behrens et al., 2014; Miller et al., 2015). Contudo, este tipo de abordagem, que incorpora múltiplas escalas dos MDEs para previsão de classes de solos ainda é limitada nos estudos de MDS, como relatado por Drăguţ et al. (2011); Li et al., (2017) e Behrens et al., (2018). Vários autores mostraram que as abordagens existentes para o mapeamento de solo que incorporam dados em escala múltipla fornecem melhor precisão da predição (Behrens et al., 2010; Miller et al., 2015; Sun et al., 2017). No entanto, a maioria das abordagens é limitada em termos do número de escalas e/ou da escala máxima, eficiência computacional e/ou restrita a um subconjunto limitado de atributos do terreno (Behrens et al., 2014).

Portanto, estudos que integram atributos em várias escalas devem ser considerados como uma ferramenta importante para o mapeamento de solos e para o entendimento dos processos pedogenéticos, cujas múltiplas escalas podem fornecer diferentes tipos de informações com níveis de detalhes diferenciados. Tais conhecimentos podem ser úteis, sobretudo, para diminuir o custo do MDS e selecionar um conjunto de atributos com melhor poder explicativo e capacidade preditiva na previsão de classes de solos. Neste sentido, o objetivo do presente estudo foi analisar o efeito das diferentes resoluções espaciais dos MDEs e atributos derivados do terreno e suas implicações para aplicação em modelos preditivos do solo. Para isso, o estudo apresentou uma abordagem relacionando os MDEs com um conjunto de ferramentas estatísticas e técnicas de aprendizado de máquinas, que são ainda pouco exploradas no MDS.

2. Material e métodos

2.1. Descrição da área de estudo

A área de estudo localiza-se no município de Mariana, estado de Minas Gerais - Brasil e compreende uma área de aproximadamente 144 km² (Figura 1). Esta área destaca-se por ser uma das poucas em que existe disponível uma base de dados composta por MDEs em diferentes resoluções espaciais.

Segundo a classificação de Koeppen, o clima é do tipo Cwa caracterizado como tropical semiúmido, com duas estações climáticas bem definidas: (1) inverno seco, com temperaturas

inferiores a 18ºC; (2) verão úmido, com temperaturas superiores a 22ºC (Garcia, 2007; Varajão 2009). A vegetação predominante são os Campos Rupestres em relevo marcado por cristas, escarpas, encostas do Quadrilátero Ferrífero e Planaltos dissecados do Rio Piracicaba e do alto Rio Doce (IBGE, 2001; Garcia, 2007). As principais classes de solos encontradas são Latossolos, Argissolos, Cambissolos e Neossolos Litólicos (UFV et al., 2010).

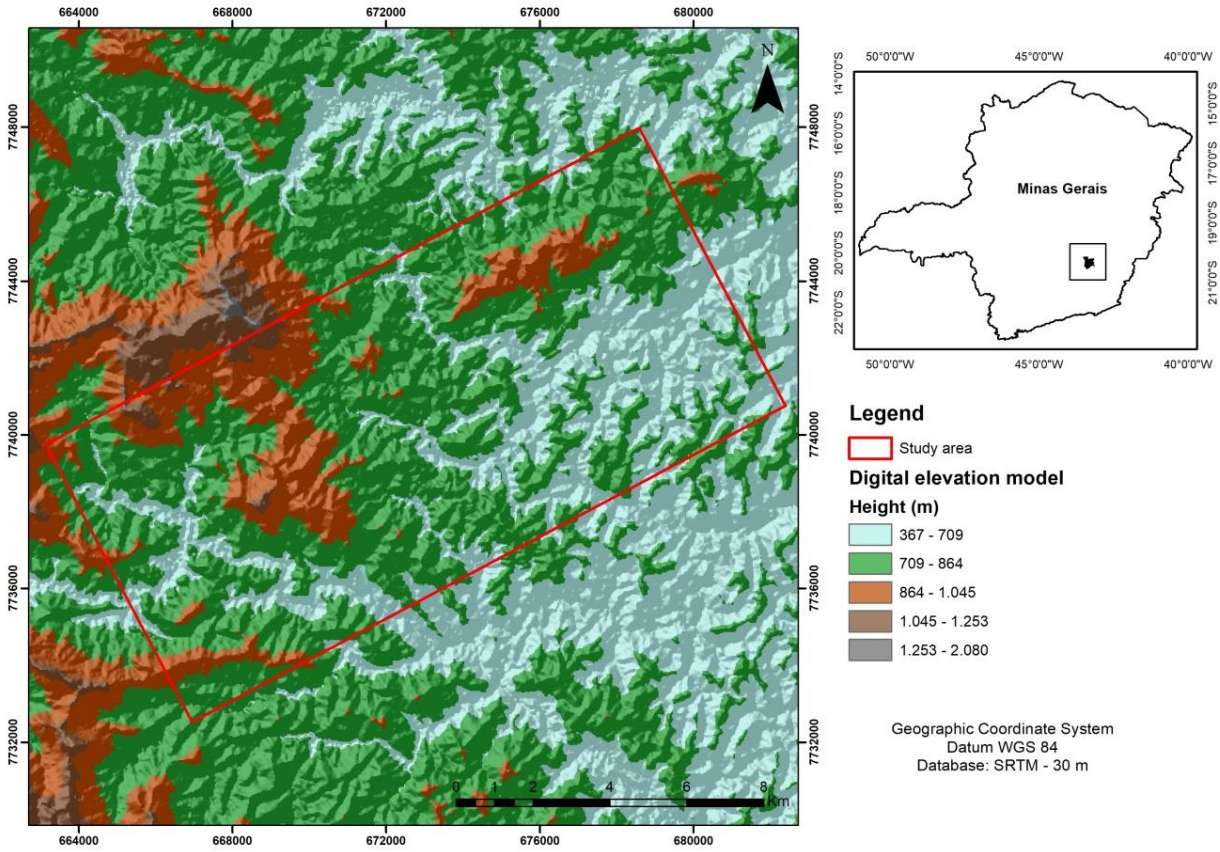


Figura 1. Mapa de localização da área de estudo.

2.2. Caracterização dos modelos digitais de elevação

2.2.1 SRTM DEM

O SRTM (*Shuttle Radar Topography Mission*) consiste em um projeto realizado pela National Aeronautics and Space Administration – NASA em parceria com Agências Espaciais da Alemanha e Itália. A missão utilizou um radar imageador por técnicas de InSAR (interferometria de radar abertura sintética), a bordo do ônibus espacial Endeavour, que recobriu aproximadamente 80% da área do globo terrestre entre latitudes 56º S e 60º N, com a captura de dados por dois sensores de radar ativo, o SIR-C (*Spaceborn Imaging Radar*) e sensores do tipo X-SAR (*Synthetic Aperture Radar*). Nas técnicas de interferometria a elevação é obtida através do cálculo da diferença de fase do sinal retroespalhado, tomada sob

geometrias distintas. Inicialmente seus dados foram disponibilizados com resolução de 1 segundo de arco (≅ 30 m) apenas para os Estados Unidos, sendo o restante da cobertura possuindo uma resolução de 3 segundos de arco. Contudo, em setembro de 2014, os dados SRTM com a resolução de 1 segundo de arco foram disponibilizados para todo o planeta (NASA, 2019).

2.2.2 ASTER GDEM

O ASTER (*Advanced Spaceborne Thermal Emission and Reflection Radiometer*) foi viabilizado através de uma parceria entre a NASA e o Ministério da Economia, Comércio e Indústria do Japão - METI. Os MDEs do instrumento ASTER são denominados de ASTER GDEM (*Global Digital Elevation Model*) que recobrem aproximadamente 99% da superfície terrestre entre as latitudes 83° N e 83° S. As imagens ASTER GDEM possuem resolução espacial 1 segundo de arco (≅ 30 m) e exatidão de elevação no MDE entre 7 e 14 m. Para a obtenção dos dados de elevação, o instrumento opera no modo estereoscópico, com as bandas 3N (nadir) e 3B (backward), correspondentes à faixa do infravermelho próximo (0,78 – 0,86 μm) (Rodrigues et al., 2010).

2.2.3 ALOS DEM

O satélite ALOS foi lançado em 2006 pela Japan Aerospace Exploration Agency – JAXA, e era equipado com três instrumentos: dois instrumentos ópticos, o PRISM (Panchromatic Remote-sensing Instrument for Stereo Mapping) e AVNIR-2 (Advanced Visible and Near-Infrared Radiometer type 2) e um radar polarimétrico de abertura sintética de banda larga PALSAR (Phased Array L-band Synthetic Aperture Radar). O PALSAR é um radar de abertura sintética que opera na Banda L, capaz de obter imagens diurnas ou noturnas e em quaisquer condições atmosféricas. Através destes dados do PALSAR foram criados produtos *radiometrically terrain-corrected* (RTC), no qual foi possível corrigir a geometria e a radiometria do radar de abertura sintética, gerando produtos MDE de alta resolução (12.5 m) e de baixa resolução (30 m) para todas as áreas terrestres globais, exceto na Antártida, Groelândia, Islândia e norte da Eurásia (JAXA, 2019).

2.2.4 LIDAR DEM

O sistema LIDAR (*Light Detection and Ranging*) é uma tecnologia que permite a obtenção de informações tridimensionais acerca da superfície terrestre com alta precisão, cujo princípio de

funcionamento está baseado na emissão de um pulso laser sob uma determinada taxa de frequência de repetição (Santos, 2006). O LIDAR fornece como produto uma nuvem maciça de pontos com coordenadas tridimensionais da superfície terrestre. Para obtenção do MDE, são considerados apenas os dados de elevação, deste modo, esta nuvem é submetida a um processo de filtragem para remoção de pontos da superfície, tais como vegetação e edificações. Neste trabalho, foram utilizados dados do levantamento com sensor aerotransportado realizado pela companhia Vale S/A. O equipamento utilizado foi o ALS50, com abertura de 70º e frequência de até 150 kHz. Os pontos foram transformados em um MDE com resolução de 2 x 2 m através da média da altitude dos pontos incluídos na área da célula.

2.3 Obtenção dos atributos do terreno

Foram utilizadas três resoluções espaciais a partir de quatro diferentes MDEs: (1) Levantamento altimétrico a laser do LIDAR com resolução de 2 m; (2) ALOS PALSAR com resolução de 12.5 e 30 m; (3) SRTM com resolução de 30 m; e (4) ASTER GDEM com resolução de ≅ 30 m.

A obtenção dos atributos do terreno foi realizada através do programa computacional R (R CORE TEAM, 2017), utilizando os pacotes *"rsaga"* (Brenning, 2008), *"raster"* (Robert, 2019) e *"rgrass7"* (Bivand et al., 2019), onde foram derivados 34 atributos do terreno de cada MDE, totalizando 170 atributos, descritos na Tabela 1.

Tabela 1. Atributos do terreno derivados dos modelos digitais de elevação

Atributos do terreno	Abreviações	Breve descrição
Aspect	ASP	Orientação das vertentes
Convergence index	CI	Índice de convergência/divergência em relação ao escoamento superficial
Cross sectional curvature	CSC	Curvatura transversal
Diurnal anisotropic heating	DAH	Medida contínua da energia dependente de exposição
Easterness	E	Seno do aspecto
Flow line curvature	FLC	Curvatura da linha de fluxo
General curvature	GC	Curvatura geral
Gradient	G	Corresponde ao gradiente hidrológico
Longitudinal curvature	LC	Curvatura longitudinal
Mass balance index	MBI	Índice de balanço entre erosão e deposição
Maximal curvature	MAXC	Curvatura máxima na seção normal local
Digital elevation model	DEM	Representa a elevação em cada célula do modelo
Mid-slope position	MSP	Representa a distância em relação ao topo e vale, variando entre 0 e 1

Minimal curvature	MINC	Curvatura mínima referente a seção normal local
Multiresolution index of ridge top flatness	MRRTF	Indica posições planas em áreas de alta altitude
Multiresolution index of valley bottom flatness	MRVBF	Indica superfícies planas no fundo do vale
Normalized height	NH	Distância vertical entre a base e o cume do declive normalizada
Northerness	N	Cosseno do aspecto
Plan curvature	PLANC	Plano de curvatura
Profile curvature	PROC	Descreve o segundo mecanismo de acumulação
Real surface área	RSA	Cálculo real da área da célula
Slope	S	Representa a declividade angular local
Slope height	SH	Distância vertical entre a base e o cume do declive
Standardized height	STANH	Distância vertical entre a base e o cume do declive padronizada
Surface specific points	SSP	Indica diferenças entre pontos específicos de mudança da superfície
Tangencial curvature	TANC	Descreve o primeiro mecanismo de acumulação
Terrain ruggedness index	TRI	Índice quantitativo da heterogeneidade da topografia
Terrain surface convexity	TSC	Convexidade da superfície do terreno
Terrain surface texture	TST	Textura da superfície do terreno
Total curvature	TC	Curvatura total
Topographic position index	TPI	Diferença entre a elevação de um ponto com a elevação do entorno
Valley depth	VD	Cálculo da distância vertical ao nível de base da rede de drenagem
Vector ruggedness measure	VRM	Mede a variação na rugosidade do terreno
Topographic wetness index	TWI	Descreve a tendência de cada célula em acumular água em função do relevo

2.4 Análises multivariadas

2.4.1 Análise de correlação

Para avaliar o grau de correlação entre as variáveis utilizou-se o coeficiente de correlação de Pearson (*r*) (Pearson, 1895), o qual avalia a relação linear das variáveis, medindo a intensidade (fraca ou forte) e a direção da correlação (positiva ou negativa). O cálculo de *r* é obtido através da Equação 1:

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2 | n\sum Y - (\sum Y)^2]}} \quad \text{(Eq. 1)}$$

Sendo: Y = valores dispostos no eixo vertical; a = ordenada à origem, ou intercessão no eixo dos Y; b = coeficiente angular; X = valores dispostos no eixo horizontal; n= número de períodos observados; r = índice de correlação.

O valor de *r* pode variar de -1 a +1, quanto maior for o valor absoluto do coeficiente, mais forte é a relação entre as variáveis. Um valor absoluto de 1 indica uma relação linear perfeita. Caso esteja próximo de + 1 existe correlação positiva entre as variáveis, se o valor de *r* estiver próximo de -1 a correlação entre as variáveis é negativa. A correlação próxima de 0 indica que não há relação linear entre as variáveis.

As correlações de Pearson foram calculadas com base nos atributos derivados a partir de dois MDEs: (1) o LIDAR 2 m – modelo com menor tamanho de pixel disponível, e (2) o SRTM 30 m – modelo com maior tamanho de pixel, um dos mais empregados globalmente em pesquisas científicas e disponível gratuitamente.

2.4.2 Análise de cluster

A análise de cluster é um método que permite agrupar observações em grupos homogêneos em função do grau de similaridade (Hair Junior et al., 1998; Fávero et al., 2009). Quando representados graficamente, as observações dentro dos agrupamentos estarão próximas, enquanto que as observações de diferentes grupos estarão distantes (Hair Junior et al., 1998).

Foi aplicado o método K-means, um dos procedimentos não hierárquicos mais conhecidos, no qual cada cluster é representado por seu centróide, que corresponde à média dos pontos atribuídos ao cluster. Cada um dos atributos restantes é atribuído ao seu centróide mais próximo, onde o mais próximo é definido usando a distância euclidiana entre o atributo e a média do agrupamento.

Os atributos foram então agrupados em função do grau de sensibilidade à resolução em: sensíveis à resolução (SR), pouco sensíveis à resolução (LSR) e altamente sensíveis à resolução (HSR) (Schunemann, 2016).

2.4.3 Análise de componentes principais

Os atributos do terreno foram submetidos à análise de componentes principais (PCA), que a partir da matriz de correlação *r*, permitiu transformar um conjunto de variáveis Z_1, Z_2, ... , Z_p em um novo conjunto de variáveis Y_1 (CP1), Y_2 (CP2), ... , Y_p (CPp). Assim, tem-se um novo conjunto de p variáveis não correlacionadas entre si e arranjadas em ordem decrescente de variância (Cruz et al., 2004).

O método baseia-se na premissa de que os primeiros componentes principais contenham a maior variabilidade dos dados originais; sendo possível descartar os demais componentes e reduzir o número de variáveis. Neste trabalho, foram utilizados apenas os dois primeiros componentes, pois foram considerados suficientes para explicar a maior parte da variabilidade dos dados. Os resultados foram representados bidimensionalmente em gráficos denominados biplot PC1 x PC2.

Todas as análises foram realizadas no software R 3.5.3 (R Core Team, 2019) a partir de um script utilizando os pacotes "*rgdal*" (Bivand et al., 2018), "*raster*" (Robert, 2019), "*ggplot2*" (Wickham et al., 2018), "*GGally*" (Schloerke et al., 2018), "*gplots*" (Warnes et al., 2019), "*dplyr*" (Wickham et al., 2019), "*factoextra*" (Kassambara e Mundt, 2017) e "*factoMineR*" (Husson et al*., 2018*).

2.5 Predição e validação

A predição das classes de solos foi realizada utilizando cada agrupamento dos atributos do terreno, os quais foram apresentados na Tabela 2 e definidos conforme sensibilidade à resolução (LSR, SR, HSR) discutidos na Tabela 3, em relação às diferentes resoluções espaciais dos MDEs (SRTM – 30 m, ALOS PALSAR - 12,5 m e LIDAR 2 m).

Tabela 2. Conjunto de atributos utilizados nos modelos preditivos conforme os agrupamentos de sensibilidade à resolução

Agrupamentos		
LSR	**SR**	**HSR**
Diurnal anysotropic heating	Aspect	Cross sectional curvature
Easterness	Gradient	Flow line curvature
Digital elevation model	Mid slope position	General curvature
Normalized height	Multiresolution index of valley bottom flatness	Longitudinal curvature
Northerness	Real surface área	Maximal curvature
Slope height	Slope	Minimal curvature
Standardized height	Terrain ruggedness index	Plan curvature
Topographic position index	Topographic wetness index	Profile curvature
		Tangencial curvature
		Total curvature
		Mass balance index
		Surface specific points

Foram utilizados os algoritmos Random Forest (RF) e Support Vetor Machine (SVM). Os algoritmos foram implementados nos pacotes e1071 (Meyer et al., 2019) e RandomForest (Breiman et al., 2018), respectivamente, no software R (R Core Team, 2019).

Random Forest

O RF foi desenvolvido por Breiman (2001) como extensão do programa CART (Classification and Regression Trees). Consiste em uma técnica não paramétrica que combina previsões feitas por múltiplas árvores de decisões, onde cada árvore é gerada baseada nos valores de um conjunto independente de vetores aleatórios (Tan et al., 2009). Cada um destes conjuntos é criado por um tipo de amostragem chamado de bootstrap (Han et al., 2011). Para tanto são definidos três parâmetros: o número de árvores (ntree), o número mínimo de dados em cada nó terminal (nodesize) e o número de variáveis utilizadas em cada árvore (mtry) (Liaw e Wiener, 2002). O mtry é o único parâmetro que requer julgamento especial (Breiman, 2002), o qual é utilizado para a otimização do modelo no pacote Caret (Kuhn et al., 2020).

Support Vetor Machine

O SVM compreende um conjunto de técnicas de aprendizado supervisionado, proposto por Cortes e Vapnik (1995), baseado na utilização de hiperplanos para a separação ideal entre as classes de um conjunto de dados (Hastie et al., 2009), maximizando a margem entre os pontos das duas classes mais próximas, chamados "vetores de suporte" e levando a uma melhor probabilidade de generalização (Ließ et al., 2016). O desempenho da generalização do SVM depende de uma boa configuração dos hiperparâmetros (Cherkassky e Ma, 2004). Os parâmetros usados no ajuste do modelo no pacote Caret (Kuhn et al., 2020) inclui a penalidade (custo) que controla o trade-off entre erros de margem e erros de treinamento e a largura do kernel (sigma) que controla o grau de não linearidade do modelo (Naghibi et al., 2017).

Foram extraídas 1.800 amostras de classes de solos no segundo nível categórico do Sistema Brasileiro de Classificação de Solos (Embrapa, 2018), oriundas do mapa de solos legado do município de Mariana (1:50.000), sendo estas divididas em 75% para treinamento e teste e 25 % para validação.

