UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE AGRONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DO SOLO

MAPEAMENTO DIGITAL DE CLASSES DE SOLOS NAS BACIAS DOS RIOS SANTO CRISTO E LAJEADO GRANDE

ALCINEI RIBEIRO CAMPOS

(Tese de doutorado)

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE AGRONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DO SOLO

MAPEAMENTO DIGITAL DE CLASSES DE SOLOS NAS BACIAS DOS RIOS SANTO CRISTO E LAJEADO GRANDE

ALCINEI RIBEIRO CAMPOS
Mestre em Agronomia – Solos e Nutrição de Plantas (UFPI)
Engenheiro Agrônomo (UFPI)

Tese apresentada como um dos requisitos à obtenção do Grau de Doutor em Ciência do Solo

Porto Alegre (RS) Brasil
Julho de 2018

CIP - Catalogação na Publicação

Ribeiro Campos, Alcinei
MAPEAMENTO DIGITAL DE CLASSES DE SOLOS NAS BACIAS DOS RIOS SANTO CRISTO E LAJEADO GRANDE / Alcinei Ribeiro Campos. -- 2018.
102 f.
Orientadora: Elvio Giasson.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Agronomia, Programa de Pós-Graduação em Ciência do Solo, Porto Alegre, BR-RS, 2018.

1. Mapeamento digital de solos. 2. Modelos preditores. 3. Variáveis preditoras. 4. Levantamento de solos. I. Giasson, Elvio, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

ALCINEI RIBEIRO CAMPOS
Engenheiro Agrônomo - UFPI
Mestre em Agronomia - Solos e Nutrição de Plantas - UFPI

TESE

Submetida como parte dos requisitos
para obtenção do Grau de

DOUTOR EM CIÊNCIA DO SOLO

Programa de Pós-Graduação em Ciência do Solo
Faculdade de Agronomia
Universidade Federal do Rio Grande do Sul
Porto Alegre (RS), Brasil

Aprovado em: 30.07.2018
Pela Banca Examinadora

ELVIO GIASSON
Orientador-PPG Ciência do Solo

PAULO CÉSAR DO NASCIMENTO
Departamento de Solos/UFRGS

IVAN LUIZ ZILLI BACIC
EPAGRI/SC

ALEXANDRE TEN CATEN
UFSC

Homologado em: 01.02.2019
Por

CARLOS GUSTAVO TORNQUIST
Coordenador do
Programa de Pós-Graduação em
Ciência do Solo

CARLOS ALBERTO BISSANI
Diretor da Faculdade
de Agronomia

Dedico este trabalho as meus Pais e
Professores que foram meus guias
ao longo dessa jornada.

AGRADECIMENTOS

À Deus por permitir a conclusão de mais essa etapa;

À minha família pelo apoio em mais essa conquista;

À Universidade Federal do Rio Grande do Sul e ao Programa de Pós-Graduação em Ciência do Solo, pela oportunidade de realização do curso de Doutorado;

Ao meu orientador Elvio Giasson, pela orientação, amizade e confiança;

À todos os professores do PPGCS/UFRGS, pelas importantes contribuições em mais essa fase da minha formação acadêmica.

À CAPES e CNPq pelo apoio financeiro e pela concessão da bolsa de estudos;

Aos amigos do PPGCS/UFRGS, em especial ao grupo de Gênese e Classificação de Solos.