A otimização de cada hiperparâmetro dos modelos foi testada utilizando 5 valores (tuneLenght) definidos aleatoriamente pelo pacote Caret, sendo avaliados pela validação cruzada 5 fold. Para cada modelo, o processo foi repetido 50 vezes com seu próprio subconjunto de variáveis e comparado pelos valores médios dos parâmetros de precisão. O processo de várias repetições é importante para determinar a variabilidade da previsão já que

diferentes grupos de conjuntos de dados de treinamento e validação podem gerar resultados de precisão diferentes (Kuhn e Johnson, 2013).

Para avaliar o desempenho dos algoritmos, foi utilizada a matriz de confusão, a partir da qual foi derivado o índice Kappa (κ). Este índice consiste em uma medida da precisão da classificação, responsável pelo acordo de chance, descrito na equação 2.

$$\kappa = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}(x_{i+} * x_{+i})}{N^2 - \sum_{i=1}^{r}(x_{i+} * x_{+i})} \quad \text{(Eq. 2)}$$

em que: κ = índice de exatidão Kappa; r = número de linhas da matriz; Xii = número de observações na linha i e coluna i; Xi+ e X+i = totais marginais da linha i e coluna i, respectivamente; N = número total de observações. Valores de κ maiores que 0,80 representam concordância forte, valores entre 0,4 e 0,8 representam concordância moderada e valores abaixo de 0,4 representam baixa concordância (Congalton e Green, 1998).

3. Resultados e Discussão

3.1 Correlação entre os atributos

A maioria das correlações entre os atributos derivados dos diferentes MDEs com base no LIDAR 2 m são positivas e fracas com valores de r < 0.4 (Figura 2). Apenas os atributos "*vector ruggedness measure*" – VRM e "*terrain surface texture*" - TST apresentaram correlação negativa em relação aos demais MDEs. Os atributos associados às curvaturas ("*total curvature"* - TC, *"flow line curvature*" – FLC e "*plan curvature*" - PLANC) apresentaram menor intensidade das relações tanto para as resoluções mais finas como para as mais grosseiras, com os menores valores de correlação.

Essas correlações positivas fracas entre as curvaturas foram observadas também por outro autores como Deng et al. (2007) e Neuman et al. (2018) que constataram menores valores de r, como por exemplo, para "*plan and profile curvature*", independente da base utilizada, sendo estas consideradas mais sensíveis à resolução entre os atributos testados.

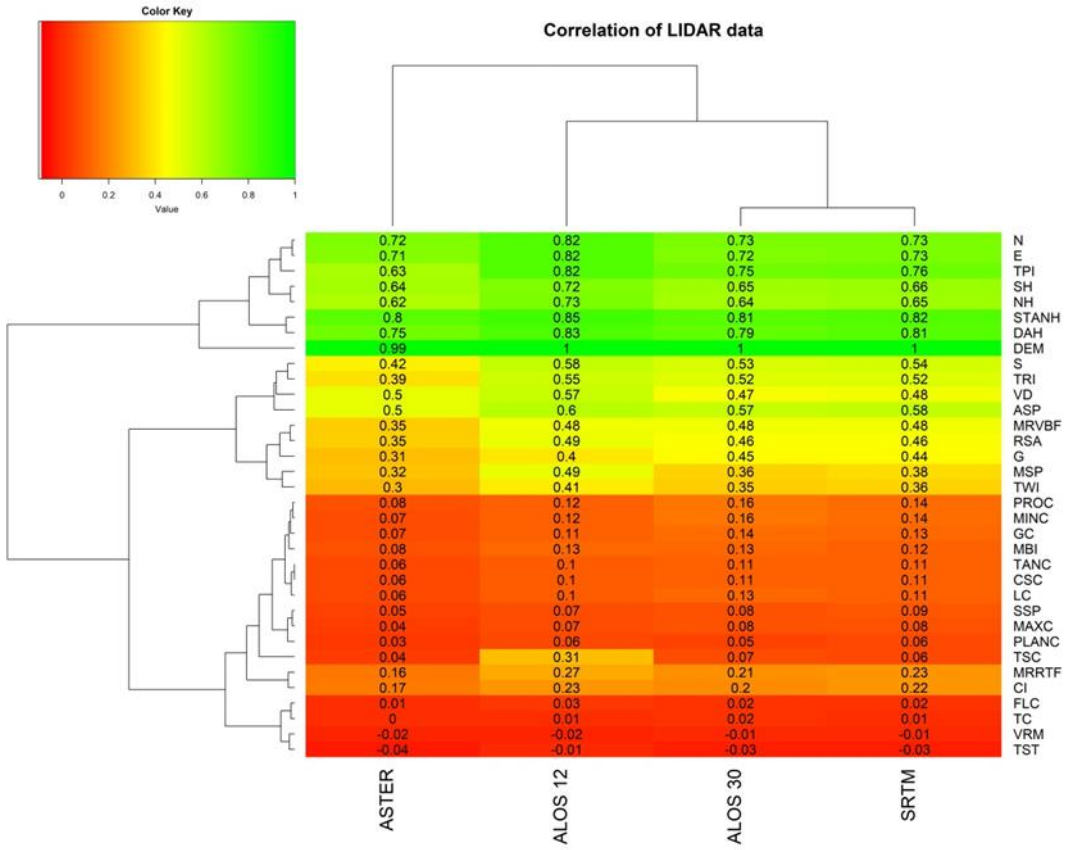


Figura 2. Heatmap ilustrando as correlações de Pearson entre os atributos derivados dos diferentes MDEs com base no LIDAR 2 m.

As curvaturas descrevem a forma geral da vertente em todas as direções (côncava, retilínea ou convexa). Neste sentido, os valores positivos indicam que a superfície é convexa. Este fato influencia, sobretudo, na concentração e dispersão dos fluxos na paisagem, o que incide diretamente sobre a velocidade do fluxo superficial, taxa de erosão/deposição e conteúdo de água no solo (Wilson e Gallant, 2000).

Os atributos "*easterness"* - E, "*northerness"* - N, "*topographic position index*" - TPI, "*slope height"* - SH, *"normalized height" - NH,* "*standardized height"* - SH, *"diurnal anisotropic heating"* - DAH *e "digital elevation model"* - DEM apresentaram correlação positiva moderada a muito forte com valores de $r > 0.6$ em relação a todos os modelos avaliados. Esses valores altos de correlação, principalmente para os "*DEMs*", indicam que estes atributos são pouco sensíveis à mudança do tamanho da célula, não ocorrendo incremento na qualidade das variáveis ao mudar a resolução de referência.

Por outro lado, ao considerar o SRTM como referência, a maioria das correlações entre os atributos são positivas e fortes com valores de $r > 0.6$ (Figura 3). Enquanto na Figura 2 os

atributos "*vector ruggedness measure*" – VRM e "*terrain surface texture*" - TST apresentaram correlação negativa em relação aos demais MDEs, quando considerado o SRTM de referência, estes atributos só tiveram valores negativos em relação ao LIDAR. Vale destacar que o atributo *"digital elevation model"* – DEM apresentou maior valor absoluto de coeficiente (> 0.9), com correlação linear muito forte entre os atributos, tanto para o LIDAR quanto para o SRTM como modelo de referência.

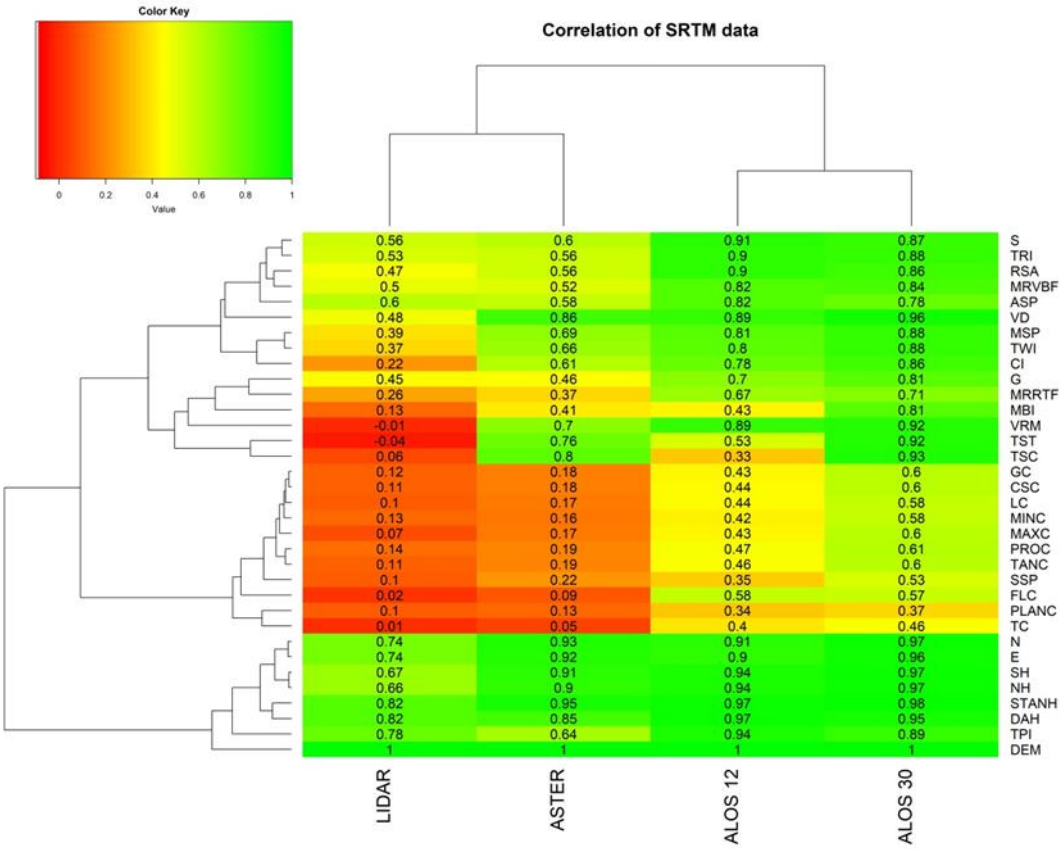


Figura 3. Heatmap ilustrando as correlações de Pearson entre os atributos derivados dos diferentes MDEs com base no SRTM 30 m.

Observa-se ainda que os MDEs apresentam correlação entre si em função da resolução espacial. Os atributos derivados do SRTM apresentaram maior correlação com os atributos derivados do ALOS 30 e ALOS 12, e menor correlação com os atributos derivados do LIDAR 2 m (Figura 3).

Estes valores de correlação entre os MDEs são justificáveis tendo em vista que tanto o SRTM quanto o ALOS 30 apresentam resolução espacial equivalente de 30 m. Além disso, no processo de geração do DEM ALOS 12, os dados sofreram um *downsampling*, ou seja, foram

reamostrados para a resolução 30 m, a fim de corrigir geometricamente os dados, sendo em seguida reamostrados novamente (*upsampling*) para a resolução 12,5 m (JAXA, 2019).

Por outro lado, embora o ASTER GDEM apresente aproximadamente a mesma resolução espacial que o SRTM e o ALOS 30, as correlações foram geralmente inferiores quando comparados a estes MDEs, o que pode ser explicado pelo fato que os dados derivados do ASTER GDEM podem conter erros ou anomalias oriundas do próprio processo de obtenção dos pares estereoscópicos, no qual os dados de elevação são obtidos através de duas visadas diferentes.

3.2 Agrupamento dos atributos do terreno

Diversos estudos relataram que os modelos relacionados à morfometria são sensíveis à resolução do DEM (Walker e Willgoose, 1999; Kumar et al., 2000; Thompson et al., 2001; Wu et al., 2007; Schunemann, 2016). A Tabela 3 apresenta o agrupamento dos atributos do terreno para os três grupos definidos quanto à sensibilidade a resolução (LSR, SR, HSR) em relação aos diferentes MDEs. Observa-se que a maioria dos atributos permaneceu no mesmo grupo de sensibilidade à resolução mesmo com a alteração da resolução espacial de referência dos MDEs.

O mesmo resultado foi observado por Schunemann (2016) ao avaliar a influência da mudança do tamanho das células dos MDEs na qualidade das variáveis morfométricas com resoluções espaciais de 30,0; 20,0; 10,0; 5,0; 1,0; 0,5 e 0,2 metros, na região situada na península Keller, inserida na denominada Antártica Marítima, com altitude máxima correspondente a 380 m. O autor aponta que ao mudar o tamanho da célula de referência, não ocorrem grandes alterações na classificação gerada pelos agrupamentos (cluster = LSR, SR e HSR) das diferentes variáveis estudadas.

Contudo, Kienzle (2004) observou em seu estudo que todas as variáveis do terreno testadas variam significativamente com uma mudança no tamanho da célula do DEM. Deng et al. (2007) também observaram que as variáveis respondem à mudança de resolução de maneiras caracteristicamente diferentes, especialmente quando a resolução é aumentada no intervalo de 5 a 50 m. Sørensen and Seibert (2007) realizaram o mapeamento dos vários índices topográficos calculados para diferentes MDEs e mostraram que ocorre uma clara variação

com a resolução, cujo deslocamento de uma resolução de 5 a 10 m afetou consideravelmente os índices topográficos computados.

Tabela 3. Agrupamento dos atributos conforme sensibilidade à resolução em função da correlação de Pearson

Atributos do terreno	**Coeficiente de correlação de Pearson**				
	LIDAR 2 m	***ALOS 12.5 m***	***ALOS 30 m***	***ASTER***	***SRTM***
Aspect	SR	SR	SR	SR	SR
Convergence index	HSR	SR	SR	SR	SR
Cross sectional curvature	HSR	HSR	HSR	HSR	HSR
Flow line curvature	HSR	HSR	HSR	HSR	HSR
General curvature	HSR	HSR	HSR	HSR	HSR
Longitudinal curvature	HSR	HSR	HSR	HSR	HSR
Maximal curvature	HSR	HSR	HSR	HSR	HSR
Minimal curvature	HSR	HSR	HSR	HSR	HSR
Plan curvature	HSR	HSR	HSR	HSR	HSR
Profile curvature	HSR	HSR	HSR	HSR	HSR
Tangencial curvature	HSR	HSR	HSR	HSR	HSR
Total curvature	HSR	HSR	HSR	HSR	HSR
Diurnal anysotropic heating	LSR	LSR	LSR	LSR	LSR
Easterness	LSR	LSR	LSR	LSR	LSR
Gradient	SR	SR	SR	SR	SR
Mass balance index	HSR	HSR	HSR	HSR	HSR
Digital elevation model	LSR	LSR	LSR	LSR	LSR
Mid-slope position	SR	SR	SR	SR	SR
Multiresolution index of ridge top flatness	HSR	SR	HSR	HSR	SR
Multiresolution index of valley bottom flatness	SR	SR	SR	SR	SR
Normalized height	LSR	LSR	LSR	LSR	LSR
Northerness	LSR	LSR	LSR	LSR	LSR
Real surface área	SR	SR	SR	SR	SR
Slope	SR	SR	SR	SR	SR
Slope height	LSR	LSR	LSR	LSR	LSR
Standardized height	LSR	LSR	LSR	LSR	LSR
Surface specific points	HSR	HSR	HSR	HSR	HSR
Terrain ruggedness index	SR	SR	SR	SR	SR
Terrain surface convexity	HSR	SR	SR	SR	SR
Terrain surface texture	HSR	SR	SR	SR	SR
Topographic position index	LSR	LSR	LSR	SR	LSR
Valley depth	SR	LSR	LSR	LSR	LSR
Vector ruggedness measure	HSR	SR	SR	SR	SR
Topographic wetness index	SR	SR	SR	SR	SR

Legenda: HSR = altamente sensível à resolução; LSR = pouco sensível à resolução; SR = sensível à resolução (Adaptado de Schunemann, 2016).

Dentre o agrupamento LSR, os atributos "*Diurnal anysotropic heating*" - DAH, "*Easterness*" - E, "*Digital elevation model*" - DEM, "*Normalized height*" - NH, "*Northerness*" - N, "*Slope height*" - SH e "*Standardized height*" - STANH não mudaram de grupo de sensibilidade com a mudança do MDE de referência. Estes atributos indicam que mesmo com a mudança de resolução, não há alteração de comportamento; corroborando com os resultados de

correlações positivas moderada a muito forte observados anteriormente. Sendo possível então a utilização do atributo com uma baixa resolução nos modelos de predição e obtendo como vantagem, por exemplo, evitar o excesso de detalhes e facilitar o processamento computacional.

No agrupamento SR os atributos *"Aspect"* - A*, "Gradient"* - G*, "Mid-slope position*" - MSP, *"Multiresolution index of valley bottom flatness*" - MRVBF, *"Real surface área"* - RSA*, "Slope"* - S*, "Terrain ruggedness index"* - TRI *e "Topographic wetness index"* – TWI permaneceram no mesmo grupo ao mudar a resolução.

Para o grupo HSR, os atributos "*Cross sectional curvature*" - CSC, "*Flow line curvature*" - FLC, *"General curvature"* - GC*, "Longitudinal curvature"* - LC*, "Maximal curvature"* - MAXC*, "Minimal curvature"* - MINC*, "Plan curvature"* - PLANC*, "Profile curvature"*- PROC*, "Tangencial curvature*" - TANC, "*Total curvature*" - TC, *"Mass balance index*" - MBI *e "Surface specific points"* - SSP apresentaram alta sensibilidade considerando todos os MDEs. Em geral, estes mesmos atributos apresentaram baixas correlações lineares, como observadas nas Figuras 2 e 3.

De acordo com Schunemann (2016) atributos agrupados como sensíveis a altamente sensíveis podem ser utilizados nos modelos de predição com diferentes resoluções, pois poderá fornecer informações diferentes ao preditor e aumentar o poder de predição. Destaca-se que, sobretudo os atributos associados às curvaturas, apresentam-se como altamente sensíveis à resolução, o que pode estar relacionado ao fato de que as curvaturas são as variáveis que constituem a segunda derivada de um DEM.

Alguns autores apontam que uma melhor opção seria escolher resoluções distintas para variáveis diferentes porque a dependência da escala varia entre os atributos e classes de solos, e combinações de múltiplas resoluções podem fornecem resultados diferenciados e produzir modelos com melhor desempenho (Behrens et al., 2010, 2014, 2018; Miller et al., 2015; Sun et al., 2017).

3.3 Análise de componentes principais

A PCA possibilitou a transformação dos 170 atributos do terreno em um novo conjunto de PC, representado pelo biplot PC1 x PC2, cada qual retendo parte da variabilidade original dos

dados. Observa-se que na Figura 4 representada pelo biplot PC1 x PC2 utilizando o SRTM 30 m como referência, a PC1 retém a maior parte da variância existente entre os atributos do terreno (85.69 %), enquanto a PC2 explica a segunda maior variação dos dados (9.9 %).

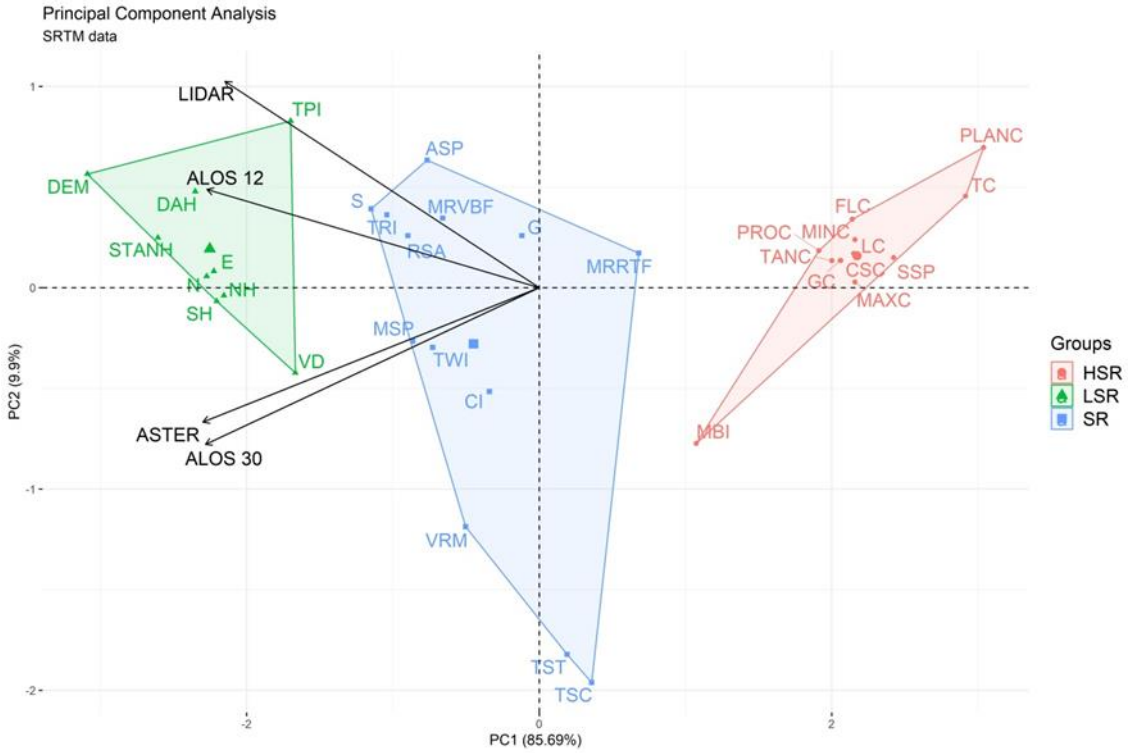


Figura 4. Biplot PC1 x PC2 com SRTM 30 m como referência.