MAPEAMENTO DIGITAL DE CLASSES DE SOLOS NAS BACIAS DOS RIOS SANTO CRISTO E LAJEADO GRANDE[1]/

Autor: Alcinei Ribeiro Campos
Orientador: Prof. Elvio Giasson

RESUMO

O Mapeamento Digital de Solos (MDS) tem ganhado destaque como uma alternativa às abordagens tradicionais empregadas nos levantamentos de solos, entretanto, ainda não possui uma metodologia definida. Dentre os aspectos envolvidos no MDS, ainda não há recomendação de métodos eficientes para seleção das variáveis mais relevantes, nem técnicas que permitam aumentar a eficiência no uso de perfis de solos georreferenciados na predição dos solos, assim, a presente tese teve como objetivos gerais estudar técnicas que podem ser aplicadas para aumentar a eficiência das metodologias empregadas no MDS. A tese está dividida em três estudos. Os estudos foram realizados nas bacias dos rios Santo Cristo e do Lajeado Grande, noroeste do Rio Grande do Sul. O Estudo 1 avaliou três métodos de seleção de variáveis preditoras, aplicados em 40 variáveis ambientais buscando identificar as variáveis mais relevantes para predição da ocorrência dos solos, assim como do método mais eficiente para seleção destas variáveis preditoras. Neste estudo concluiu-se que a seleção recursiva *wrapper* selecionou o subconjunto de variáveis com maior eficiência na predição da ocorrência dos solos. O segundo estudo avaliou as variáveis preditoras em múltiplos níveis de suavização, para isso foram aplicados diferentes tamanhos de filtro de média no modelo digital de elevação a partir dos quais foram geradas as variáveis preditoras. Neste estudo concluiu-se que a aplicação dos filtros com tamanhos 20x20, 25x25 e 30x30 resultou em variáveis com maior eficiência na predição dos solos. O terceiro estudo avaliou o uso de *buffer* para coleta de amostras vizinhas aos perfis de solos georreferenciados disponíveis nas áreas de estudo. Foram testados cinco raios de *buffers* para coleta dos pixels amostrais. Neste estudo concluiu-se que a utilização dos pixels amostrais coletados nos *buffers* não alterou de forma expressiva a acurácia geral dos mapas preditos na bacia do rio Lajeado Grande, mas permitiu um ganho de 15,6% de concordância no mapa predito da bacia do rio Santo Cristo. Como conclusão geral constatamos que os procedimentos metodológicos testados aumentaram o desempenho das técnicas utilizadas na predição de ocorrência dos solos e podem ser utilizadas em áreas com disponibilidade de dados de solos na forma de mapas legados ou perfis georreferenciados.

[1]/ Tese de Doutorado em Ciência do Solo. Programa de Pós-Graduação em Ciência do Solo, Faculdade de Agronomia, Universidade Federal do Rio Grande do Sul. Porto Alegre. (86 p.) Julho, 2018. Trabalho realizado com apoio financeiro da CAPES/CNPq

DIGITAL SOIL CLASS MAPPING IN THE WATERSHEDS OF THE RIVERS SANTO CRISTO AND LAJEADO GRANDE[1]/

Author: Alcinei Ribeiro Campos
Adviser: Prof. Elvio Giasson

ABSTRACT

Digital Soil Mapping (DSM) has gained prominence as an alternative to the traditional approaches employed in soil surveys, however, it still does not have a defined methodology. Between the aspects involved in the MDS, still does not recommend efficient methods in the selection of the most relevant variables, nor techniques that allow a more efficient use of the georeferenced soil profiles in soil prediction, so the present thesis had as general objectives to study techniques that can be applied to increase the efficiency of the methodologies used in the DSM. The thesis is divided into three studies. The studies were performed in watersheds of the rivers Santo Cristo and Lajeado Grande. The first study evaluated three predictive variable selection methods, applied to 40 environmental variables. The *wrapper* recursive selection was found to select the subset of variables more efficiently. The second study assessed the predictive variables in multiple levels of smoothing by applying different sizes of average filter in the digital elevation model. The use of the filter 20x20, 25x25 and 30x30 resulted in more efficient variables in soil prediction. The third study evaluated the use of *buffers* to collect samples neighboring the georeferenced soil profiles neighbors available. Five buffer radii were tested in the collection of sample pixels. The conclusion was that using the sample pixels collected in the *buffers* did not alter significantly the general accuracy of the predicted maps in the Lajeado Grande river basin, but it resulted in a gain of 15.6% agreement on the predicted map of Santo Cristo watershed. As a general conclusion we verified that the methodological procedures tested increased the performance of the techniques used in predicting soil occurrence and can be used in areas with availability of soil data in the form of legacy maps or georeferenced profiles.

[1]/ Doctoral thesis in Soil Science, Graduate Program in Soil Science, Faculty of Agronomy, Federal University of Rio Grande do Sul, Porto Alegre. (86 p.) July, 2018. Research supported by CNPq and CAPES

SUMÁRIO

Página

RELAÇÃO DE TABELAS

Página

RELAÇÃO DE QUADROS

Página

RELAÇÃO DE FIGURAS

Página

RELAÇÃO DE ABREVIATURAS

AA	Amostragem Aleatória
AD	Árvores de Decisão
AE	Amostragem Estratificada
AG	Acurácia Geral
AM	Acurácia do Mapeador
ASTER/GDEM	*Advanced Spaceborne Thermal Emission and Reflection Radiometer/ Global Digital Elevation Model*
CFS	Correlation-based Feature Selection
CSE	Consistency Subset Evaluation
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
FP	Falso Positivo
FN	Falso Negativo
IBGE	Instituto Brasileiro de Geografia e Estatística
LG	Bacia hidrográfica do rio Lajeado Grande
MAE	Erro Médio Quadrático
MCC	Coeficiente de Correlação de Matthews
MDE	Modelo Digital de Elevação
MDS	Mapeamento Digital de Solos
PREC	Precisão
REC	*Recall*
RF	Random Forest
RMS	Erro médio quadrado
RMSE	Raiz Quadrada do Erro Médio
RNA	Redes Neurais Artificiais
SiBCS	Sistema Brasileiro de Classificação de Solos
SC	Bacia hidrográfica do rio Santo Cristo
SIG	Sistemas de Informações Geográficas
SNPA	Serviço Nacional de Pesquisas Agronômicas
TN	Verdadeiro Negativo
TP	Verdadeiro Positivo
UM	Unidade de Mapeamento

1. INTRODUÇÃO

Os levantamentos de solos são indispensáveis para o planejamento da ocupação e uso das terras e sustentabilidade do meio ambiente. A aplicabilidade dos mapas de solos para a gestão dos recursos naturais depende da sua escala, que deve se aproximar ao máximo do fenômeno a ser representado, que pode ser variações nas propriedades ou classes dos solos (DALMOLIN et al., 2004; FIGUEIREDO et al., 2008; BEHRENS et al., 2014).

Atualmente no Brasil, a maioria dos levantamentos de solos existentes disponibilizam mapas em pequenas escalas, inadequadas para o planejamento do uso e manejo de solos (CARVALHO; NUNES; ANTUNES, 2013; SANTOS et al., 2013). A deficiência de mapas de solos em escalas maiores é explicada em grande parte pela falta de investimento em pesquisas nessa área. Além disso o método de levantamento tradicional utilizado demanda muito tempo, trabalho e recursos humanos com elevado conhecimento técnico das relações solo-paisagem, fatores que tem limitado a produção destas informações.

As dificuldades para execução de levantamentos de solos pelos métodos tradicionais colaboraram para o surgimento de novas abordagens nos levantamentos de solos, dentre elas, o Mapeamento Digital de Solos (MDS). O MDS teve sua maior evolução associado aos avanços alcançados nos últimos anos na indústria espacial, com a coleta de dados da superfície terrestre juntamente com o desenvolvimento de equipamentos e programas computacionais que permitiu a criação de métodos de mapeamento baseados em modelos matemáticas (MCBRATNEY et al., 2003).

O MDS é um método fundamentado no uso de equações de predição, baseadas no modelo *SCORPAN* (MCBRATNEY et al., 2003), que pode fornecer informações sobre classes ou propriedades dos solos.

A maioria dos estudos realizados no MDS tem buscado através de modelos digitais de predição a reprodução de mapas legados, permitindo assim, a criação de modelos matemáticos digitais que podem reproduzir o modelo utilizado pelo pedólogo para delimitar as classes de solos, aumentando a capacidade de extrapolação desses modelos para outras áreas fisiograficamente semelhantes (GRINAND et al., 2008; GIASSON et al., 2013; BAGATINI; GIASSON; TESKE, 2016).