Contudo, o LIDAR 2 m de referência, apresentou uma melhor representatividade dos dados no espaço bidimensional (Figura 5). A PC1 explica 99.14 % da variabilidade dos dados e a PC2 explica 0.51 %. Esta maior percentagem da variabilidade entre as dimensões principais demonstra uma alta concentração de informações nesta componente e indica que os atributos do terreno possuem um grau elevado de independência.

Já as variáveis que não se correlacionam com nenhuma componente ou correlacionadas com as últimas dimensões, são variáveis com baixa contribuição, ou seja, com baixa variabilidade ou redundantes, e podem ser removidas para reduzir e simplificar a análise dos resultados, sem necessariamente uma perda significativa de informação.

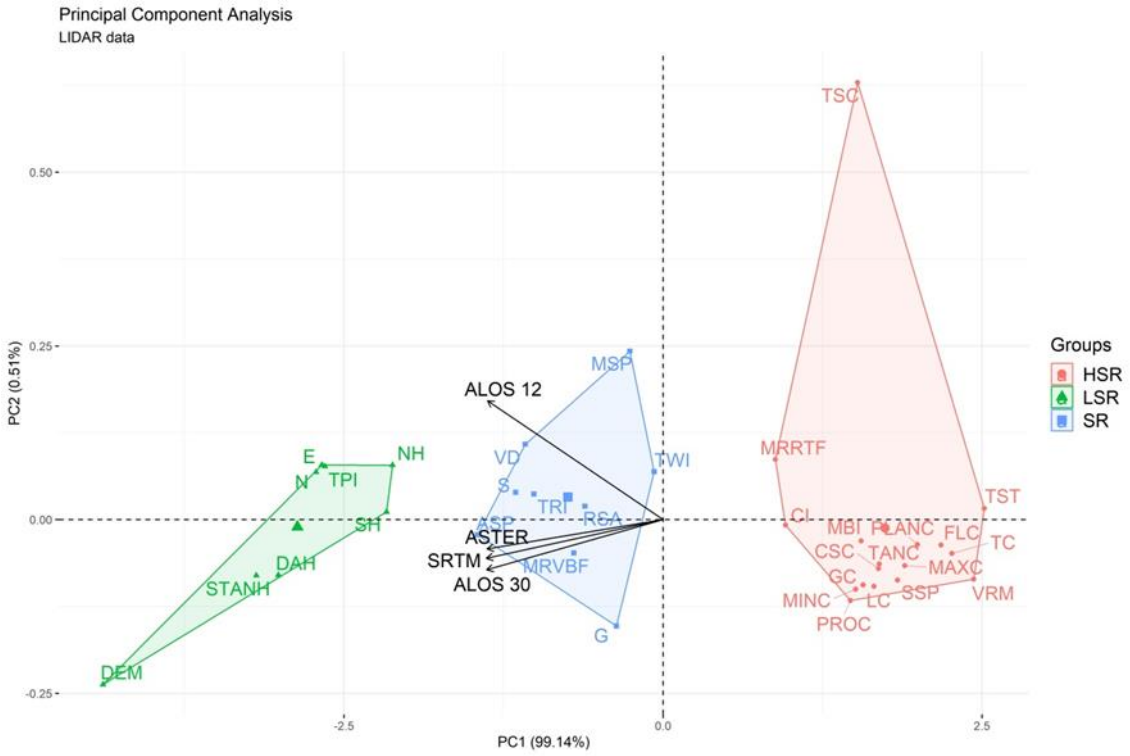


Figura 5. Biplot CP1 x CP2 com LIDAR 2 m como referência.

Desse modo, ao alterar o DEM de referência, ocorre um rearranjo na distribuição dos atributos dentro dos agrupamentos, influenciando, no padrão das formas representadas pelos biplots, dada a variação dos valores de correlação dos atributos. De modo geral, pode-se observar também que os agrupamentos apresentam alta homogeneidade interna, principalmente, o agrupamento dos atributos altamente sensíveis à resolução; e alta heterogeneidade externa entre agrupamentos.

Entretanto, apesar das diferenças entre o tamanho das células dos MDEs de referência, não houve mudança no padrão de contribuição dos agrupamentos por componentes. Em ambos os casos, o agrupamento dos atributos altamente sensíveis à resolução explicam a maior variabilidade de dados na PC1, ou seja, os atributos com alta variância dominam a componente principal. Logo, ao utilizar a PC1 seria possível reduzir o conjunto de dados com pouca perda de informação acerca da variabilidade dos dados.

Os MDEs estão negativamente correlacionados com a PC1 (eixo horizontal), cuja alta correlação entre os atributos é representada pelos ângulos dos vetores < 90º. Na Figura 4, os atributos derivados do ALOS 30 e ASTER apresentaram maior correlação entre si, enquanto na Figura 5 os atributos derivados do SRTM, ALOS 30 e ASTER estão altamente

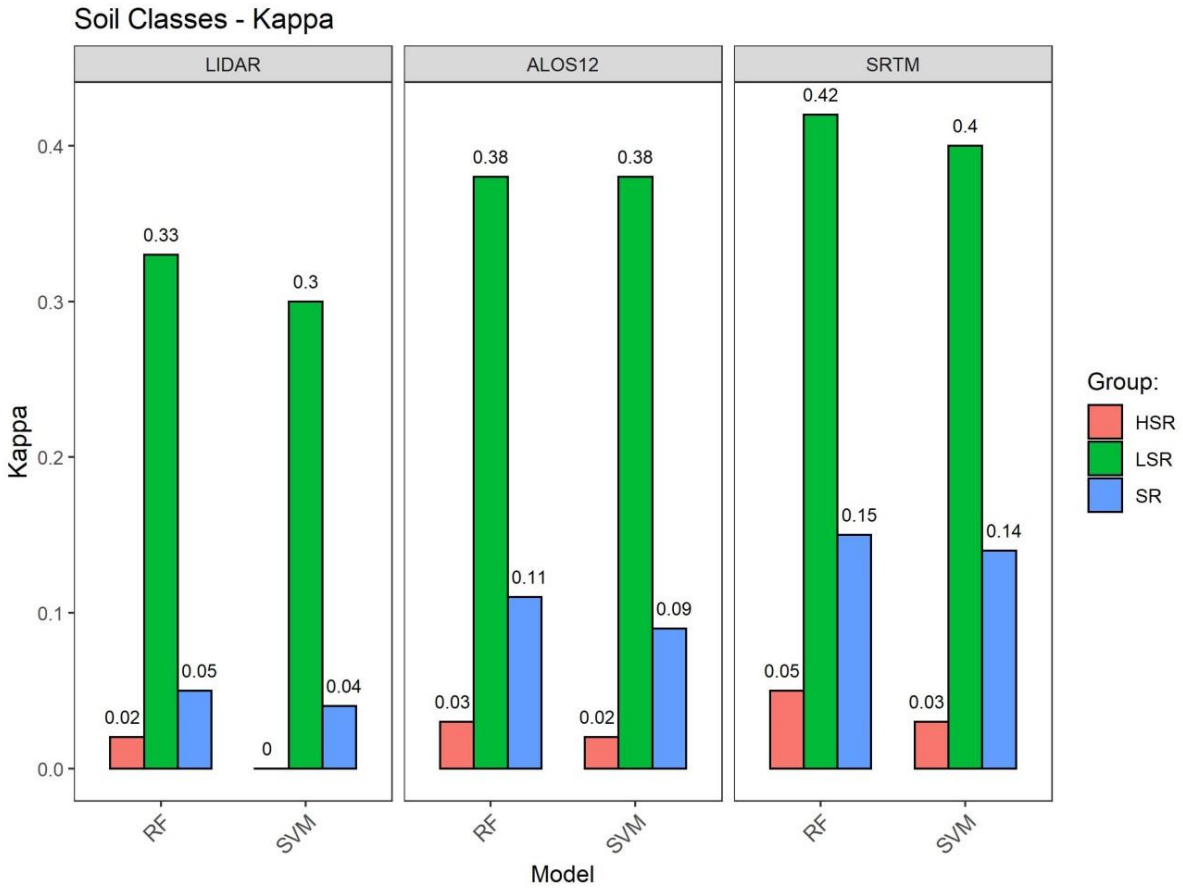
correlacionados entre si, dado o tamanho da célula de 30 m dos MDEs. O comprimento do vetor aproxima-se da variância da variável original, com isso pode-se observar que o LIDAR e ALOS 12, nas Figuras 4 e 5, respectivamente, são independentes dos demais MDEs e apresentam os vetores mais longos, indicando maior variância relativa dos dados. mais longos, indicando maior variância relativa dos dados.

3.4 Predição das classes de solos

Os resultados encontrados no presente estudo reforçam a ideia de que as diferentes resoluções dos atributos do terreno possuem capacidade preditiva distinta para a predição de classes de solos, o que demonstra claramente a importância dos estudos sobre análise do terreno em múltiplas escalas (Behrens et al., 2010). Observa-se que ocorre um aumento na precisão da previsão na medida em que se aplica DEM com resoluções mais baixas. O DEM de resolução mais alta (LIDAR 2 m) apresentou o menor desempenho para ambos os modelos preditivos avaliados (RF e SVM), enquanto o DEM de baixa resolução (SRTM – 30 m) apresentou melhor precisão. Houve pouca diferença das precisões entre a aplicação do DEM de média (ALOS – 12,5 m) e baixa resolução (SRTM – 30 m) (Figura 6).

Com a crescente disponibilidade de MDEs derivados do LiDAR, houve uma suposição de que os atributos do terreno derivados de resoluções espaciais finas produzirão correlações mais fortes com as propriedades do solo (Maynard and Johnson, 2014). No entanto, os resultados encontrados ressalta a premissa de que nem sempre a melhor resolução disponível do DEM apresentará melhores resultados (Thompson et al., 2001; Leempoel et al.,2015). Sørensen and Seibert (2007) relataram também que em alguns casos, um DEM de baixa resolução pode ser mais útil para análises e modelagem da paisagem. Samuel Rosa et al. (2015) observaram que as previsões podem ser degradadas ao usar a versão mais detalhada das covariáveis. Com isso, estes resultados demonstram que a utilização do LIDAR não apresenta vantagens para a predição de classes de solos, principalmente, devido ao alto custo do DEM, alto custo computacional e a pouca ou nenhuma melhoria dos modelos que limitam fortemente sua aplicação no MDS.

Figura 6. Desempenho dos modelos para os diferentes agrupamentos (HSR, LSR e SR) e as diferentes resoluções espaciais.



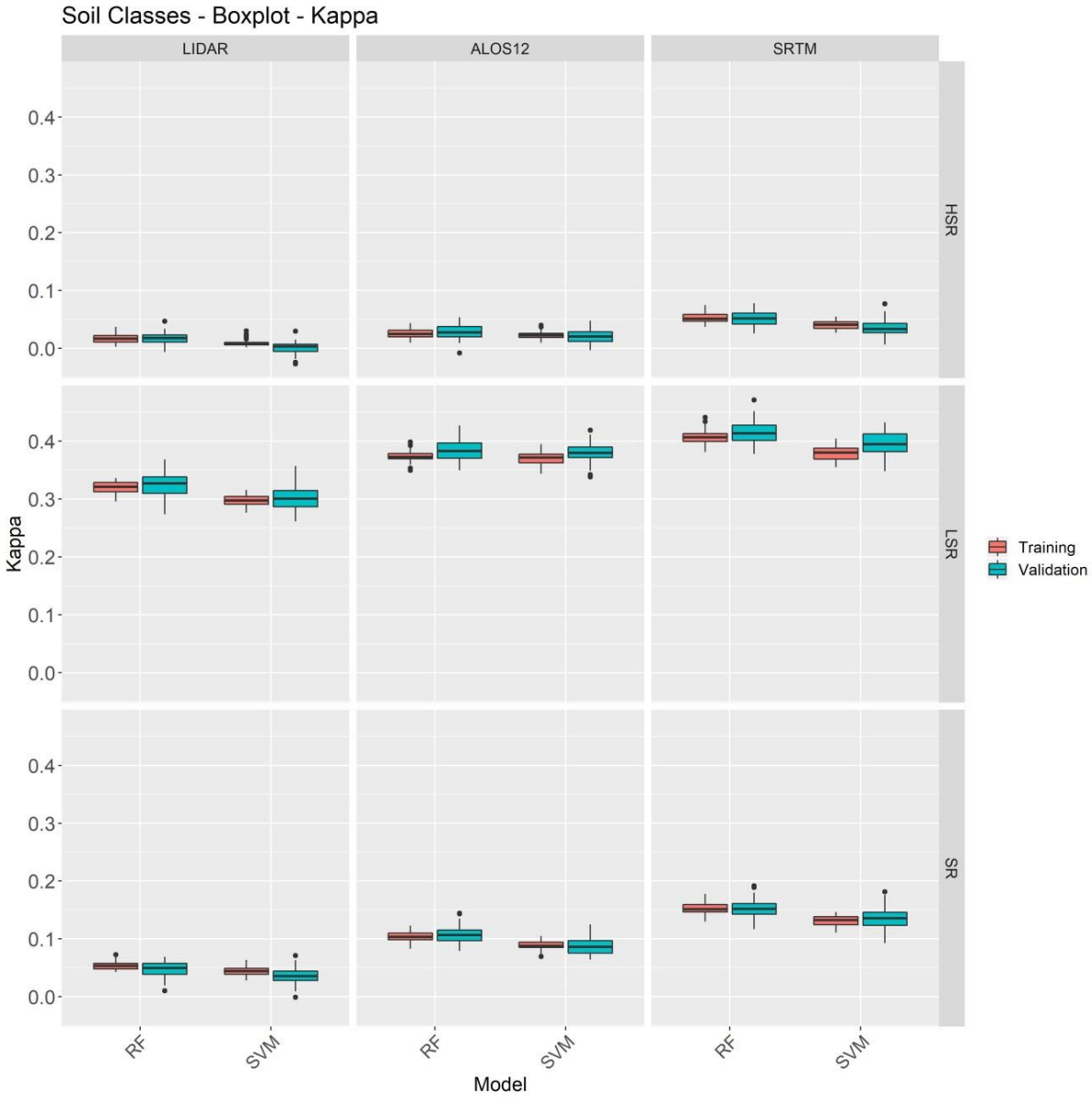
Foi observado também que para as diferentes resoluções abordadas tivemos diferentes contribuições dos atributos do terreno, agrupados com base na sensibilidade à resolução, na precisão dos modelos de predição. Os atributos agrupados como pouco sensíveis à resolução espacial (LSR) foram os que apresentaram melhor capacidade preditiva para ambos os modelos e nas diferentes resoluções espaciais, especialmente para o RF ao utilizar o SRTM, com valor de Kappa igual a 0.42, que é considerado um valor de concordância moderada. Portanto, ao selecionar estes atributos, seria possível diminuir a complexidade e o tempo computacional de um modelo e reduzir a redundância e o armazenamento de dados.

Entretanto, conforme destacado por Schunemann (2016), os atributos altamente sensíveis à resolução (HSR) podem fornecer informações diferenciadas com a mudança da resolução espacial. Contudo, para o presente estudo, estes atributos apresentaram baixa ou quase nula capacidade preditiva nas diferentes resoluções espaciais, o que pode estar associado com o

excesso de detalhes que podem gerar muito ruído e, consequentemente, a perda do poder explicativo (Cavazzi et al. 2013).

A Figura 7 mostra que os modelos são precisos e apresentam baixo desvio padrão entre os resultados. Tanto para os dados de treinamento quanto para validação, os modelos não apresentaram diferença significativa de desempenho, ou seja, os modelos não apresentaram *overfitting*, que ocorre quando um modelo se ajusta muito bem ao conjunto de dados anteriormente observado, mas se mostra ineficaz para prever novos resultados.

Fig. 7. Boxplot apresentando o desempenho dos dados de treinamento e validação para os modelos nos diferentes agrupamentos e resoluções espaciais.



Neste caso, pode-se afirmar que as limitações das previsões não estão associadas aos modelos utilizados, mas a seleção de um conjunto ideal de atributos para a predição das classes de solos na área de estudo. Contudo, a questão de uma resolução ideal ainda precisa ser mais fortemente discutida, uma vez que depende das finalidades as quais os atributos possam ser utilizados e das propriedades da paisagem de interesse (Hengl, 2006; Deng et al., 2007; Sørensen and Seibert; 2007; Maynard and Johnson, 2014).

4. Conclusões

O estudo destaca a importância da análise do terreno em múltiplas escalas, pois atributos do terreno em diferentes resoluções possuem capacidade preditiva distinta para a predição de classes de solos. A utilização dos atributos LSR derivados do SRTM DEM apresentaram melhor precisão para os algoritmos utilizados e possuem como vantagens o baixo custo e à facilidade do processamento computacional. Contudo, os resultados apresentados foram realizados em uma área relativamente pequena e de relevo complexo, sendo necessárias pesquisas adicionais em áreas com diferentes contextos ambientais, para obter métodos alternativos que integrem atributos do terreno em diferentes resoluções espaciais nos modelos de MDS com boa precisão e baixo custo. Cabe destacar que para selecionar um DEM ideal é necessário, sobretudo, considerar as características da área de estudo, os objetivos e escala de trabalho, além dos recursos financeiros, humanos e de infraestrutura disponíveis para o projeto.

5. Bibliografia

Ballabio, C., Fava, F., Rosenmund, A., 2012. A plant ecology approach to digital soil mapping, improving the prediction of soil organic carbon content in alpine grasslands. Geoderma. 187-188:102-116. https://doi.org/10.1016/j.geoderma.2012.04.002.

Behrens, T., Zhu, A.X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma. 155, 175-185. https://doi.org/10.1016/j.geoderma.2009.07.010

Behrens, T., Schmidt, K., Ramirez-Lopez, L., Gallant, J., Zhu, A.-X., Scholten T., 2014. Hyper-scale digital soil mapping and soil formation analysis. Geoderma. 213, 578-588. https://doi.org/10.1016/j.geoderma.2013.07.031

Behrens, T., Schmidt, K., MacMillan, R.A., Viscarra Rossel, R.A., 2018. Multiscale contextual spatial modelling with the Gaussian scale space. Geoderma. 310, 128-137. https://doi.org/10.1016/j.geoderma.2017.09.015

Bivand, R., Keitt, T., Rowlingson, B., 2018. rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 1.3-6. Available at https://CRAN.Rproject.org/package=rgdal (accessed 15 Jan 2019)

Bivand, R., Krug, R., Neteler, M., Jeworutzki, S., 2019. rgrass7: Interface Between GRASS 7 Geographical Information System and R. R package version 0.2-1. Available at https://cran.r-project.org/web/packages/rgrass7/index.html (accessed 15 Jan 2019)

Brenning, A., 2008. Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models. In: Böhner, J., Blaschke, T., Montanarella, L. (eds.), SAGA – Seconds Out (= Hamburger Beiträge zur
Physischen Geographie und Landschaftsökologie) 19, 23–32.

Breiman, L., 2001. Random Forests. Machine Learning. 45:5-32.

Breiman, L., 2002. Manual on setting up, using, and understanding random forests v3.1. Statistics Department University of California Berkeley, CA, USA.

Breiman, L., Cutler, A., Liaw, A., Wiener, M., 2018. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.6-14.

Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., Fealy, R., 2013. Are fine resolution digital elevation models always the best choice in digital soil mapping? Geoderma. 195-196, 111-121. https://doi.org/10.1016/j.geoderma.2012.11.020

Cherkassky, V., Ma, Y., 2004. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Netw. 17:113-126. https://doi.org/10.1016/S0893-6080(03)00169-2

Congalton, R., Green, K., 1998. Assessing the accuracy of remotely sensed data: principles and practices. Lewis Publishers, Boca Raton.