Os principais componentes do MDS são os modelos de predição, os esquemas de amostragem, as variáveis preditoras e a variável resposta (classes ou propriedades de solos), sendo que a qualidade dos produtos gerados no MDS depende diretamente da combinação destes componentes. Diante disso, estudos vêm sendo realizados objetivando elucidar a contribuição dos componentes do MDS na predição de classes de solo. No entanto, como pode ser constatado na literatura, há uma grande variabilidade na combinação dos componentes envolvidos no MDS, consequentemente, ainda não existe um padrão nas metodologias empregadas, o que resulta em diferentes eficiências das técnicas utilizadas e qualidade dos produtos gerados pelo MDS. Neste contexto, são necessárias mais pesquisas para desenvolver novas metodologias eficientes ou reforçar os bons resultados obtidos com algumas metodologias já testadas em outros estudos (ABDEL-KADER, 2011; TEN CATEN et al., 2011a; BAGATINI; GIASSON; TESKE, 2015; FRANCO et al., 2015).

A disponibilidade e qualidade da variável resposta são importantes para treinamento e validação dos modelos preditores, assim, ganham destaque às poucas áreas no Brasil que possuem mapas de solos com grande escala de detalhamento, sendo que o Rio Grande do Sul possui áreas mapeadas em escala igual ou maior que 1:50.000 (GIASSON et al., 2011), e permitem a validação das metodologias aplicada no MDS com informações dos solos de elevado nível de detalhe.

Portanto, devido à demanda de mapas de solos e da necessidade de mais estudos para desenvolvimento e validação das metodologias utilizadas no MDS, este trabalho de tese teve como objetivo geral estudar os componentes que influenciam a acurácia final dos mapas obtidos pelo método do Mapeamento Digital de Solos, sendo estruturada da seguinte forma:

Capitulo I - Revisão bibliográfica sobre o atual cenário do Brasil quanto à disponibilidade de mapas de solos e do quadro atual de desenvolvimento e aplicação do Mapeamento Digital de Solos, buscando entender os fatores que influenciam o desempenho das metodologias aplicadas no MDS.

Capitulo II – Estudo 1 - Seleção de variáveis ambientais para treinamento de algoritmos preditores aplicados no Mapeamento Digital de Solos. Diante do grande número de variáveis ambientais que podem ser utilizadas no MDS, a seleção de variáveis eficientes na discriminação das classes de solos é indispensável para o máximo desempenho dos modelos preditores, assim, o Estudo 1 teve como objetivos gerais avaliar a relevância de variáveis ambientais utilizadas como preditoras e o desempenho de algoritmos aplicados na seleção destas variáveis ambientais para uso na predição de ocorrência das classes de solos;

Capitulo III – Estudo 2 - Avaliação de variáveis preditoras geradas de modelos digitais de elevação suavizados na predição de ocorrência de solos. Na predição de ocorrência dos solos as variáveis preditoras devem apresentar escala adequada a representar as variações das classes de solos na paisagem, nesse contexto, sem sempre as variáveis no seu maior nível de detalhamento serão as mais eficientes para predição de ocorrência dos solos, assim, o Estudo 2 buscou entender o efeito da aplicação de filtro de média para suavização do modelo digital de elevação e a eficiência das variáveis geradas a partir destes MDEs na predição de ocorrência das classes de solos.

Capitulo IV – Estudo 3 - Predição de classes de solos com dados coletados em pixels delimitados por *buffers* em perfis de solo georreferenciados. A maioria dos estudos de MDS tem buscado a reprodutibilidade de mapas convencionais de solos, no entanto, a exigência de um mapa legado reduz as áreas de aplicação do MDS, uma vez que sempre será necessária a realização do levantamento tradicional para produção destes mapas. Nesse contexto, o terceiro Estudo teve como objetivo testar a capacidade do Mapeamento Digital de Solos apenas com informações contidas em perfis de solos georreferenciados e em pixels vizinhos as os perfis de solos, buscando assim, reduzir a dependência de mapas legados para treinamento e validação dos modelos preditores.

2. CAPITULO I - REVISÃO BIBLIOGRÁFICA

2.1. Levantamentos de Solos no Brasil

Os mapas de solos são instrumentos imprescindíveis para o planejamento e gerenciamento correto dos recursos naturais. As informações neles contidas permitem ao poder público a proposição de políticas territoriais, a construção de instrumentos jurídico-administrativos e formulação de diretrizes para preservação e recuperação dos recursos naturais, dentre eles o solo. No Brasil, os mapas de solos começaram a ser produzidos no início na década de 1940, com o primeiro Esboço Agro-Geológico do Estado de São Paulo, publicado em 1943 (CARVALHO; NUNES; ANTUNES, 2013).

A partir de 1947, com a criação da Comissão de Solos do Serviço Nacional de Pesquisas Agronômicas (SNPA) do Ministério da Agricultura, ocorreu um aumento nos levantamentos de solos em todo território nacional (CARVALHO; NUNES; ANTUNES, 2013). Porém, assim como em outros países, a partir da década de 1980 houve redução dos recursos econômicos destinados à cartografia dos solos, causando uma drástica redução tanto da produção de mapas de solos como do quadro de técnicos capacitados para realizá-los (HARTEMINK; MCBRATNEY, 2008; EMBRAPA, 2016), consequentemente, resultando em deficiência de mapas em escala adequada ao manejo dos solos no País.

Atualmente, todo o território brasileiro possui levantamentos esquemáticos de solos, publicados em escala que varia de 1:1.000.000 a 1:5.000.000. Aproximadamente 85% do território possui mapas produzidos nos levantamentos de reconhecimento de baixa intensidade publicados em escala de 1:250.000 a 1:750.000, e menos de 8,4% do território possuí mapas produzidos nos levantamentos de média intensidade, com publicação em escala de 1:100.000 a 1:250.000, de acordo com os dados integrados

a bases de dados da Embrapa Solos e IBGE (SANTOS et al., 2013). Vale ressaltar que existem outros levantamentos realizados por outras instituições (ex. Universidades), que em função de sua extensão ou outras limitações não foram integrados nas bases de dados da Embrapa ou IBGE.

Como visto a maioria dos mapas disponíveis não apresenta nível de detalhe adequado para orientação das atividades no meio rural (EMBRAPA, 2016), refletindo a realidade do Brasil sobre a disponibilidade de informações dos solos e evidenciando a necessidades de novos levantamentos para produção destas informações. Diante do déficit de informações de solos e para atender a um Acórdão do Tribunal de Contas da União (n° 1942/2015) que versa sobre Governança de Solos, em 2016 foi criado o Programa Nacional de Solos do Brasil (PRONASOLOS) com o objetivo de aumentar o nível de conhecimento dos solos brasileiros, possibilitando sua governança por parte do poder público, valorizando o manejo sustentável dos recursos naturais, e possibilitando um desenvolvimento agropecuário sustentável (EMBRAPA, 2016).

No Brasil o principal método utilizado na produção dos mapas de solos é o levantamento pedológico convencional. Este método de levantamento é realizado por meio de prospecções nas áreas a serem mapeadas, demandando pessoal técnico experiente, cartas topográficas e fotos aéreas em escala adequada, boa disponibilidade de infraestruturas como rodovias ou estradas e são lentos e caros, o que tem dificultado sua execução para a produção de informações de solos em escalas mais detalhadas ou para extensas áreas (MCBRATNEY et al., 2003; SANTOS et al., 2013; EMBRAPA, 2016; FLACH; CORR, 2017). Dessa forma os levantamentos pedológicos convencionais, em geral, não atendem em tempo hábil as demandas atuais destas informações.

As dificuldades de execução dos levantamentos convencionais de solos podem ser confirmadas pelo tempo requerido no projeto PRONASOLOS para produção das informações dos solos, o qual estima que de 10 a 30 anos alcançará 250 mil km^2 mapeados em escala 1:25.000, um milhão de km^2 em escala 1:50.000 e 6,9 milhões de km^2 em escala 1:100.000, o que corresponde a aproximadamente 3%, 12% e 80% do território brasileiro (EMBRAPA, 2016). Este cenário indica que mesmo após execução do projeto PRONASOLOS, o Brasil ainda não disponibilizará em todo seu território de mapas pedológicos como constatados em outros países, a exemplo os EUA, que já possuem todo

seu território com mapas de solos publicados na escala entre 1:20.000 e 1:40.000 (EMBRAPA, 2016), e a Dinamarca que já disponibiliza de mapa de solo na escala 1:25.000 para maior parte do seu território (GEUS, 1998).