Cortes, C., Vapnik, V., 1995. Support-vector networks, Mach. Learn. 20:273-297.

Cruz, C.D., Regazzi, A.J., Carneiro, P.C.S., 2004. Modelos biométricos aplicados ao melhoramento genético, ed. UFV, Viçosa.

Deng, Y., Wilson, J.P., Bauer, B.O., 2007. DEM resolution dependencies of terrain attributes across a landscape. International Journal of Geographical Information Science. 21, 187-213. https://doi.org/10.1080/13658810600894364

Drăguţ, L., Eisank, C., Strasser, T., 2011. Local variance for multi-scale analysis in geomorphometry. Geomorphology. 130, 162-172. https://doi.org/10.1016/j.geomorph.2011.03.011

EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária. 2018. Sistema Brasileiro de Classificação de Solos. Rio de Janeiro: Embrapa Solos.

Fávero, L.P., Belfiore, P.P., Silva, F.L., Chan, B.L., 2009. Análise de dados: modelagem multivariada para tomada de decisões. Campus, Rio de Janeiro.

Garcia, L.C., 2007. Fenologia de espécies da canga em Barão de Cocais, Quadrilátero Ferrífero de Minas Gerais. Master degree dissertation - Universidade Federal de Minas Gerais, Belo Horizonte, 123p.

Hair Junior, J.F., Black, W.C., Babin, B.J., Anderson, R.E., 1998. Multivariate data analysis. Prentice Hall, Upper Saddle River.

Han, J., Kamber, M., Pei, J., 2011. Data mining: concepts and techniques, 3ed. San Francisco: Morgan Kaufmann Publishers.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer, New York, 745p.

Hengl, T., 2006. Finding the right pixel size. Computers and Geosciences. 32, 1283 1298. https://doi.org/10.1016/j.cageo.2005.11.008

Husson, F., Josse, J., Le, S., Mazet, J. 2018. FactoMineR: Multivariate Exploratory Data Analysis and Data Mining. R package version 1.41. https://cran.r-project.org/web/packages/FactoMineR/index.html (accessed 07 Jun 2019)

IBGE. 2001. Base cartográfica vetorial contínua do Brasil ao milionésimo - BCIM. IBGE, Rio de Janeiro.

JAXA - Japan Aerospace Exploration Agency. About ALOS – PALSAR. 2019. Available at https://www.eorc.jaxa.jp/ALOS/en/about/palsar.htm (accessed 21 Mai 2019)

Kassambara, A., Mundt, F. 2017. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5. Available at https://cran.r-project.org/web/packages/factoextra/index.html (accessed 07 Jun 2019)

Kempen, B., Brus, D.J., Stoorvogel, J.J., 2011. Three-dimensional mapping of soil organic matter content using soil type–specific depth functions. Geoderma. 62, 107-123. https://doi.org/10.1016/j.geoderma.2011.01.010

Kerry, R., Oliver, M.A., 2011. Soil geomorphology: identifying relations between the scale of spatial variation and soil processes using the variogram. Geomorphology. 130, 40-54. https://doi.org/10.1016/j.geomorph.2010.10.002

Kienzle, S., 2004. The effect of DEM raster resolution on first order, second order and compound terrain derivatives. Transactions in GIS. 8, 83–112. https://doi.org/10.1111/j.1467-9671.2004.00169.x

Kumar, P., Verdin, K.L., Greenlee, S.K., 2000. Basin level statistical properties of topographic index for North America. Advances in Water Resources. 23, 571-578. http://dx.doi.org/10.1016/S0309-1708(99)00049-4

Kuhn, M., Johnson, K., 2013. Applied predictive modeling. New York, NY: Springer.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T., 2020. caret: Classification and Regression Training. R package version 6.0-85.

Leempoel, K., Parisod, C., Geiser, C., Daprà, L., Vittoz, P., Joost, S., 2015. Very high-resolution digital elevation models: are multi-scale derived variables ecologically relevant? Methods in Ecology and Evolution. 6, 1373–1383. https://doi.org/10.1111/2041-210X.12427

Li, X., Zhang, Y., Jin, X., He, Q., & Zhang, X., 2017. Comparison of digital elevation models and relevant derived attributes. Journal of Applied Remote Sensing. 11(4), 046027. https://doi.org/10.1117/1.JRS.11.046027

Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R News, 2:18-22.

Ließ, M., Schmidt, J., Glaser, B., 2016. Improving the spatial prediction of soil organic carbon stocks in a complex Tropical Mountain landscape by methodological specifications in machine learning approaches. PLoS One. 11:1-22. https://doi.org/10.1371/journal.pone.0153673

Mattivi, P., Franci, F., Lambertini, A., Bitelli, G., 2019. TWI computation: a comparison of different open source GISs. Open Geospatial Data, Software and Standards, 4:6. https://doi.org/10.1186/s40965-019-0066-y

Maynard, J., Johnson, M., 2014. Scale-dependency of LiDAR derived terrain attributes in quantitative soil-landscape modeling: effects of grid resolution vs. neighborhood extent. Geoderma. 230, 29-40. https://doi.org/10.1016/j.geoderma.2014.03.021

McBratney, A.B., Mendonça-Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma. 117, 3-52. https://doi.org/10.1016/S0016-7061(03)00223-4

Mendonça-Santos, M.L., Santos, H.G., Dart, R.G., Pares, J.G., 2007. Modelagem e Mapeamento Digital de Estoque de Carbono Orgânico na Camada Superficial dos Solos (0-10 cm) do Estado do Rio de Janeiro. Embrapa, Boletim de pesquisa e desenvolvimento.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C., Lin, C.C., 2019. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-3.

Miller, B.A., Koszinski, S., Wehrhan, M., Sommer, M., 2015. Impact of multi-scale predictor selection for modeling soil properties. Geoderma. 239, 97-106. https://doi.org/10.1016/j.geoderma.2014.09.018

Miot, H.A., 2018. Análise de correlação em estudos clínicos e experimentais. Jornal Vascular Brasileiro. 17, 275-279. https://doi.org/10.1590/1677-5449.174118

Moore, I.D., Gessler, P.E., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Science Society American Journal. 57, 443-452. https://doi.org/10.2136/sssaj1993.572NPb

Naghibi, S.A., Ahmadi, K., Daneshi, A., 2017. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resources Management. 31:2761–2775. https://doi.org/10.1007/s11269-017-1660-3

NASA - National Aeronautics and Space Administration. 2019. Shuttle Radar Topography Mission. Available at https://www2.jpl.nasa.gov/srtm/ (accessed 31 Mai. 2019)

Neuman, G., Silveira, C.T., Sampaio, T.V.M., 2018. Análise da influência da escala na obtenção dos atributos topográficos derivados de MDE. Raega - O Espaço Geográfico em Análise. 43, 179-199. https://doi.org/10.5380/raega

Oksanen, J., Sarjakoski, T., 2005. Error propagation of DEM-based surface derivatives. Computers & Geosciences. 31:1015-1027. https://doi.org/10.1016/j.cageo.2005.02.014

Pearson, K., 1895. Contributions to the mathematical theory of evolution, II: skew variation. Philosophical Transactions of the Royal Society of London, A. 186, 343-414.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at https://www.R-project.org/

Robert, J.H., 2019. raster: Geographic Data Analysis and Modeling. R package version 2.8-19. Available at http://CRAN.R-project.org/package=raster (accessed 15 Jan 2019)

Rodrigues, T.L., Debisasi, P., Souza, R.F. de., 2010. Avaliação da adequação de produtos ASTERGDEM no auxilio ao mapeamento sistemático Brasileiro. III SIMGEO (Simpósio Brasileiro de Ciências Geodésicas e Tecnologia da Geoinformação). Available at https://www3.ufpe.br/cgtg/SIMGEOIII/IIISIMGEO_CD/artigos/CartografiaeSIG/Cartografia/A_17.pdf (accessed 31 Mai. 2019)

Samuel-Rosa, A., Dalmolin, R.S.D., Miguel, P., 2013. Building predictive models of soil particle-size distribution. Revista Brasileira de Ciência do Solo. 37, 422-430. http://dx.doi.org/10.1590/S0100-06832013000200013

Samuel-Rosa, A., Heuvelink, G., Vasques, G., e Anjos, L., 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma. 243, 214-227. https://doi.org/10.1016/j.geoderma.2014.12.017

Santos, D.R., 2006. Automação da resseção espacial de imagens com uso de hipóteses de rodovias como apoio de campo derivadas do sistema de varredura laser. Doctoral thesis – Universidade Federal do Paraná, Curitiba.

Schloerke, B., Crowley, J., Cook, D., 2018. GGally: Extension to 'ggplot2'. R package version 1.4.0. Available at https://cran.r-project.org/web/packages/GGally (accessed 15 Jan 2019)

Schunemann, A.L., 2016. Geotecnologias para mapeamento digital na Antártica marítima. 2016. Doctoral thesis - Universidade Federal de Viçosa, Viçosa.

Sirtoli, A.E., Silveira, C.T., Montovani, L.E., Silva, C.R., Ribeiro, S.R.A., OKA-FIORI, C., 2008. Atributos topográficos secundários no mapeamento de pedoformas. Geociências. 21:63-77.

Smith, M.P., Zhu, A-X., Burt, J.E., Stiles, C., 2006. The effects of DEM resolution and neighborhood size on digital soil survey. Geoderma. 137, 58-69. https://doi.org/10.1016/j.geoderma.2006.07.002

Sørensen, R., Seibert, J., 2007. Effects of DEM resolution on the calculation of topographical indices: TWI and its componentes. J. Hydrol., 347, 79-89. https://doi.org/10.1016/J.JHYDROL.2007.09.001

Sun, X.L., Wang, H.L., Zhao, Y.G., Zhang, C., Zhang, G.L., 2017. Digital soil mapping based on wavelet decomposed components of environmental covariates. Geoderma. 303, 118–132. https://doi.org/10.1016/j.geoderma.2017.05.017

Tan, P.N., Steinbach, M., Kumar, V., 2009. Introdução ao Data Mining. Rio de Janeiro: Editora Ciência Moderna Ltda, 896 p.

Teske, R., Giasson, E., Bagatini, T., 2014. Comparação do uso de modelos digitais de elevação em mapeamento digital de solos em Dois Irmãos, RS, Brasil. Revista Brasileira de Ciência do Solo. 38, 1367-1376. http://dx.doi.org/10.1590/S0100-06832014000500002

Thompson, J.A., Bell, J.C., Butler, C.A., 2001. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modelling. Geoderma. 100, 67-89. https://doi.org/10.1016/S0016-7061(00)00081-1

Universidade Federal de Viçosa (UFV), Fundação Centro Tecnológico de Minas Gerais (CETEC-MG), Universidade Federal de Lavras (UFLA), Fundação Estadual do Meio Ambiente (FEAM). 2010. Mapa de solos do Estado de Minas Gerais: legenda expandida. Fundação Estadual do Meio Ambiente, Belo Horizonte. Available at http://www.feam.br/noticias/1/1355-mapa-de-solos (accessed 05 Jun. 2019)

Varajão, C.A.C., Salgado, A.A.R., Varajão, A.F.D.C., Braucher, R., Colin, F., Nalini Júnior, H.A., 2009. Estudo da evolução da paisagem do Quadrilátero Ferrífero (Minas Gerais, Brasil) por meio da mensuração das taxas de erosão (10be) e da pedogênese. Revista Brasileira de Ciência Solo. 33, 1409-1425. http://dx.doi.org/10.1590/S0100-06832009000500032

Viscarra Rossel, R.A., 2011. Fine-resolution multiscale mapping of clay minerals in Australian soils measured with near infrared spectra. J. Geophys. Res. F: Earth Surf., 116, 1-15. https://doi.org/10.1029/2011JF001977

Walker, J.P., Willgoose, G.R., 1999. On the effect of digital elevation model accuracy on hydrology and geomorphology. Water Resources Research. 7, 2259-2268.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W.H.A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., Venables, B. 2019. gplots: Various R Programming Tools for Plotting Data. Available at https://cran.r-project.org/web/packages/gplots/index.html (accessed 15 Jan 2019)

Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C., Woo, K., 2018. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 3.1.0. Available at https://cran.r-project.org/web/packages/ggplot2 (accessed 15 Jan 2019)

Wickham, H., François, R., Henry, L., Muller, K., 2019. dplyr: A Grammar of Data Manipulation. R package version 0.8.3. Available at https://cran.r-project.org/web/packages/dplyr/index.html (accessed 15 Jan 2019)

Wilson, J.P., Gallant, J.C., 2000. Terrain Analysis: Principles and Applications. Jonh Wiley & Sons, Inc., Nova Jersey. ISBN 0-471-32188-5.

Wu, S., Li, J., Huang, G.H., 2007. Modeling the effects of elevation data resolution on the performance of topography-based watershed runoff simulation. Environmental Modelling and Software. 22, 1250-1260. http://dx.doi.org/10.1016/j.envsoft.2006.08.001

Zhu, A-X., 2008. Spatial scale and neighborhood size in spatial data processing for modeling the natural environment. Mount, N.J., Harvey, G.L., Aplin, P., Priestnall, G. (eds.), Representing, Modeling and Visualizing the Natural Environment: Innovations in GIS 13, CRC Press, Florida. https://doi.org/10.3354/meps11378

Artigo 3

AVALIAÇÃO DO MÉTODO HIPERCUBO LATINO CONDICIONADO (cLHS) PARA SELEÇÃO DE LOCAIS DE AMOSTRAGEM

1. Introdução

O esquema de amostragem é uma das primeiras etapas de delineamento do mapeamento digital de solos (MDS) e tem sido cada vez mais abordado pelos diversos autores (Bagatini et al., 2015; Szatmári et al., 2015; Teske et al., 2015; Stumpf et al., 2016), uma vez que determinam em grande parte a precisão e influencia diretamente o custo e a eficiência da pesquisa (Hengl et al. 2003; Brungard e Boettinger, 2010).

Contudo, as estratégias de amostragem já adotadas em levantamentos convencionais de solos podem ser ineficientes para o MDS. Haja vista que são consideradas como direcionadas, não aleatórias e não fornecem estimativas estatísticas representativas da variabilidade ambiental da área de estudo, sendo apontadas como um método subjetivo por parte dos pedólogos, para sustentar os modelos mentais de distribuição dos solos na paisagem (Hengl, 2003; Brungard e Boettinger, 2010).

Diferentes métodos de amostragem têm sido aplicados, como exemplo, a amostragem aleatória (Teske et al., 2014; Silva et al., 2017); amostragem aleatória estratificada (Grimm et al., 2008; Machado et al.; 2018); amostragem em cluster (Mora-Vallejo et al., 2008); amostragem intencional ou direcionada (Arruda et al., 2013; Wang et al., 2016) e o método do hipercubo latino condicionado (cLHS) (Minasny e McBratney, 2006; Malone et al., 2019). Nos últimos anos, o cLHS tem sido apontado como uma estratégia robusta de amostragem para seleção de amostras representativas, o qual utiliza como principal condicionante as variáveis ambientais e suas distribuições multivariadas (Minasny e McBratney, 2006).

Entretanto, a seleção do conjunto de variáveis ambientais a serem inseridas no cLHS ainda é pouco discutida nos estudos que utilizam este método. Para a amostragem ser considerada eficiente é necessária a utilização de todas as variáveis disponíveis associadas à função SCORPAN, mas que aplicam alguma medida de redução de dimensão previamente (Ließ, 2020), haja vista que a qualidade do MDS estará, dentre outros fatores, diretamente

relacionado com a qualidade das variáveis utilizadas para predição de classes e/ou propriedades dos solos. Além disso, a amostragem deve contemplar a variabilidade dos solos existentes na área de estudo e apresentar boa operacionalidade, que consiste na acessibilidade dos pontos em campo, com menor demanda de custo e tempo; bem como considerar a possibilidade de redução do número de amostras a serem coletadas a partir do conhecimento prévio dos dados legados de solo (Stumpf et al., 2016).

Desta forma, o objetivo do presente estudo foi avaliar a utilização do método cLHS para seleção de locais de amostragem a serem utilizados no mapeamento digital de solos e analisar o desempenho operacional com base nas suas potencialidades e restrições.

2. Material e métodos

2.1 Área de estudo

A área de estudo está inserida na Bacia Hidrográfica do Rio Doce, no município de Mariana - Minas Gerais, com uma área total 144 km² (Figura 1). Compreende os distritos municipais de Padre Viegas, Cachoeira do Brumado, Barro Branco, Mainart, Vargem e Serra.

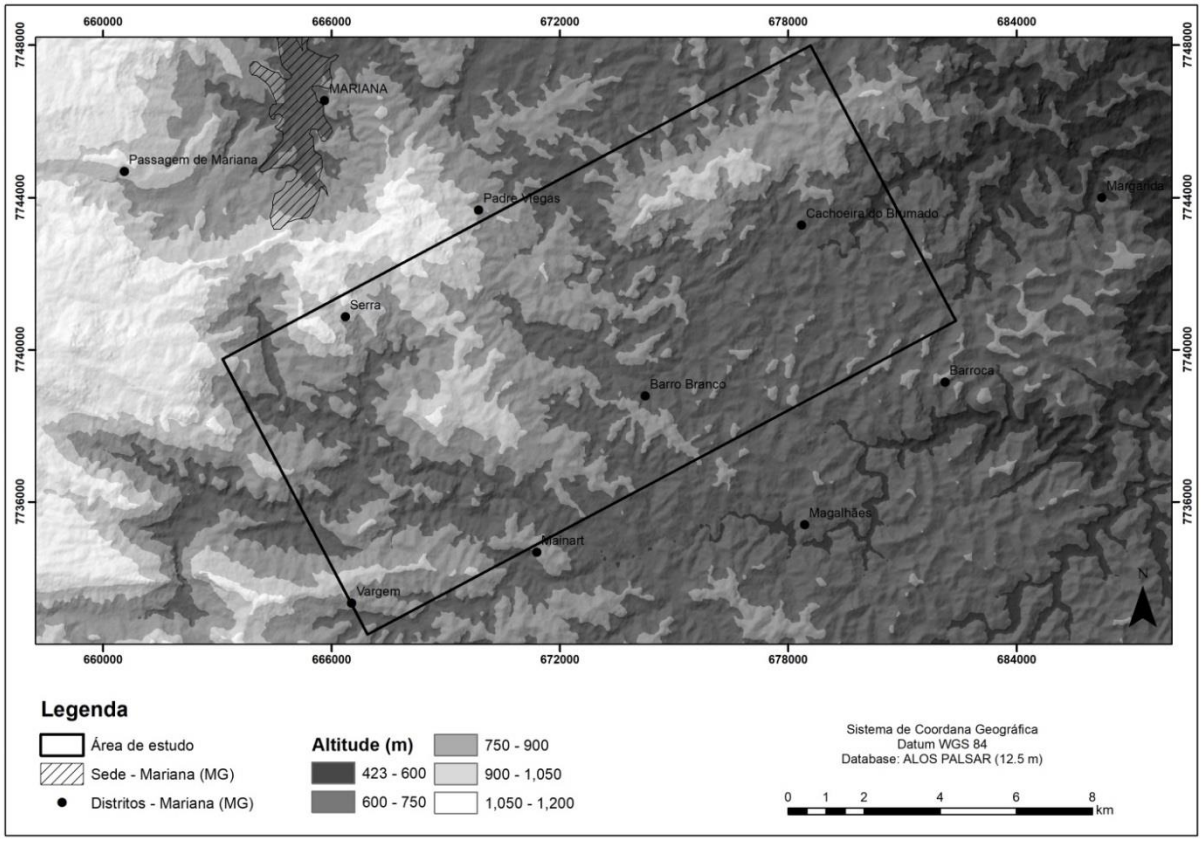


Figura 1. Mapa de localização da área de estudo.

A área destaca-se por apresentar uma heterogeneidade geológica composta por litotipos dos Supergrupos Rio das Velhas e Minas; Grupos Nova Lima, Maquiné, Piracicaba, Sabará e Itacolomi; Complexos Santa Barbará, Mantiqueira e Monsenhor Isidro e; Suíte Alto Maranhão (Codemig, 2014). A vegetação predominante são os Campos Rupestres em relevo complexo marcado por cristas, escarpas, encostas do Quadrilátero Ferrífero e Planaltos dissecados do Rio Piracicaba e do alto Rio Doce (IBGE, 2001; Garcia, 2007).