O cenário atual e as perspectivas de futuro para disponibilidade de mapas de solos no Brasil, mesmo com os investimentos previstos no projeto PRONASOLOS, indicam necessidade de novas abordagens para produção dos mapas de solos (FLACH; CORR, 2017). Dentre as novas abordagens podemos destacar o Mapeamento Digital de Solos (MDS), o qual foi proposto como método para fornecer mapas de classes ou propriedades dos solos em maior escala de detalhe e menor tempo de execução. O uso do MDS como ferramenta auxiliar na cartografia dos solos pode permitir uma redução nas dificuldades e no tempo demandado pelos levantamentos convencionais para produção destas informações (MCBRATNEY et al., 2003).

O uso do MDS poderá aumentar a disponibilidade de mapas de solos no Brasil, principalmente se aplicado ao projeto PRONASOLOS, no qual já há previsão de inserção de materiais e procedimentos geotecnológicos inovadores de eficácia comprovada (EMBRAPA, 2016). No entanto, o uso do MDS como método eficiente e confiável para produção de mapas de solos ou auxiliar nos levantamentos convencionais ainda sofre resistência no Brasil, sendo que os trabalhos realizados até o presente momento tiveram caráter exploratório (COELHO; GIASSON, 2010; TEN CATEN et al., 2011a; BAGATINI; GIASSON; TESKE, 2015).

A falta de uso do MDS no Brasil como ferramenta auxiliar nos levantamentos convencionais de solos reforça a importância dos estudos para desenvolvimento e validação de metodologias eficientes para produção destas informações. De modo geral, a validação das metodologias utilizadas no MDS permitirá ao Brasil seguir outros países que já começam a utilizar esta técnica como método para atualização das suas informações de solos (SUBBURAYALU; SLATER, 2013; PAHLAVAN RAD et al., 2014; SUBBURAYALU; JENHANI; SLATER, 2014).

2.2. Mapeamento digital de solos

2.2.1. Histórico e evolução

O MDS é definido como a criação, e população de sistemas de informação espacial do solo através do uso de métodos observacionais de campo e laboratório integrados a modelos matemáticos de inferência espacial e não-espacial das classes ou propriedades do solo (LAGACHERIE; MCBRATNEY; VOLTZ, 2006). O MDS foi formalizado por McBratney et al. (2003) a partir da equação $S_{(c,p)}$= f (s.c.o.r.p.a.n.) mais conhecida como modelo *SCORPAN*, onde as classes (S_c) ou propriedades do solo (S_p) são definidas em função (f) do próprio solo (s), do clima (c), dos organismos (o), do relevo (r), do material de origem (p), do tempo (a) e da localização geográfica (n).

O modelo *SCORPAN* pode ser visto como uma evolução do modelo CLORPT (clima, organismos, relevo, material de origem e tempo) proposto por Jenny em 1941 (MCBRATNEY et al., 2003; DALMOLIN; TEN CATEN, 2015), e tem com finalidade descrever, classificar e estudar os padrões de variação dos solos na paisagem através de técnicas pedométricas (MENDONÇA-SANTOS; SANTOS, 2003). Dessa forma, as informações dos solos podem ser preditas a partir dele próprio (com dados legados) associado a variáveis ambientais que caracterizem os fatores de formação constantes no modelo (MCBRATNEY et al., 2003).

Estudos vêm sendo realizados no MDS buscando a aplicação das técnicas pedométricas e desenvolvimentos de metodologias que possam produzir resultados acurados em relação aos produtos obtidos pelo método de levantamento convencional de solos (GIASSON et al., 2006; TEN CATEN et al., 2011a; AFSHAR; AYOUBI; JAFARI, 2018). A eficiência das técnicas pedométricas na predição dos solos depende diretamente da combinação de três componentes (algoritmos preditores, variáveis preditoras e esquemas de amostragem), sendo que estes componentes são integrados por meio de modelos preditores e a eficiência de cada componente influencia diretamente a qualidades dos mapas obtidos no MDS.