2.2 Base de dados

Foram extraídas 1.700 amostras de classes de solos no segundo nível categórico do Sistema Brasileiro de Classificação de Solos (Embrapa, 2018), oriundas do mapa de solos legado do município de Mariana (1:50.000) para treinamento dos modelos preditivos. Foi utilizado um conjunto inicial de variáveis associadas aos fatores de formação do solo (Tabela 1) proposto por Jenny (1941).

Tabela 1. Conjunto inicial das covariáveis utilizadas

Fatores de formação do solo	Covariáveis	Fonte
Relevo	Aspect, convergence index, cross sectional curvature, flow line curvature, general curvature, longitudinal curvature, maximal curvature, minimal curvature, plan curvature, profile curvature, tangencial curvature, total curvature, classification curvature, diurnal anisotropic heating, easterness, geomorph, gradient, mass balance index, digital elevation model, mid-slope position, multiresolution index of ridge top flatness, multiresolution index of valley bottom flatness, normalized height, northerness, real surface área, slope, slope height, standardized height, surface specific points, terrain ruggedness index, terrain surface convexity, terrain surface texture, topographic position index, valley depth, vector ruggedness measure, topographic wetness index	LIDAR 2m ALOS PALSAR 12,5 m ALOS PALSAR 30 m ASTER GDEM ≅ 30 m SRTM 30 m
Material de origem	Mapa Geológico do Município de Mariana (2011)	Base cartográfica: IBGE (2001), CPRM (1993)
Organismos	Normalized difference vegetation index - NDVI, soil-adjusted vegetation index - SAVI, Clay minerals e Iron oxides	LANDSAT 8 SENTINEL 2

Os atributos do terreno foram extraídos a partir de três resoluções espaciais derivadas de quatro diferentes MDEs: (1) Levantamento altimétrico a laser do LIDAR com resolução de 2

m; (2) ALOS PALSAR com resolução de 12.5 e 30 m; (3) SRTM com resolução de 30 m; e (4) ASTER GDEM com resolução de ≅ 30 m. A obtenção dos atributos do terreno foi realizada através do programa computacional R (R CORE TEAM, 2017), utilizando os pacotes *"rsaga"* (Brenning, 2008), *"raster"* (Robert, 2019) e *"rgrass7"* (Bivand et al., 2019).

As informações referentes ao material de origem foram extraídas do Mapa Geológico do Município de Marina na escala 1.50.000. As variáveis associadas ao fator organismos foram obtidas através de dados espectrais dos satélites Landsat 8 e Sentinel 2 para os meses de janeiro (úmido) e setembro (seco). Dados climáticos e aerogeofísicos, ainda que disponíveis para a área de estudo, não foram considerados, uma vez que não possuíam escalas compatíveis com os objetivos do presente estudo.

2.3 Seleção das variáveis

Para gerar as variáveis a serem utilizadas como condicionantes no método cLHS, primeiramente foram removidas do conjunto inicial de dados as variáveis altamente correlacionadas (correlação linear > 95%). Posteriormente, foi realizado o treinamento 100 vezes de cada modelo de aprendizado supervisionado comumente utilizado para predição de solos (Gradient boosting machine - GBM, Random Forest - RF, Support Vector Machine - SVM, K Nearest Neighbor - kNN e C5.0), considerando os diferentes MDEs (LIDAR 2m; ALOS PALSAR 12,5 m; ALOS PALSAR 30 m; ASTER GDEM ≅ 30 m e SRTM 30 m) a fim de obter uma variância nos dados. A importância do conjunto de variáveis foi obtida a cada vez que os modelos eram treinados. Com isso, foi realizada a soma total dos valores das importâncias, dentre as quais foram selecionadas as 10 melhores a serem utilizadas no cLHS: 1 - MDE, 2 - Geologia, 3 - Iron oxide (Sentinel/Setembro), 4 - Clay minerals (Landsat/Janeiro), 5 - Terrain surface texture, 6 - Terrain surface convexity, 7 - Standardized height, 8 – Clay minerals (Sentinel/Setembro), 9 - SAVI (Sentinel/Setembro) e 10 - Slope height.

2.4 Esquema de amostragem cLHS

O método cLHS consiste em: dado K variáveis com X1,...,Xk, sendo a faixa de variação de cada uma, X é dividido em *n* prováveis intervalos iguais (estratos); para cada variável uma amostra aleatória é tomada para cada estrato. As amostras obtidas para cada variável por estrato são confrontadas umas com as outras, de forma aleatória ou seguindo alguma regra

previamente especificada. A regra neste caso é de que a amostragem possa refletir a mesma representação dos estratos para todas as variáveis consideradas. Ao final, tem-se um número de amostras que cobrem os *n* estratos para todas as variáveis analisadas (Carvalho Junior et al., 2014). Desta forma, o cLHS foi aplicado utilizando 500.000 iterações para a determinação de 200 pontos amostrais, considerados a princípio suficientes para captar toda a variabilidade espacial de solos da área de estudo. Foram incorporadas como restrições ao cLHS a distância euclidiana das estradas e a exclusão das áreas urbanas e áreas de mineração.

2.5 cLHS modificado

Tendo em vista a necessidade de otimizar a amostragem em campo em relação a demanda de custo e tempo, uma vez que a área de estudo localiza-se em uma região de relevo complexo com limitações de vias de acesso; foi aplicado no presente estudo o método cLHS modificado, a fim de fornecer maior possibilidade e flexibilidade para coleta das amostras.

Para tanto, foram definidos cinco pontos alternativos para cada ponto original do conjunto cLHS (Figura 2) utilizando o método k-Means, o qual consiste em um método de agrupamento baseados em centroides multivariados (no caso, os pontos originais cLHS) e visa minimizar a distância quadrática média entre objetos e os valores mais próximos do centróide (pontos alternativos) (Brus et al., 2006). A variação multivariada dentro do cluster é otimizada para ser o menor possível para cada cluster, agrupando valores de atributos muito semelhantes para cada cluster e pequenas distâncias espaciais entre eles para conjuntos de dados estruturados espacialmente (Burrough et al., 2000).

Os pontos alternativos seriam utilizados caso os pontos originais determinados pelo cLHS fossem inacessíveis por motivos como impedimento pelas condições de relevo, propriedades privadas sem autorização de acesso, dentre outros. Foi permitida uma 'tolerância radial' para cada um dos pontos, onde a equipe de campo poderia amostrar se a amostragem naquele local exato não fosse possível, desde que o ponto estivesse presente dentro de um mesmo domínio de classe de solo, de acordo com os conhecimentos dos pedólogos (Kidd et al., 2015).

Todo processo computacional foi realizado no programa R (R CORE TEAM, 2019), utilizando os pacotes "*rgdal*" (Bivand et al., 2018), "*raster*" (Robert, 2019), "*dplyr*" (Wickham et al., 2019), "*caret"* (Kuhn et al., 2020), *"gbm"* (Greenwell et al., 2019),

"*randomForest*" (Breiman e Cutler, 2018), *"clhs"* (Roudier et al., 2019), "*ggplot2*" (Wickham et al., 2018), *"esquisse"* (Meyer et al., 2020) e "*scales*" (Wickham e Seidel, 2019).

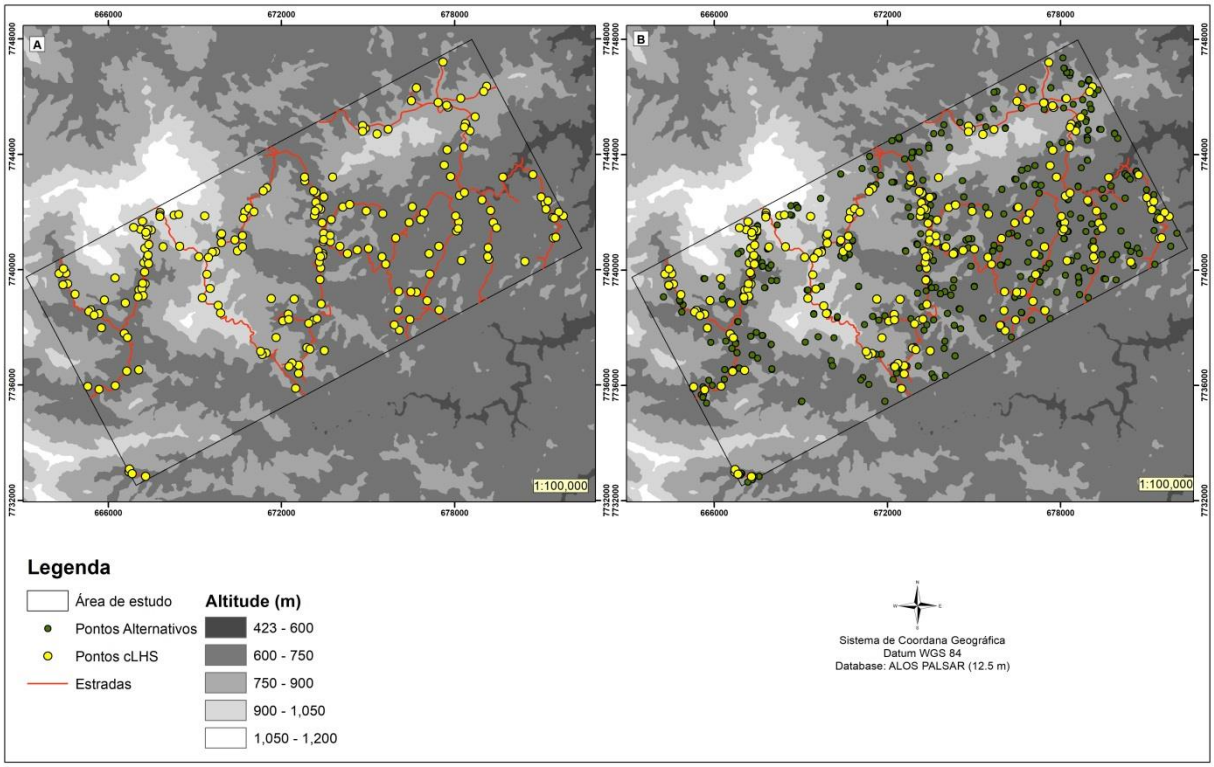


Figura 2. A: Pontos cLHS alocados na área de estudo. B: Pontos alternativos aos pontos cLHS.

2.6 Coleta de amostras in situ

As atividades de campo foram realizadas em quatro dias, cuja área total foi dividida em três subáreas com equipes de campo distintas. Para tanto, dentre os 200 pontos amostrais determinados pelo cLHS, foram realizados 150 pontos de observação (sem coleta) e 50 perfis de solos foram descritos, coletados (Santos et al., 2013) e classificados até segundo nível categórico (Embrapa, 2018).

3. Resultados e Discussão

3.1 Importância das variáveis

Considerando a soma total de importância, para cada modelo e resolução espacial, a variável *"MDE"* foi considerada como a mais importante assim como a variável *"geologia"* (Tabela 2). O MDE representa a altitude da célula em relação a um plano de referência. Sua importância já é reconhecida nos estudos de MDS, uma vez que possui influência sobre o clima, a vegetação e a energia potencial (Wilson e Gallant, 2000).

Tabela 2. Importância das covariáveis por modelos preditivos e modelos digitais de elevação

Nº	Covariáveis	Random Forest					SVM					kNN					C5.0					GBM					Soma
		A1	*A3*	*AS*	*L*	*S*	*A1*	*A3*	*AS*	*L*	*S*	*A1*	*A3*	*AS*	*L*	*S*	*A1*	*A3*	*AS*	*L*	*S*	*A1*	*A3*	*AS*	*L*	*S*	
1	MDE	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	2500
2	GEOLOGIA	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	2500
3	IRON – S (SETEMBRO)	100	100	100	100	98	100	100	100	100	100	100	100	100	100	100	67	79	77	87	58	100	100	100	100	98	2364
4	CLAY - L (JANEIRO)	100	100	100	100	100	100	34	85	100	43	100	34	85	100	43	88	65	80	95	85	100	100	100	100	100	2137
5	TERRAIN SURFACE TEXTURE	2	100	97	NA	100	46	100	95	NA	97	46	100	95	NA	97	64	94	93	NA	96	17	100	98	NA	100	1637
6	TERRAIN SURFACE CONVEXITY	5	99	100	NA	100	31	100	95	2	99	31	100	95	2	99	87	86	94	6	96	8	97	100	1	99	1632
7	STANDARDIZED HEIGHT	97	93	100	31	97	100	100	100	100	100	100	100	100	100	100	2	6	4	2	10	38	14	73	14	40	1621
8	CLAY - S (SETEMBRO)	98	47	1	100	35	99	44	61	97	24	99	44	61	97	24	57	26	9	80	27	70	44	39	100	79	1462
9	SAVI - S (SETEMBRO)	78	1	NA	97	6	100	97	100	99	98	100	97	100	99	98	34	15	7	78	18	16	1	1	90	10	1440
10	SLOPE HEIGHT	11	11	7	NA	26	100	100	100	84	100	100	100	100	84	100	61	33	44	40	42	50	12	25	25	23	1378
11	IRON - L (SETEMBRO)	73	11	28	100	32	85	25	55	86	56	85	25	55	86	56	28	24	28	71	35	66	18	25	69	42	1264
12	CLAY - S (JANEIRO)	100	98	99	100	70	NA	NA	NA	1	NA	NA	NA	NA	1	NA	67	49	56		72	96	33	41	98	12	1082
13	NORTHERNESS	84	100	100	NA	98	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	22	86	84	6	43	20	100	96	NA	88	927
14	VALLEY DEPTH	NA	4	63	NA	26	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	32	60	72	14	71	35	88	21	NA	48	534
15	NORMALIZED HEIGHT	NA	NA	NA	NA	NA	38	84	9	NA	69	38	84	9	NA	69	1	2	10	1	4	NA	NA	NA	NA	NA	418
16	EASTERNESS	34	32	2	NA	9	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	43	29	25	1	34	36	78	46	NA	4	373
17	SAVI - L (JANEIRO)	9	NA	3	100	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	13	8	17	61	17	32	2	28	72	NA	363
18	SAVI - L (SETEMBRO)	8	4	NA	28	2	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	8	19	8	21	6	42	4	2	50	27	229
19	IRON - S (JANEIRO)	1	NA	NA	3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	26	12	14	21	13	41	NA	3	16	19	169
20	IRON - L (JANEIRO)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	19	9	8	9	6	25	8	2	48	4	138
21	GRADIENT	NA	NA	NA	NA	NA	NA	7	NA	46	NA	NA	7	NA	46	NA	3	11	1	8	1	NA	NA	NA	2	NA	132

22	MRVBF	NA	NA	NA	NA	NA	NA	9	NA	24	9	NA	9	NA	24	9	1	2	NA	9	1	NA	NA	NA	NA	NA	97
23	ASPECT	NA	NA	NA	22	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	7	8	12	4	12	4	NA	NA	6	7	82
24	SLOPE	NA	NA	NA	NA	NA	NA	NA	NA	35	3	NA	NA	NA	35	3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	76
25	SAVI - S (JANEIRO)	NA	NA	NA	19	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4	8	16	9	NA	NA	NA	3	NA	59
26	DIURNAL ANISOTROPIC HEATING	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	15	6	10	10	7	4	1	NA	2	NA	55
27	CLAY - L(SETEMBRO)	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	1	NA	NA	NA	NA	4	26	4	14	1	NA	NA	NA	3	NA	54
28	SURFACE SPECIFIC POINTS	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	6	11	7	18	12	NA	NA	NA	NA	NA	54
29	REAL SURFACE ÁREA	NA	NA	NA	NA	NA	NA	NA	NA	20	2	NA	NA	NA	20	2	NA	2	1	NA	1	NA	NA	NA	NA	NA	48
30	TOPOGRAPHIC POSITION INDEX	NA	NA	NA	NA	NA	NA	NA	NA	4	NA	NA	NA	NA	4	NA	7	1	6	13	12	NA	NA	NA	NA	NA	47
31	CURVATURE FLOW LINE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	11	5	10	8	NA	NA	NA	NA	NA	NA	34
32	CURVATURE TOTAL	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	7	8	NA	1	1	NA	NA	NA	NA	NA	17
33	GEOMORPHONS	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	4	6	3	2	2	NA	NA	NA	NA	NA	17
34	MRRTF	NA	NA	NA	NA	NA	NA	NA	NA	2	NA	NA	NA	NA	2	NA	2	3	NA	4	2	NA	NA	NA	NA	NA	15
35	CONVERGENCE INDEX	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	1	2	1	2	NA	NA	NA	NA	NA	7
36	CURVATURE MAXIMAL	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	3	NA	1	3	NA	NA	NA	NA	NA	NA	7
37	NDVI - S (JANEIRO)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	6
38	CURVATURE MINIMAL	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	2	2	NA	NA	NA	NA	NA	NA	5
39	CURVATURE PLAN	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	1	NA	1	2	NA	NA	NA	NA	NA	5
40	CURVATURE GENERAL	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	1	NA	1	1	NA	NA	NA	NA	NA	4
41	CURVATURE PROFILE	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	1	2	NA	NA	NA	NA	NA	NA	NA	4
42	MID SLOPE POSITION	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	NA	1	NA	NA	NA	NA	NA	1	NA	3
43	CURVATURE LONGITUDINAL	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	2	NA	NA	NA	NA	NA	NA	2
44	CURVATURE TANGENCIAL	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1	1	NA	NA	NA	NA	NA	2

Legenda: A1 – ALOS 12,5 m; A3 – ALOS 30 m; AS – ASTER 30 m; L – LIDAR 2 m; S – SRTM 30 m; NA – variável não apresentou importância significativa.

Dentre os atributos derivados do MDE, o *"terrain surface texture*", *"terrain surface convexity", "standardized height"* e "*slope height"* também foram considerados importantes neste estudo. Estas covariáveis estão relacionadas com as formas do terreno, que influenciam diretamente no tempo de exposição dos materiais de origem e na intensidade e direção do fluxo da água, sendo então, as características geomorfológicas capazes de regular os processos pedogenéticos.

Esta relação pode ser entendida através do conceito de catena proposta por Milne (1935), a qual pode ocorrer à formação de diferentes solos dependendo da sua posição na paisagem, devido à movimentação da matéria e ao movimento e distribuição de água, na qual podem ocorrer áreas de erosão nos locais mais íngremes e áreas de acúmulo nos locais mais planos. Desta forma, em regiões que apresentam maiores declividades e formatos convexos, os processos de erosão são mais intensos e ocorrem solos menos desenvolvidos, onde é comum encontrar afloramentos de rochas e solos como Cambissolo Háplicos, Neossolos Regolíticos e Litólicos (Kampf e Curi, 2012). Já nas áreas onde as declividades são mais suaves os solos são mais desenvolvidos, como os Latossolos e Argissolos (Campos, 2012).

Nota-se ainda que as covariáveis "*Terrain surface texture*" e "*Terrain surface convexity*" são sensíveis à mudança de resolução espacial. Estas covariáveis apresentam menor importância nas resoluções mais finas, independente do modelo utilizado, sobretudo para o LIDAR, as quais não apresentam importância significativa. Este fato ressalta a proposição de que para altas resoluções, alguns atributos do terreno podem apresentar um excesso de detalhes, o que pode gerar muito ruído e até mesmo invalidar a acurácia da predição (Cavazzi et al., 2013).

Por conseguinte, segundo Gray et al. (2014) poucos trabalhos consideram a utilização de mapas de geologia (ou litologia) como covariáveis ambientais em mapeamentos digitais de solos. Este fato está associado muitas vezes à ausência destas informações em escala adequada para a área de estudo ou pela falta de entendimento das possíveis relações com a formação do solo. Contudo, a heterogeneidade litológica da área de estudo marcada pela presença de anfibólito, gabro, quartzito, gnaisse, granito, xisto, dentre outros e a disponibilidade de material, possibilitaram a sua utilização e resultaram também na maior influência desta variável. Outros estudos de MDS também apontam a variável geologia como

determinante na predição de classes e/ou propriedades dos solos (Crivelenti et al., 2009; Chagas et al., 2011; Lemercier et al., 2012; Dias, 2015).

Esta heterogeneidade litológica condiciona solos com diferentes tipos de minerais de argila e outros constituintes; o que explica a importância da geologia associada também aos índices espectrais do solo *"clay minerals", "iron oxides"* e "*SAVI*". O "*clay minerals*" evidencia a presença de minerais de argila como ilita, caulinita e montmorilonita nos solos (Sabins, 1997; Chagas et al., 2013). O índice "*iron oxide*" destaca a presença de óxidos e sulfatos de ferro nos solos. Minerais de ferro como goethita e hematita têm baixos valores de reflectância na banda 1 (azul) e elevados valores na banda 3 (vermelho) do Landsat, consequentemente, os solos tendem a apresentar elevados valores da relação entre as bandas 3 e 1, que varia de acordo com a composição mineralógica do solo (Sirtoli, 2008). Chagas (2006) aponta que o uso dos índices "*clay minerals*" e "*iron oxide*" apresenta significativa potencialidade para auxiliar no mapeamento de solos.

Alguns trabalhos também relataram as variáveis derivadas do fator organismo como as mais importantes nos modelos de predição. Taghizadeh-Mehrjardi et al. (2016) citam o NDVI (Índice de vegetação de diferença normalizada) e SAVI (índice de vegetação ajustada ao solo) como variáveis importantes na previsão do conteúdo de COS em solos superficiais. Zeraatpisheh et al. (2017) observaram que o SAVI, índice de argila e NDVI foram as covariáveis mais importantes nos níveis de ordem, subordens e grande grupo para prever tipos de solo em uma região semiárida no Irã.

Contudo, no presente estudo, dentre os índices espectrais, o NDVI não apresentou maior importância, uma vez que o NDVI não possui relação direta com a pedogênese do solo, mas reflete muitas características da vegetação. Este índice realça o comportamento espectral da vegetação e permite avaliar o vigor vegetativo dos estágios sucessionais de determinada vegetação ou cultura. Nota-se ainda que para os dados espectrais devirados do satélite Landsat 8 assim como do Sentinel 2, as variáveis espectrais apresentaram maior importância ao utilizar dados derivados do mês de setembro (mais seco), fato este que demonstra a influência da umidade no comportamento espectral do solo, sendo a reflectância maior quando a umidade do solo diminui.

Outros estudos também retratam a utilização de atributos do terreno, dados de sensores remotos e informações sobre litologia como preditores e com elas a obtenção de melhores resultados (Sirtoli, 2008; Chagas et al., 2010). Sirtoli (2008) complementa que a utilização destas variáveis numa abordagem por RNA possibilitam tornar o levantamento de solos mais rápido e menos dependente da experiência e subjetividade da pessoa que o realiza.

3.2 Desempenho operacional do cLHS

O desempenho do cLHS em relação à sua aplicabilidade prática nos levantamentos e mapeamentos de solos foi abordado em alguns estudos, como por exemplo, Roudier et al., (2012); Clifford et al., (2014) e Ließ (2015; 2020); cujos autores abordam algumas limitações do método e formas de otimização. No presente estudo, algumas restrições operacionais em atividades práticas de campo foram identificadas e descritas na Tabela 3.

Tabela 3. Restrições operacionais do cLHS durante as atividades de campo

Restrição	Descrição	Número de locais
Perturbações	Locais fisicamente perturbados pela construção de infraestruturas e/ou terraplenagem	4
Acesso físico	Locais de acesso limitado por barreiras físicas (ex. porteiras) e/ou de acesso privado	4
Condições do relevo	Terreno muito íngreme não possibilitando o acesso seguro ao local	23
Vegetação	Locais onde a vegetação densa dificulta o acesso	19
Sem acesso	Falta de acesso/Ausência de estradas e/ou vias alternativas para o ponto selecionado	31

*Alguns pontos apresentaram mais de uma restrição.

As restrições aos locais de amostragem estão associadas, sobretudo, a falta de acesso aos pontos selecionados (n: 31), uma vez que a área de estudo apresenta relevo complexo com infraestrutura viária precária, o que dificulta o deslocamento e até mesmo oferece risco à segurança física do pedólogo. Mesmo com a exclusão das áreas urbanas correspondentes aos distritos municipais e as áreas de mineração, quatro pontos foram selecionados em áreas de construções de infraestruturas e propriedades privadas, os quais não foram possíveis ser

identificados em laboratório, apenas nos trabalhos de campo. Além disso, as partes mais elevadas da área estão associadas com área de vegetação natural formada por florestas estacional semidecidual associadas a áreas de extrativismo vegetal e produção de carvão, nas quais se localizaram outros quatro pontos inacessíveis.

Thomas et al. (2012) também encontraram dificuldades de acesso devido à presença de terrenos complexos, distâncias de viagem e cobertura vegetal ao amostrar paisagens montanhosas e com muita vegetação. Cambule et al. (2013) criticaram as técnicas de MDS, incluindo a cLHS, por serem impraticáveis e proibitivamente caras em grandes regiões com dificuldades de acesso devido à falta de estradas ou terrenos difíceis. Kidd et al. (2015) relataram que nas áreas do projeto, as limitações de acesso eram devido às restrições físicas, como portões trancados, cercas e trilhos, além dos proprietários de terras que se recusaram a colaborar com a pesquisa.

Deste modo, dos 200 pontos amostrais determinados pelo cLHS, aproximadamente 30% não puderam ser acessados em campo. Kidd et al. (2015) relataram um porcentagem ainda maior, o total de 44% locais inacessíveis na área de estudo utilizando o cLHS. Os autores consideram este valor uma proporção substancial (e surpreendente), uma vez que indica que a amostragem 'estrita' pode ser problemática em áreas intensivamente usadas, onde existem impedimentos ao acesso e amostragem física.

Atualmente, não existem normas ou metodologias para o estabelecimento do tamanho “ideal” de amostras a serem aplicadas no cLHS. Brungard e Boettinger (2010) sugeriram um tamanho ideal de 200 - 300 amostras em uma área de estudo de 300 km² usando o cLHS para aplicação em modelos preditivos, os quais seriam capaz de fornecer uma melhor aproximação da distribuição de todas as covariáveis ambientais. Malone et al. (2019) recomendam para um cenário prático um tamanho mínimo ideal de 110 amostras em uma área equivalente a 220 km² para capturar a variação nos dados auxiliares, utilizando o cLHS em uma região da Austrália. Entretanto, o tamanho amostral ideal dependerá da complexidade da área de estudo, principalmente em termos geológico e geomorfológico. Quanto maior for essa complexidade, maior a necessidade amostral para representar todas as variações da paisagem.

Sendo assim, conforme os autores supracitados, ao considerar uma proporção de 0,5 – 1,0 amostras/km² recomendado para o cLHS, os pontos acessados atenderiam a este critério de tamanho mínimo ideal, com proporção de 0,97 amostras/km², para predição de classes de solos. Brungard e Boettinger (2010) complementam que o tamanho ideal da amostra também depende do modelo a ser utilizado para prever a distribuição do solo. Alguns modelos preditivos, como árvores de classificação e regressão e alguns métodos geoestatísticos "consomem muita informação" e requerem grandes quantidades de dados de entrada.

Nos estudos de Roudier et al. (2012); Mulder et al. (2013) e Kidd et al. (2015) os autores adaptaram o cLHS para torná-lo mais adequado para áreas de difícil acesso, devido à restrições operacionais e a necessidade subsequente de navegar para um local alternativo. No presente estudo, o método do cLHS modificado utilizando pontos alternativos permitiu maior acessibilidade, possibilitando desta forma o acesso a cerca de 17 % dos locais de amostragem. No entanto, para os pontos do cLHS que tiveram acesso restrito como descritos na Tabela 2, os pontos alternativos a eles também estavam localizados em áreas de difícil acesso; tornando inviável, da mesma forma, o acesso. Ma et al. (2019) também ressaltam essa questão em seu estudo, que devido à imprevisibilidade das condições de campo, amostras alternativas também podem não estar disponíveis em alguns casos. Contudo, como destacado por Clifford et al. (2014) e Malone et al. (2019) modificações empregadas no cLHS não garantem acessibilidade, apenas aumentam sua probabilidade de acesso.

De acordo com Stumpf et al. (2016) como consequência das restrições e limitações descritas, a maioria dos projetos de amostragem se concentra apenas em refletir a variação da propriedade alvo do solo. Geralmente, eles não consideram melhorias de operacionalidade e eficiência em termos de acessibilidade, uso integrativo de amostras herdadas e otimização do tamanho do conjunto de amostras (Lagacherie, 2008; Cambule et al., 2013). Portanto, os avanços no levantamento de dados do solo para o MDS dependem do tratamento abrangente dos potenciais estatísticos, operacionais e de eficiência dos projetos de amostragem.

4. Considerações Finais

A seleção do conjunto de variáveis a serem utilizadas no método cLHS ainda é negligenciada em muitos estudos. Este processo é de fundamental importância para garantir uma melhor representação da variabilidade ambiental, que vai depender da disponibilidade de variáveis

bem como da complexidade da área. O método cLHS possibilita a realização de levantamentos e mapeamentos de solos em campo de forma eficiente, mesmo que ocorram problemas operacionais de amostragem devido às restrições de acesso. Por meio dessa técnica é possível a identificação de pontos amostrais com suas coordenadas pré-definidas e a restrição prévia de áreas inacessíveis, como as áreas urbanas e de mineração. A flexibilidade da amostragem, como apresentada no presente estudo pelo método cLHS modificado, aumentou a possibilidade de acesso aos locais de amostragem e assegurou a redução de tempo e custo para alcançar um local completamente novo e alternativo nos trabalhos de campo. Desta forma, foi possível obter um tamanho mínimo ideal de pontos amostrais a serem utilizados em modelos preditivos nos estudos de MDS. No entanto, poucos estudos abordam problemas e restrições operacionais para seleção de amostragem utilizando o cLHS, ainda são necessárias pesquisas futuras a fim de aperfeiçoar a identificação de locais e coleta de amostras, considerando, por exemplo, a densidade do tamanho da amostra e como isso pode afetar a distribuição das previsões finais do MDS.

5. Bibliografia

Arruda GP, Dematte JAM, Chagas CS. Mapeamento digital de solos por redes neurais artificiais com base na relação solo-paisagem. R. Bras. Ci. Solo. 2013; 37:327-338. http://dx.doi.org/10.1590/S0100-06832013000200004.

Bagatini T, Giasson E, Teske R. Selection of sampling density based on data from areas already mapped for training decision tree models in digital soil mapping. R. Bras. Ci. Solo. 2015; 39:960-967. https://10.1590/01000683rbcs20140289

Bivand R, Keitt T, Rowlingson B. 2018. rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 1.3-6. Available at https://CRAN.Rproject.org/package=rgdal (accessed 15 Jan 2019)

Breiman L, Cutler A. 2018. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.6-14. Available at https://cran.r-project.org/web/packages/randomForest/index.html

Brenning A. 2008. Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models. In: Böhner, J., Blaschke, T., Montanarella, L. (eds.), SAGA – Seconds Out (= Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie) 19, 23–32.

Brungard CW, Boettinger JL. Conditioned Latin Hypercube Sampling: Optimal sample size for Digital Soil Mapping of Arid Rangelands in Utah, USA. In: Boettinger JL, Howell DW, Moore AC, Hartemink AE, Kienast-Brown S, eds. Digital Soil Mapping. Bridging Research, Environmental Application and Operation. Dordrecht, Springer, 2010. p.67-75. (Progress in Soil Science, 2)

Brus DJ, Gruijter JJ, van Groenigen JW. Chapter 14. Designing spatial coverage samples using the k-means clustering algorithm. Developments in Soil Science. 2006; 31:183–192.

https://10.1016/S0166-2481(06)31014-8.

Burrough PA, van Gaans PFM, MacMillan RA. High-resolution landform classification using fuzzy k-means. Fuzzy Sets and Systems. 2000; 113:37-52. https://doi.org/10.1016/S0165-0114(99)00011-1

Cambule AH, Rossiter DG, Stoorvogel JJ. A methodology for digital soil mapping in poorly-accessible áreas. Geoderma. 2013; 192:341-353. https://doi.org/10.1016/j.geoderma.2012.08.020

Campos MCC. Relações solo-paisagem: conceitos, evolução e aplicações. Ambiência. 2012; 8:963 – 982. https://10.5777/ambiencia.2012.05.01rb

Carvalho Junior W, Chagas CS, Lagacherie P, Calderano Filho B, Bhering SB. Evaluation of statistical and geostatistical models of digital soil properties mapping in tropical mountain regions. R. Bras. Ci. Solo. 2014; 38:706-717. http://dx.doi.org/10.1590/S0100-06832014000300003

Cavazzi, S., Corstanje, R., Mayr, T., Hannam, J., Fealy, R., 2013. Are fine resolution digital elevation models always the best choice in digital soil mapping? Geoderma. 195-196, 111-121. https://doi.org/10.1016/j.geoderma.2012.11.020

Chagas CS. Mapeamento digital de solos por correlação ambiental e redes neurais em uma bacia hidrográfica de domínio de mar de morros [thesis]. Viçosa: Universisade Federal de Viçosa; 2006.

Chagas CS, Carvalho Júnior W, Bhering SB. Integração de dados do Quickbird e atributos do terreno no mapeamento digital de solos por redes neurais artificiais. R. Bras. Ci. Solo. 2011; 35:693-704. http://dx.doi.org/10.1590/S0100-06832011000300004

Chagas CS, Vieira CAO, Fernandes Filho EI. Comparison between artificial neural networks and maximum likelihood classification in digital soil mapping. R. Bras. Ci. Solo. 2013; 37:339-351. http://dx.doi.org/10.1590/S0100-06832013000200005

Clifford D, Payne J, Pringle M, Searle R, Butler N. Pragmatic soil survey design using flexible Latin hypercube sampling. Comput. Geosci. 2014; 67:62-68. https://doi.org/10.1016/j.cageo.2014.03.005

Codemig - Companhia de desenvolvimento econômico de Minas Gerais – Mapa Geológico. 2014. Available at https://www.portaldageologia.com.br

Crivelenti RC, Coelho RM, Adami SF, Oliveira SRM. Mineração de dados para a inferência de relações solo-paisagem em mapeamentos digitais de solo. Rev. Agro. Bras. 44:1707-1715, 2009. http://dx.doi.org/10.1590/S0100-204X2009001200021

Dias LMS. Predição de classes de solo por atributos do meio físico e de sensoriamento remoto em área da Bacia Sedimentar do São Francisco [dissertation]. Campinhas: Instituto Agronômico; 2015.

Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA. Sistema Brasileiro de Classificação de Solos. Rio de Janeiro: Embrapa Solos, 2018.

Garcia LC. Fenologia de espécies da canga em Barão de Cocais, Quadrilátero Ferrífero de Minas Gerais [dissertation]. Belo Horizonte: Universidade Federal de Minas Gerais; 2007.

Gray J, Bishop T, Wilford J. Lithology as a powerful covariate in digital soil mapping. In: Arrouays D, Mckenzie N, Hempel J, Forges AR, Mcbratney A, eds. GlobalSoilMap: Basis of the Global Spatial Soil Information System. Leiden: CRC Press, 2014. p. 433–439.

Grimm R, Behrens T, Märker M, Elsenbeer H. Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis. Geoderma. 2008; 146:102–113. https://doi.org/10.1016/j.geoderma.2008.05.008

Greenwell B, Boehmke B, Cunningham J. 2019. gbm: Generalized Boosted Regression Models. R package version 2.1.5. Available at https://cran.r-project.org/web/packages/gbm/index.html

Hengl T, Rossiter DG, Stein A. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Aust J Soil Res. 2003; 41:1403–1422. https://doi.org/10.1071/SR03005

IBGE. 2001. Base cartográfica vetorial contínua do Brasil ao milionésimo - BCIM. IBGE, Rio de Janeiro.

Jenny H. Factors of soil formation. A system of quantitative pedology [Internet]. Dover Publ. 1941.

Kämpf N, Curi N. Formação e evolução do solo (Pedogênese). In: Ker JC, Schaefer CEGR, Vidal-Torrado P. Pedologia: fundamentos. Sociedade Brasileira de Ciência do Solo.Viçosa-MG, 2012, p.207-302.

Kidd D, Malone B, McBratney AB, Minasny B, Webb M. Operational sampling challenges to Digital Soil Mapping in Tasmania, Australia. Geoderma Regional. 2015; 4:1-10. . https://10.1016/j.geodrs.2014.11.002

Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, Benesty M, Lescarbeau R, Ziem A, Scrucca L, Tang Y, Candan C, Hunt T. 2020. caret: Classification and Regression Training. R package version 6.0-85. Available at https://cran.r-project.org/web/packages/caret/index.html

Lagacherie P. Digital soil mapping: a state of the art. In: Hertmink AE, McBratney A, Mendonça-Santos ML. Digital soil mapping with limited data. Springer; 2008. p. 3-14.

Lemercier B, Lacoste M, Loum M, Walter C. Extrapolation at regional scale of local soil knowledge using boosted classification trees: A two-step approach. Geoderma. 2011; 171-172:75-84. https://doi.org/10.1016/j.geoderma.2011.03.010

Ließ M. Sampling for regression-based digital soil mapping: Closing the gap between statistical desires and operational applicability. Spat Stat. 2015; 13:106–122. https://doi.org/10.1016/j.spasta.2015.06.002

Ließ M. At the interface between domain knowledge and statistical sampling theory: Conditional distribution based sampling for environmental survey (CODIBAS). Catena. 2020; 187:1-10. https://doi.org/10.1016/j.catena.2019.104423

Ma T, Wei T, Qin CZ, Zhu AX, Qi F, Liu J, Zhao F, Pan H. In-situ recommendation of alternative soil samples during field sampling based on environmental similarity. Earth Science Informatics. 2019; 1-15. https://doi.org/10.1007/s12145-019-00407-x

Machado IR, Giasson E, Campos AR, Costa JJF, Silva EB da, Bonfatti BR. Spatial Disaggregation of Multi-Component Soil Map Units Using Legacy Data and a Tree-Based Algorithm in Southern Brazil. R. Bras. Ci. Solo. 2018; 42:1-14. http://dx.doi.org/10.1590/18069657rbcs20170193

Malone BP, Minansy B, Brungard C. Some methods to improve the utility of conditioned Latin hypercube sampling. PeerJ. 2019; 7:1-17. https://10.7717/peerj.6451

Maynard J, Levi MR. Hyper-temporal remote sensing for digital soil mapping: Characterizing soil-vegetation response to climatic variability. Geoderma. 2017; 285:94-109. https://doi.org/10.1016/j.geoderma.2016.09.024

Meyer F, Perrier V, Carroll I. 2020. esquisse: Explore and Visualize Your Data Interactively. R package version 0.2.3. Available at https://cran.r-project.org/web/packages/esquisse/index.html

Milne G. Some suggested units of classification and mapping particularly for East African soils. Soil Research. 1935; 4:183–198.

Minasny B, McBratney AB. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput Geosci. 2006; 32:1378–1388. https://doi.org/10.1016/j.cageo.2005.12.009

Mitchell TM. Machine Learning. New York; McGraw-Hill. 1997.

Mora-Vallejo A, Claessens L, Stoorvogel J, Heuvelink GBM. Small scale digital soil mapping in Southeastern Kenya. CATENA. 2008; 76:44–53. https://doi.org/10.1016/j.catena.2008.09.008

Mulder VL, De Bruin S, Schaepman M. Representing major soil variability at regional scale by constrained Latin hypercube sampling of remote sensing data. Int. J. Appl. Earth Obs. Geoinf. 2013; 21:301-310. https://doi.org/10.1016/j.jag.2012.07.004

Pinheiro HSK. Métodos de Mapeamento Digital Aplicados na Predição de Classes e Atributos dos Solos da Bacia Hidrográfica do Rio Guapi Macacu, RJ [thesis]. Rio de Janeiro: Universidade Federal Rural do Rio de Janeiro; 2015.

Robert JH. 2019. raster: Geographic Data Analysis and Modeling. R package version 2.8-19. Available at http://CRAN.R-project.org/package=raster (accessed 15 Jan 2019)

Roudier P, Hewitt AE, Beaudette DE. A conditioned latin hypercube sampling algorithm incorporating operational constraints. In: Minasny B, Malone BP, McBratney AB, eds. Digital soil assessments and beyond. London, CRC Press/Balkema, 2012. p.227-232.

Roudier P, Brugnard C, Beaudette D, Louis B. 2019. clhs: Conditioned Latin Hypercube Sampling. R package version 0.7-2. Available at https://cran.r-project.org/web/packages/clhs/index.html

Sabins FF. Remote sensing: Principles and Interpretation. 3nd ed. New York: W.H. Freeman and Company. 1997. 432p.

Santos RD, Lemos RC, Santos HG, Ker JC, Anjos LHC, Shimizu SH. Manual de descrição e coleta de solo no campo. 6.ed. rev. e ampl. Viçosa: SBCS, 2013. 100p.

Syed ME. Attribute weighting in K-nearest neighbor classification [thesis]. Finland: University of Tampere; 2014.

Silva SHG, Teixeira AF dos S, Menezes MD de, Silva SHG, Moreira FM de S, Curi N. Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence spectrometer (pXRF). Ciênc. agrotec. 2017; 41:648-664. http://dx.doi.org/10.1590/1413-70542017416010317

Sirtoli AE. Mapeamento de solos com auxílio de atributos do terreno, índices espectrais e geologia integrados por redes neurais artificiais [thesis]. Curitiba: Universidade Federal do Paraná; 2008.

Stumpf F, Schmidt K, Behrens T, Schönbrodt-Stitt S, Buzzo G, Dumperth C, Wadoux A, Xiang W, Scholten T. Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. J. Plant Nutr. Soil Sci. 2016; 179:499–509. https://10.1002/jpln.201500313

Szatmári G, Barta K, Pásztor L. An application of a spatial simulated annealing sampling optimization algorithm to support digital soil mapping. Hungarian Geographical Bulletin. 2015; 64:35-48. https://doi.org/10.15201/hungeobull.64.1.4

Taghizadeh-Mehrjardi R, Nabiollahi K, Kerry R. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma. 2016; 266:98–110. https://doi.org/10.1016/j.geoderma.2015.12.003

Teske R, Giasson E, Bagatini T. Comparação do uso de modelos digitais de elevação em mapeamento digital de solos em Dois Irmãos, RS, Brasil. R. Bras. Ci. Solo. 2014; 38:1367-1376. http://dx.doi.org/10.1590/S0100-06832014000500002

Teske R, Giasson E, Bagatini T. Comparação de esquemas de amostragem para treinamento de modelos preditores no mapeamento digital de classes de solos R. Bras. Ci. Solo. 2015; 39:14–20. http://dx.doi.org/10.1590/S0100-06832014000500002

Thomas M, Odgers N, Ringrose-Voase A, Grealish G, Glover M, Dowling T. Soil survey design for management-scale digital soil mapping in a mountainous southern Philippine catchment. In: Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping. Sydney, Australia, CRC Press, 2012, p. 233.

Wang S, Wang Q, Adhikari K, Jia S, Jin X, Liu H. Spatial-Temporal Changes of Soil Organic Carbon Content in Wafangdian, China. Sustainability. 2016; 8:1-16. https://doi.org/10.3390/su8111154

Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K., Wilke C, Woo K. 2018. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 3.1.0. Available at https://cran.r-project.org/web/packages/ggplot2 (accessed 15 Jan 2019)

Wickham H, François R, Henry L, Muller K. 2019. dplyr: A Grammar of Data Manipulation. R package version 0.8.3. Available at https://cran.r-project.org/web/packages/dplyr/index.html (accessed 15 Jan 2019)

Wickham H, Seidel D. scales: Scale Functions for Visualization. 2019. R package version 1.1.0. Available at https://cran.r-project.org/web/packages/scales/index.html

Wilson JP, Gallant JC, 2000. Terrain Analysis: Principles and Applications. Jonh Wiley & Sons, Inc., Nova Jersey. ISBN 0-471-32188-5.

Yang F, White M, Michaelis AR, Ichii K, Hashimoto H, Votava P, Zhu AX, Nemani RR. Prediction of continental-scale evapotranspiration by combining MODIS and AmeriFlux data through support vector machine. IEEE Trans. Geosci. Remoto Sens. 2006; 44:3452 – 3461. https://doi.org/10.1109/TGRS.2006.876297

Zeraatpisheh M, Ayoubi S, Jafari A, Finke P. Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. Geomorphology. 2017; 285:186–204. https://doi.org/10.1016/j.geomorph.2017.02.015

Artigo 4

TÉCNICAS EM MAPEAMENTO DIGITAL DE SOLOS UTILIZANDO CONJUNTO DE DADOS DESBALANCEADO

1. Introdução

O mapeamento digital de solos (MDS) surgiu da necessidade de gerar informações pedológicas com menor demanda de tempo, custos e recursos humanos; o qual se baseia no uso de modelos numéricos para inferência das variações espaciais e temporais nos tipos de solos e em suas propriedades; a partir de observações de campo, do conhecimento dos solos e do uso de variáveis ambientais correlacionadas (Lagacherie e McBratney, 2007).

O MDS permite a quantificação e avaliação da precisão e exatidão das informações pedológicas geradas, tornando os procedimentos mais rápidos, reprodutíveis e os resultados comparáveis (McBratney e DeGruijter, 1992). Contudo, a precisão do MDS pode ser afetada pelo desbalanceamento de classes dos solos, o qual é considerado como um dos desafios do *machine learning*, cuja ocorrência se dá quando uma classe, geralmente a classe de interesse do ponto de vista da tarefa de aprendizagem (classe positiva ou minoritária), é subrepresentada no conjunto de dados. Em outras palavras, o número de instâncias negativas (maioria) supera a quantidade de instâncias positivas de classe (Galar et al., 2012).

De modo ideal, espera-se que a relação quantitativa das classes de solo com covariáveis seja suficientemente distinta para diferenciação entre as classes. No entanto, em paisagens complexas, o compartilhamento de posições semelhantes da paisagem por classes de solo, pode resultar em sobreposição do espaço covariável, com a classe mais abundante dominando a classificação geral (Gopi et al., 2016). Logo, esta distribuição assimétrica tem capacidade de comprometer significativamente o desempenho dos algoritmos de aprendizado.

Com isso, algumas técnicas são utilizadas para reduzir o desequilíbrio da distribuição de classes e melhorarem a capacidade de classificação dos modelos preditivos, as quais podem ser tratadas em (1) nível de algoritmo - que tentam adaptar os algoritmos de aprendizado para direcioná-lo à classe minoritária; (2) nível de dados - geralmente considerada mais funcional, na qual é reamostrada a distribuição de classes, independente do algoritmo utilizado, com uma

etapa de pré-processamento e; (3) nível de sensibilidade aos custos - incorpora as transformações no nível dos dados (adicionando custos às instâncias) e as modificações no nível do algoritmo (modificando o processo de aprendizado para aceitar custos) (Krawczyk, 2016). A principal desvantagem dessa abordagem é a necessidade de definir custos de classificação, que geralmente não estão disponíveis nos conjuntos de dados (Garcia e Herrera, 2009; Galar et al., 2012).

Por conseguinte, ainda que o desbalanceamento de classes seja um problema reconhecido no aprendizado de máquina (Liu et al., 2009; Ma et al., 2019) e ocorra naturalmente na paisagem, esta questão é relatada em diversas áreas do conhecimento (Mazurowski et al., 2008; Lu e Wang, 2008), mas ainda tem sido pouco abordada nos estudos de mapeamento de solo. A maioria das pesquisas que utiliza classes desbalanceadas em aprendizado de máquinas se concentra na utilização de algoritmos específicos ou os desequilíbrios entre classes são binários. Contudo, observa-se que existem desequilíbrios entre várias classes (He e Garcia, 2009) que não são relatados nos trabalhos de MDS.

Diante disso, o objetivo do presente trabalho consiste em avaliar como o desbalanceamento de classes afeta o desempenho dos algoritmos de aprendizado de máquina, a fim de auxiliar no desenvolvimento de técnicas para os desafios futuros de mapeamento de solo no País.

2. Material e Métodos

2.1 Área de estudo

A área de estudo está localizada no município de Mariana - Minas Gerais, com uma área total de 144 km². A área destaca-se por apresentar uma heterogeneidade geológica composta por litotipos dos Supergrupos Rio das Velhas e Minas; Grupos Nova Lima, Maquiné, Piracicaba, Sabará e Itacolomi; Complexos Santa Barbará, Mantiqueira e Monsenhor Isidro e; Suíte Alto Maranhão (Codemig, 2014). Segundo a classificação de Koeppen, o clima é do tipo Cwa caracterizado como tropical semiúmido, com duas estações climáticas bem definidas: (1) inverno seco, com temperaturas inferiores a 18ºC; (2) verão úmido, com temperaturas superiores a 22ºC (Garcia, 2007; Varajão 2009). A vegetação predominante são os Campos Rupestres em relevo complexo marcado por cristas, escarpas, encostas do Quadrilátero Ferrífero e Planaltos dissecados do Rio Piracicaba e do alto Rio Doce (IBGE, 2001; Garcia, 2007).

2.2 Dados de solos

Foram descritas e coletadas 139 amostras de solos, conforme Santos et al. (2013), utilizando o método de amostragem cLHS (SENA et al., *in review*). Posteriormente, as amostras foram submetidas às análises físicas e químicas de rotina no Laboratório de Análises de Rotina – UFV. Os solos foram classificados até o segundo nível categórico de acordo com o Sistema Brasileiro de Classificação de Solos (Embrapa, 2018). A Tabela 1 apresenta as seis classes de solos identificadas na área de estudo, cuja distribuição de amostras por classe retrata o desbalanceamento das classes, onde as classes 1, 2 e 6 foram consideradas como classes minoritárias e a classe 3 foi considerada como a classe majoritária.

Tabela 1. Número de amostras por classe de solo

Número da Classe	Classe de Solo	Amostras/Classe
1	Argissolo Vermelho-Amarelo	7
2	Gleissolo Háplico	9
3	Cambissolo Háplico	69
4	Latossolo Vermelho-Amarelo	19
5	Latossolo Amarelo	25
6	Latossolo Vermelho	10
Total		**139**

2.3 Conjunto de variáveis

As covariáveis ambientais utilizadas incluem os atributos primários e secundários do terreno derivados do ALOS PALSAR (12,5 m): convergence index, cross sectional curvature, flow line curvature, general curvature, longitudinal curvature, maximal curvature, minimal curvature, plan curvature, profile curvature, tangencial curvature, total curvature, classification curvature, diurnal anisotropic heating, easterness, gradient, hill, hill index, landform, mass balance index, digital elevation model, mid-slope position, morphometric protection index, multiresolution index of ridge top flatness, multiresolution index of valley bottom flatness, normalized height, northerness, real surface area, slope, slope height, slope index, standardized height, surface specific points, terrain ruggedness index, terrain surface convexity, terrain surface texture, topographic position index, total, direct, diffuse and duration solar radiation, valley, valley depth, valley index, vector ruggedness measure, topographic wetness index. E os índices espectrais derivados das imagens do LANDSAT 8 e

SENTINEL 2 (Normalized difference vegetation index - NDVI, soil-adjusted vegetation index - SAVI, Clay minerals e Iron oxides).

A obtenção dos atributos do terreno foi realizada através do programa computacional R (R Core Team, 2019), utilizando os pacotes *"rsaga"* (Brenning, 2008), *"raster"* (Robert, 2019) e *"rgrass7"* (Bivand et al., 2019). A partir do conjunto de dados foram removidas as variáveis altamente correlacionadas (correlação não linear > 95 %). Posteriormente, foi utilizado o algoritmo *Recursive Feature Elimination* (RFE), implementado no pacote *Caret* (Kuhn et al., 2020a), o qual consiste em um processo iterativo que elimina os preditores menos importantes do modelo com base em uma medida inicial da importância dos preditores (Kuhn e Johnson, 2013). Após este processo, o conjunto de dados foi dividido aleatoriamente em 75% para treinamento e 25% para validação dos modelos.

2.4 Método de reamostragem

Inicialmente, os modelos foram executados com os dados desbalanceados para verificar como o desequilíbrio de classes afeta o desempenho dos modelos. Em seguida, foi aplicado métodos de reamostragem, que consiste na abordagem comumente utilizada para lidar com conjunto de dados desbalanceados, onde é realizado um pré-processamento antes do treinamento de cada classificador. A principal vantagem desse método é que não precisam ser realizadas alterações nos algoritmos de aprendizado (Ng et al., 2015). Com isso foram aplicados dois métodos:

1 – Dados Balanceados utilizando o *Oversampling* e *Undersampling*

O *Oversampling* consiste em um método não-heurístico que replica aleatoriamente as amostras da classe minoritária dos dados de treinamento, visando obter um conjunto de dados mais equilibrado a partir da distribuição original. Porém, neste método pode ocorrer a repetição das amostras existentes e aumentar a probabilidade do excesso de ajuste. Enquanto o *Undersampling* consiste em reduzir as amostras, eliminando aleatoriamente exemplos pertencentes à classe majoritária do conjunto de treinamento, com o objetivo de equilibrar o número de cada classe. Este método possui como desvantagem o fato de que podem ser descartadas amostras úteis, principalmente, para definir o limite de decisão entre as classes (Krawczyk, 2016);

2 – Dados Balanceados utilizando o *SMOTE*

Este método é considerado um dos mais eficientes para reamostragem de dados no aprendizado de máquina (Garcia et al., 2016), o qual adiciona amostras sintéticas à classe minoritária a fim de aumentar o espaço de recurso desta classe e auxilia os classificadores a melhorar sua capacidade de generalização (Chawla et al., 2002; Fernández et al., 2018).

Para o *Oversampling* e *Undersampling* foram executadas duas funções de tratamento de dados: *ubOver ()* e *ubUnder ()*, respectivamente, do pacote *unbalanced* (Dal Pozzolo et al., 2015). O algoritmo *SMOTE* foi aplicado utilizando a função *DMwR* no software R (R Core Team, 2019). A subamostragem foi realizada para a classe majoritária 3 e a sobreamostragem para as classes minoritárias 1, 2 e 6. As classes majoritárias foram subamostradas para metade e as minorias foram sobreamostradas duas vezes, para tornar a distribuição das classes mais equilibrada.

2.5 Algoritmos de aprendizado de máquina

Foram utilizados três algoritmos (C5.0, Support Vetor Machine - SVM e Random Forest - RF) para a predição das classes de solos utilizando o conjunto de dados de treinamento por 100 vezes. Os algoritmos foram implementados nos pacotes *C50* (Kuhn et al., 2020b), *e107*1 (Meyer et al., 2019) e *RandomForest* (Breiman et al., 2018), respectivamente, no software R (R Core Team, 2019).

O algoritmo C5.0 é uma versão mais avançada do modelo de classificação C4.5 de Quinlan que possui recursos adicionais, como o aumento e custos desiguais para diferentes tipos de erros (Kuhn e Johnson, 2013). O C5.0 divide a amostra tendo como base o atributo que resulta em maior ganho de informação em cada nível da árvore. A árvore construída pelo C5.0 é uma estrutura de dados recursiva formada por um nó-folha que corresponde a uma classe ou por um nó de decisão que contém um teste sobre algum atributo (Mangialardo e Duarte, 2015).

O SVM compreende um conjunto de técnicas de aprendizado supervisionado, proposto por Cortes e Vapnik (1995), baseado na utilização de hiperplanos para a separação ideal entre as classes de um conjunto de dados (Hastie et al., 2009), maximizando a margem entre os pontos das duas classes mais próximas, chamados “vetores de suporte” e levando a uma melhor

probabilidade de generalização (Ließ et al., 2016). O desempenho da generalização do SVM depende de uma boa configuração dos hiperparâmetros (Cherkassky e Ma, 2004). Os parâmetros usados no ajuste do modelo no pacote *Caret* (Kuhn et al., 2020) inclui a penalidade (custo) que controla o trade-off entre erros de margem e erros de treinamento e a largura do kernel (sigma) que controla o grau de não linearidade do modelo (Naghibi et al., 2017).

O RF foi desenvolvido por Breiman (2001) como extensão do programa CART (Classification and Regression Trees), consiste em uma técnica não paramétrica que combina previsões feitas por múltiplas árvores de decisões, onde cada árvore é gerada baseada nos valores de um conjunto independente de vetores aleatórios (Tan et al., 2009). Cada um destes conjuntos é criado por um tipo de amostragem chamado de *bootstrap* (Han et al., 2011). Para tanto são definidos três parâmetros: o número de árvores (*ntree*), o número mínimo de dados em cada nó terminal (*nodesize*) e o número de variáveis utilizadas em cada árvore (*mtry*) (Liaw e Wiener, 2002). O *mtry* é o único parâmetro que requer julgamento especial (Breiman, 2002), o qual é utilizado para a otimização do modelo no pacote *Caret*.

2.6 Avaliação do desempenho

A otimização de cada hiperparâmetro dos modelos foi testada utilizando 5 valores (*tuneLenght*) definidos aleatoriamente pelo pacote Caret, sendo avaliados pela validação cruzada 5 fold. Para cada modelo, o processo foi repetido 100 vezes com seu próprio subconjunto de variáveis e comparado pelos valores médios dos parâmetros de precisão. O processo de várias repetições é importante para determinar a variabilidade da previsão já que diferentes grupos de conjuntos de dados de treinamento e validação podem gerar resultados de precisão diferentes (Kuhn e Johnson, 2013).

Para avaliar o desempenho dos algoritmos, foi utilizada a matriz de confusão, a partir da qual foi derivada a acurácia e o índice Kappa (κ). A acurácia é calculada pela soma total dos pixels corretamente classificados divido pelo número total de pixels da matriz de confusão. O índice Kappa consiste em uma medida da precisão da classificação, responsável pelo acordo de chance, descrito na equação 1.

$$K = \frac{N\sum_{i=1}^{r} X_{ii} - \sum_{i=1}^{r}\left(X_{i+} * X_{+i}\right)}{N^2 - \sum_{i=1}^{r}\left(X_{i+} * X_{+i}\right)} \quad \text{(Eq 1)}$$

em que: K= índice de exatidão Kappa; r = número de linhas da matriz; X_{ii} = número de observações na linha i e coluna i; X_{i+} e X_{+i} = totais marginais da linha i e coluna i, respectivamente; N = número total de observações.

A contabilização do acordo de chance é considerada importante para lidar com classes altamente desequilibradas, pois a alta precisão da classificação pode resultar da classificação de todas as observações como a maior classe. Valores de κ maiores que 0,80 representam concordância forte, valores entre 0,4 e 0,8 representam concordância moderada e valores abaixo de 0,4 representam baixa concordância (Congalton e Green, 1998).

3. Resultados e Discussão

A Tabela 2 apresenta o desempenho dos três modelos preditivos para o conjunto de dados desbalanceados e balanceados utilizando os dois métodos de reamostragem. Os resultados obtidos mostram a baixa precisão do conjunto de dados desbalanceados para todos os modelos, o que indica uma baixa concordância entre os valores observados e previstos, ressaltando a premissa de que o desbalanceamento das classes afeta negativamente os resultados do aprendizado de máquina.

Tabela 2. Desempenho dos modelos para os conjuntos de dados desbalanceados e balanceados

Modelo	**Conjunto de Dados**	**Dados Desbalanceados**		**Dados Balanceados *(Over/Under)***		**Dados Balanceados *(SMOTE)***	
		Acurácia	Kappa	Acurácia	Kappa	Acurácia	Kappa
RF	Calibração	0.54	0.24	0.53	0.39	0.59	0.43
	Validação	0.54	0.19	0.42	0.12	0.52	0.19
C5.0	Calibração	0.39	0.09	0.40	0.24	0.46	0.28
	Validação	0.40	0.10	0.31	0.04	0.42	0.12
SVM	Calibração	0.48	0.11	0.44	0.28	0.59	0.44
	Validação	0.49	0.06	0.38	0.05	0.47	0.10

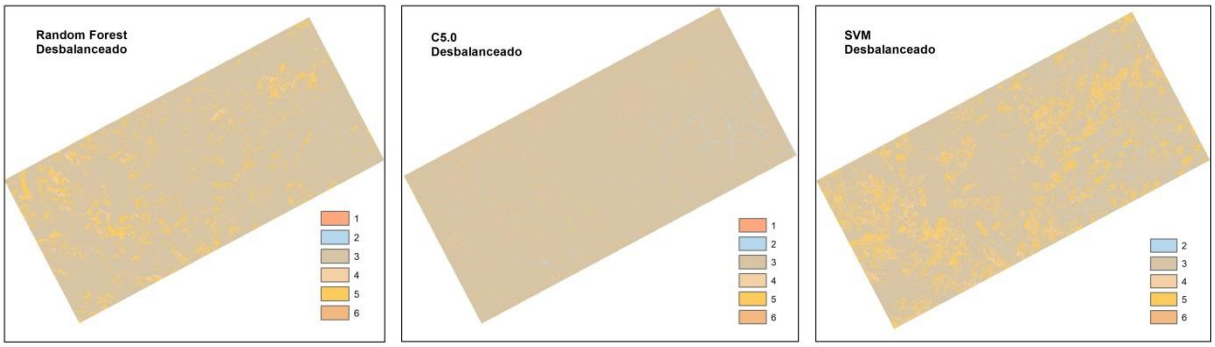
Prati et al. (2015) ao trabalharem com dados desbalanceados observaram que todos os classificadores avaliados, como C4.5, RNA e SVM, foram afetados pelo problema de desequilíbrio de classe. Sharififar et al. (2019) ao abordarem a questão do mapeamento digital de classes de solo com conjunto de dados desbalanceados, encontraram valores de Kappa para o modelo RF de -0,04 e -0,01 para os conjuntos de dados de calibração e validação, respectivamente.

Observa-se também uma menor precisão para os dados de validação nos três modelos avaliados, tanto para o conjunto de dados desbalanceados quanto para o conjunto de dados balanceados. Este fato pode estar relacionado com a presença das classes minoritárias, que possuem poucos números de observações, e que no processo de validação sofrem uma redução maior, haja vista que inicialmente o conjunto de dados é dividido para treinamento e teste. Sharififar et al. (2019) também observaram baixos valores de Kappa para validação, os autores apontam que isso não significa que o desempenho não seja aprimorado, mas que o tamanho pequeno do conjunto de dados de validação pode não ser capaz de mostrar o número real de previsões corretas das classes.

Por conseguinte, diferentemente dos resultados encontrados no presente estudo, Valadares et al. (2019) ao modelarem unidades de mapas de solo, observaram que o treinamento em bancos de dados desbalanceados superou o treinamento em bancos de dados balanceados, mostrando que neste caso não houve necessidade de balanceamento de classe, cujo algoritmo RF apresentou bom desempenho para predição das classes de solo e superou algoritmos como Árvores de Decisão únicas, RNA e Classificação Bayesiana.

A Figura 1 apresenta como o desbalanceamento de classes afeta na classificação dos solos. Observa-se que desbalanceamento pode resultar na perda de classes de solos no mapa, como constatado para o SVM desbalanceado, no qual ocorreu a perda da classe minoritária 1. Enquanto que a previsão de classes para o C5.0 com conjunto de dados desbalanceados acarretou em uma maior taxa de erro da classificação, com uma superestimação da classe majoritária 3 (dominante). Japkowicz e Stephen (2002) ao realizarem um estudo sobre desbalanceamento de classes observaram também que o algoritmo C5.0 é o mais afetado na presença de dados desbalanceados por trabalhar de forma global e não se atentar a pontos específicos.

Figura 1. Mapas produzidos pelos diferentes modelos utilizando o conjunto de dados desbalanceados.



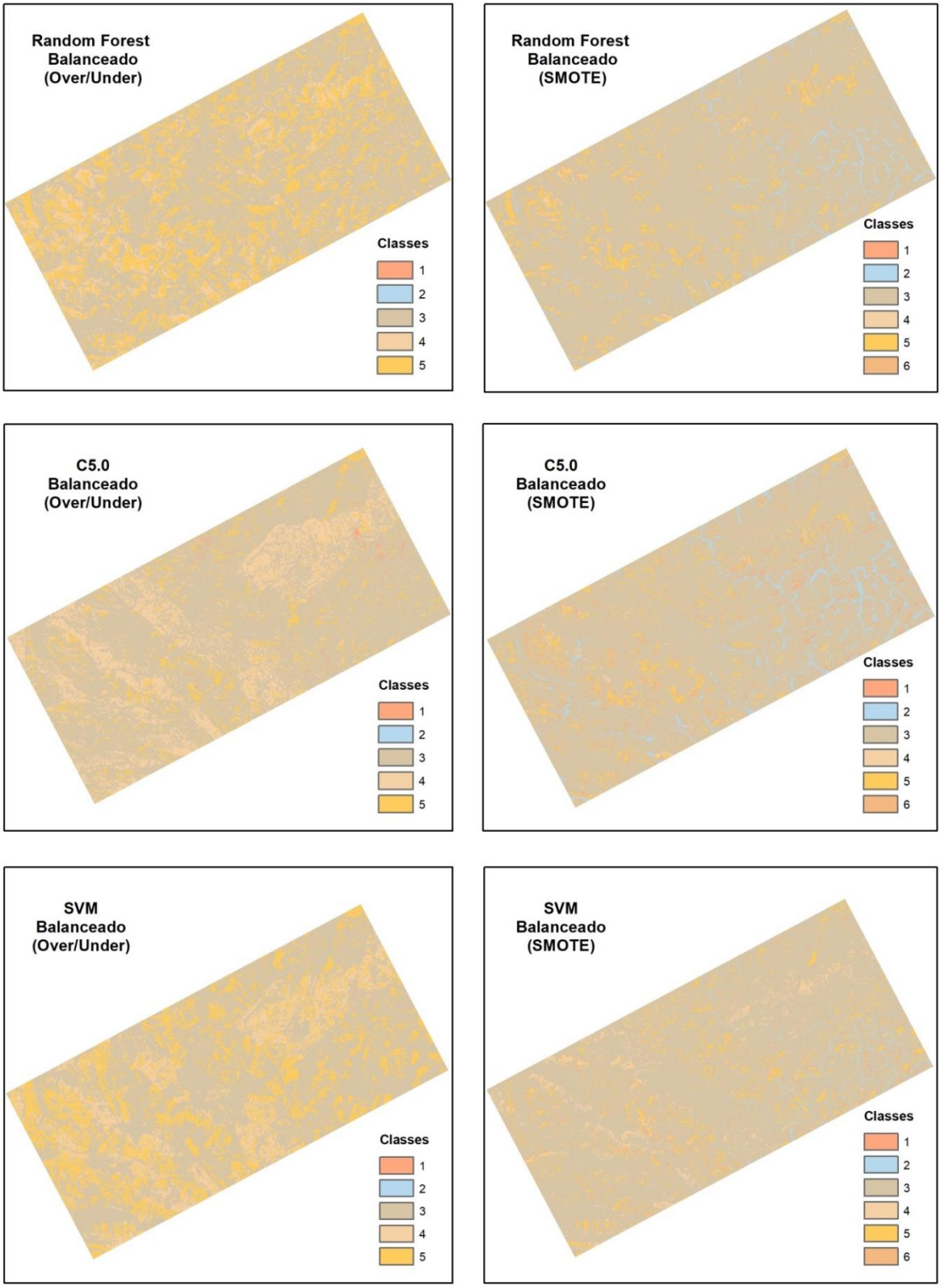
Legenda: Convenção de cores para mapas de solos - 2º nível categórico (EMBRAPA, 2018).

López et al. (2013) acrescentam que em casos mais extremos, um único exemplo mal classificado da classe minoritária pode criar uma queda significativa no desempenho. Vale descatar ainda que os valores de acurácia observados para o conjunto de dados desbalanceados (Tabela 2) evidencia o efeito da presença de que uma classe muito expressiva pode acarretar em um acerto intermediário ao acaso por parte do classificador.

De modo geral, os métodos de reamostragem contribuíram para melhorar os modelos preditivos em uma extensão razoável, conforme apresentado na Tabela 2. O conjunto de dados balanceados pelo método *SMOTE* apresentou melhor desempenho em comparação com o método *Oversampling* e *Undersampling*. Ao utilizar o método *SMOTE* foi possível obter valores moderados de índice Kappa para conjunto de dados de calibração, onde os modelos SVM e RF apresentaram melhor desempenho em comparação com o C5.0. Dias et al. (2016) ao aplicar o balanceamento de classes para a predição da distribuição de classes de solo observou também que o RF gerou o modelo de maior acurácia e teve desempenho superior ao dos algoritmos J48 e MLP.

A Figura 2 compara os mapas produzidos por ambos os métodos de reamostragem, nos quais observa-se que o método *Oversampling* e *Undersampling* omite a classe minoritária 6 nos três modelos avaliados. Sharififar et al. (2019) apontam que é difícil garantir que todas as classes sejam incluídas nos conjuntos de dados de calibração e validação sem omissão da(s) classe(s) minoritária(s). Logo, ao lidar com problemas de várias classes, pode-se facilmente perder desempenho em uma classe enquanto tentamos obtê-lo em outra (Fernández et al, 2013).

Figura 2. Mapas produzidos pelos diferentes modelos utilizando os dois métodos de reamostragem de dados.



Legenda: Convenção de cores para mapas de solos - 2º nível categórico (EMBRAPA, 2018).

Através da utilização do método *SMOTE* foi possível obter os melhores resultados, o qual está relacionado com a criação de novas amostras sintéticas usando a abordagem de k-vizinho mais próximo, com isso evita-se a perda de informação e o problema de *overfitting* dos modelos é atenuado. Além disso, com o *SMOTE* foi possível à manutenção das classes minoritárias e a diminuição da incerteza da classificação.

O desempenho do SVM foi semelhante ao do RF ao aplicar o *SMOTE*, contudo, neste caso o SVM possui como vantagem o fato de encontrar um limite de decisão no hiperplano que melhor divide os exemplos entre as classes de solos. O paradigma representativo de aprendizado baseado em kernel pode fornecer resultados de classificação relativamente robustos quando aplicados aos conjuntos de dados desequilibrados (Japkowicz e Stephen, 2002).

4. Conclusões

O presente estudo abordou o desbalanceamento de classes de solos para estudos em MDS e a aplicação de técnicas para solucionar este problema. O conjunto de dados desequilibrados apresentou baixa precisão. O método de reamostragem *SMOTE* aplicado para os modelos SVM e RF resultou em uma melhoria no desempenho dos modelos. Contudo, este estudo serve como base e direcionamento para estudos futuros, ainda são necessárias novas pesquisas relacionadas a conjunto de dados desbalanceados, particularmente para o MDS, onde devem ser consideradas técnicas alternativas para solucionar esta questão, a fim de obter uma compreensão mais aprofundada neste campo.

5. Bibliografia

Bivand R, Krug R, Neteler M, Jeworutzki S. 2019. rgrass7: Interface Between GRASS 7 Geographical Information System and R. R package version 0.2-1.

Breiman L. Random Forests. Machine Learning. 2001; 45:5-32.

Breiman L. Manual on setting up, using, and understanding random forests v3.1. Statistics Department University of California Berkeley, CA, USA. 2002

Breiman L, Cutler A, Liaw A, Wiener M. 2018. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.6-14.

Brenning A. 2008. Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models. In: Böhner, J., Blaschke, T., Montanarella, L. (eds.), SAGA – Seconds Out (= Hamburger Beiträge zur
Physischen Geographie und Landschaftsökologie) 19, 23–32.

Chawla NV, Bowyer KW, Hall LO. Kegelmeyer WP. Smote: Técnica de sobre-amostragem de minoria sintética. Journal of Artificial Intelligence Research. 2002; 16:321-357.

Cherkassky V, Ma Y. Practical selection of SVM parameters and noise estimation for SVM regression. Neural Netw. 2004; 17:113-126. https://doi.org/10.1016/S0893-6080(03)00169-2

Codemig - Companhia de desenvolvimento econômico de Minas Gerais – Mapa Geológico. 2014. Available at https://www.portaldageologia.com.br

Congalton R, Green K. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Boca Raton: CRC/Taylor & Francis, 2009. 183p.

Cortes C, Vapnik V. Support-vector networks, Mach. Learn. 1995, 20:273-297.

Dal Pozzolo A, Caelen O, Bontempi G. 2015. unbalanced: Racing for Unbalanced Methods Selection. R package version 2.0.

Dias LMS, Coelho RM, Valladares GS, Assis ACC, Ferreira EP, Silva RC. Predição de classes de solo por mineração de dados em área da bacia sedimentar do São Francisco. Pesq. agropec. bras. 2016; 51: 1396-1404. https://doi.org/10.1590/s0100-204x2016000900038

Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA. Sistema Brasileiro de Classificação de Solos. Rio de Janeiro: Embrapa Solos, 2018.

Fernández A, García S, Herrera F. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research. 2018; 61:863-905. https://doi.org/10.1613/jair.1.11192

Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Trans. Syst., Man, Cybern. C, Appl. Rev. 2012; 42:463–484. https://doi.org/10.1109/TSMCC.2011.2161285

Garcia LC. Fenologia de espécies da canga em Barão de Cocais, Quadrilátero Ferrífero de Minas Gerais [dissertation]. Belo Horizonte: Universidade Federal de Minas Gerais; 2007.

García S, Herrera F. Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy Evolutionary Computation. 2009; 17:275-306. https://doi.org/10.1162/evco.2009.17.3.275

García S, Luengo J, Herrera F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowledge-Based Systems. 2016; 98: 1–29. https://doi.org/10.1016/j.knosys.2015.12.006

Gopi SC, Suvarna B, Padmaja TM. High Dimensional Unbalanced Data Classification Vs SVM Feature Selection. Indian Journal of Science and Technology. 2016. https://doi.org/10.17485/ijst/2016/v9i30/98729

Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3ed. San Francisco: Morgan Kaufmann Publishers, 2011.

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer, New York, 2009, 745p.

He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 2009; 21:1263-1284. https://doi.org/10.1109/TKDE.2008.239

IBGE. 2001. Base cartográfica vetorial contínua do Brasil ao milionésimo - BCIM. IBGE, Rio de Janeiro.

Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Inteligent Data Analysis. 2002; 6:429-450. https://doi.org/10.3233/IDA-2002-6504

Jenny H. Factors of soil formation. A system of quantitative pedology [Internet]. Dover Publ. 1941.

Krawczyk B. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence. 2016; 5: 221–232. https://doi.org/10.1007/s13748-016-0094-0

Kuhn M, Johnson K. Applied predictive modeling. New York, NY: Springer, 2013.

Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, Benesty M, Lescarbeau R, Ziem A, Scrucca L, Tang Y, Candan C, Hunt T. 2020a. caret: Classification and Regression Training. R package version 6.0-85.

Kuhn M, Weston S, Culp M, Coulter N, Quinlan R. 2020b. C50: C5.0 Decision Trees and Rule-Based Models. R package version 0.1.3.

Lagacherie P, McBratney AB. Chapter 1 Spatial Soil Information Systems and Spatial Soil Inference Systems: Perspectives for Digital Soil Mapping. Dev Soil Sci. 2007; 31:3–22. https://doi.org/10.1016/S0166-2481(06)31001-X

Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002; 2:18-22.

Ließ M, Schmidt J, Glaser B. Improving the spatial prediction of soil organic carbon stocks in a complex Tropical Mountain landscape by methodological specifications in machine learning approaches. PLoS One. 2016; 11:1-22. https://doi.org/10.1371/journal.pone.0153673

Liu XY, Wu JX, Zhou ZH. Exploratory under sampling for class imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2009; 39:539–550. https://doi.org/10.1109/TSMCB.2008.2007853

López V, Fernández A, García S, Palade V, Herrera F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences. 2013; 250: 113-141. https://doi.org/10.1016/j.ins.2013.07.007

Lu WZ, Wang D. Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme. Sci. Total. Enviro. 2008; 395:109-116. https://doi.org/10.1016/j.scitotenv.2008.01.035

Ma Y, Minasny B, Malone BP, McBratney AB. Pedology and digital soil mapping (DSM). European Journal of Soil Science. 2019; 70:216–235. https://doi.org/10.1111/ejss.12790

Mangialardo RJ, Duarte JC. Integrating Static and Dynamic Malware Analysis Using Machine Learning. IEEE Latin America Transactions. 2015; 13: 3080-3087. https://doi.org/ 10.1109/TLA.2015.7350062

Mazurowski MA, Habas PA, Zurada JM, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Netw. 2008; 21:427-436. https://doi.org/10.1016/j.neunet.2007.12.031

McBratney AB, De Gruijter JJ. A continuum approach to soil classification by modified fuzzy k-means with extragrades. Journal of Soil Sciencev. 1992; 43:159-175. https://doi.org/10.1111/j.1365-2389.1992.tb00127.x

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang CC, Lin CC. 2019. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-3.

Naghibi SA, Ahmadi K, Daneshi A. Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resources Management. 2017; 31:2761–2775. https://doi.org/10.1007/s11269-017-1660-3

Ng WYW, Hu J, Yeung DS, Yin S, Roli F. Diversified Sensitivity-Based Undersampling for Imbalance Classification Problems. IEEE Transactions on Cybernetics. 2015; 45: 2402 – 2412 . https://doi.org/10.1109/TCYB.2014.2372060

Prati RC, Batista G, Silva DF. Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. Knowl Inform Syst. 2015; 45:247–70. https://doi.org/10.1007/s10115-014-0794-3

Robert JH. 2019. raster: Geographic Data Analysis and Modeling. R package version 2.8-19.

Santos RD, Lemos RC, Santos HG, Ker JC, Anjos LHC, Shimizu SH. Manual de descrição e coleta de solo no campo. 6.ed. rev. e ampl. Viçosa: SBCS, 2013. 100p.

Sharififar A, Sarmadian F, Malone BP, Minasny B. Addressing the issue of digital mapping of soil classes with imbalanced class observations. Geoderma. 2019; 350:84-92. https://doi.org/10.1016/j.geoderma.2019.05.016

Tan PN, Steinbach M, Kumar V. Introdução ao Data Mining. Rio de Janeiro: Editora Ciência Moderna Ltda, 2009. 896 p.

Universidade Federal de Viçosa (UFV), Fundação Centro Tecnológico de Minas Gerais (CETEC-MG), Universidade Federal de Lavras (UFLA), Fundação Estadual do Meio Ambiente (FEAM). 2010. Mapa de solos do Estado de Minas Gerais: legenda expandida. Fundação Estadual do Meio Ambiente, Belo Horizonte. Available at http://www.feam.br/noticias/1/1355-mapa-de-solos (accessed 05 Jun. 2019)

Valadares AP, Coelho RM, Oliveira SRM. Preprocessing procedures and supervised classification applied to a database of systematic soil survey. Scientia Agricola. 2019; 76: 439-447. http://dx.doi.org/10.1590/1678-992X-2017-0171

Varajão, C.A.C., Salgado, A.A.R., Varajão, A.F.D.C., Braucher, R., Colin, F., Nalini Júnior, H.A., 2009. Estudo da evolução da paisagem do Quadrilátero Ferrífero (Minas Gerais, Brasil) por meio da mensuração das taxas de erosão (10be) e da pedogênese. Revista Brasileira de Ciência Solo. 33, 1409-1425. http://dx.doi.org/10.1590/S0100-06832009000500032

Wang S, Yao X. Multiclass imbalance problems: Analysis and potential solutions. IEEE Transactions on Systems, Man, and Cybernetics, Part B(Cybernetics). 2012; 42:1119-1130. https://doi.org/10.1109/TSMCB.2012.2187280

CONCLUSÃO GERAL

O mapeamento digital de solos encontra-se em constante desenvolvimento dado o avanço na tecnologia da informação, maior capacidade de processamento dos computadores e maior disponibilidade aos dados de alta resolução espacial. O Brasil tem contribuído de forma efetiva para as pesquisas em MDS em um panorama global, apesar de existirem diversas lacunas a serem solucionadas. Com a criação do Programa Nacional de Solos do Brasil (PronaSolos), emergiu a necessidade substancial para o desenvolvimento de novas estratégias metodológicas para os levantamentos e mapeamentos de solos no País. Os principais obstáculos e dificuldades para a realização do mapeamento de solos estão associados, principalmente, ao desenvolvimento de técnicas de amostragem de solos, que consiste em uma das etapas mais onerosas do MDS. Neste sentido, o presente estudo aplicou e avaliou novas estratégias metodológicas, como a utilização do método cLHS modificado para amostragem de solos e métodos de reamostragem para lidar com conjunto de dados desbalanceados. O cLHS possibilitou a realização da amostragem em campo de forma eficiente, de modo que aumentou a possibilidade de acesso aos locais de amostragem e assegurou a redução da demanda de custo e tempo. O presente trabalho apresentou também uma visão da natureza do conjunto de dados desbalanceados e como este problema afeta o desempenho dos modelos de predição de solos; além de discutir os desafios abertos e direções futuras no MDS. Dessa forma, os resultados fornecidos neste trabalho podem fornecer uma base para estudos futuros. Contudo, a busca pelo desenvolvimento de técnicas e métodos deve ser continuada, ainda são necessárias novas pesquisas considerando diferentes cenários ambientais a fim de subsidiar os futuros desafios de mapeamento de solos no País.