Nesse contexto, estudos exploratórios vêm buscando equilibrar a associação destes componentes para obtenção do máximo desempenho do

MDS na predição das informações dos solos e contribuído para o estabelecimento deste método como ferramenta para atender as demandas atuais e futuras de informações dos solos de forma rápida e econômica (GIASSON et al., 2006; ODGERS; MCBRATNEY; MINASNY, 2011; BAGATINI; GIASSON; TESKE, 2015; TERRA; DEMATTÊ; VISCARRA ROSSEL, 2018).

Atualmente, há diversos estudos publicados na predição de ocorrência de classes de solos, nos quais é constatada uma grande variabilidade na qualidade dos mapas obtidos. A acurácia geral dos mapas preditos tem variado de 30% a 95%, com valor médio inferior a 60%. Comportamento semelhante é observado para o coeficiente Kappa que apresenta valor médio inferior a 0,52 (CHAGAS; CARVALHO JÚNIOR; BHERING, 2011; PELEGRINO et al., 2016).

Estes parâmetros acurácia geral e coeficiente Kappa, são comumente utilizados para medir a concordância geral dos mapas preditos, e os resultados atuais ainda refletem a realidade observada por ten Caten et al. (2012) indicando baixa eficiência das variadas metodologias testadas para reprodutibilidade dos mapas convencionais de solos, contribuindo para o não uso do MDS com ferramenta na execução de levantamentos de solos.

2.2.2. Modelos preditores utilizados no Mapeamento Digital de Solos

Os avanços alcançados na computação, principalmente no desenvolvimento de equipamentos computacionais com capacidade de armazenamento e processamento de dados, permitiram a criação e implementação de programas computacionais com aplicação em várias ciências. Dentre eles, os programas computacionais de geoprocessamento mais conhecidos como Sistemas de Informação Geográfica (SIGs). Os SIGs permitem armazenar e manipular grandes quantidades de dados geográficos (georreferenciados), impulsionando e facilitando a cartografia dos solos (MCBRATNEY et al., 2003; MINASNY; MCBRATNEY, 2016).

Nesse contexto, foram desenvolvidos algoritmos de aprendizagem de máquina, que podem analisar e extrair padrões de grandes conjuntos de dados, permitindo uma maior integração dos conhecimentos tácitos do pedólogo sobre as relações solos-paisagem com modelos matemáticos utilizados na predição de ocorrência das classes dos solos. Dentre as técnicas matemáticas comumente

aplicadas para predição de ocorrência das classes de solos está o uso de regressões e as técnicas de mineração de dados, das quais podemos destacar as Regressões Logísticas e os mineradores de dados *J48*, Redes Neurais Artificiais, *Simple Cart* e *Random Forest* (CHAGAS et al., 2010; GIASSON et al., 2011; TEN CATEN et al., 2011a; VAYSSE; LAGACHERIE, 2017). Estes modelos são os responsáveis por estabelecer as relações entre as variáveis preditoras e as classes de solos, sendo importante o conhecimento da arquitetura e do método de classificação destes algoritmos, pois destes dois fatores dependem a qualidade dos produtos gerados e a compreensão das relações estabelecidas entre variáveis dependentes e independentes.

O uso das regressões logísticas foi bastante explorado nos estudos inicias de predição de ocorrência de solos no Brasil, sendo que até 2012 era a técnica predominante (TEN CATEN et al., 2012). As regressões logísticas realizam a predição estimando a probabilidade de ocorrência das classes através da aplicação de regressões logísticas e tem como principal vantagem à produção de mapas com valores de probabilidade associado a cada classe de solos.

As regressões logísticas têm como desvantagem a sensibilidade à variação da proporção de ocorrência entre as classes de solos presentes nos conjuntos de treinamento e a alta complexidade na interpretação dos parâmetros estatísticos gerados pelos modelos logísticos (GIASSON et al., 2006; TEN CATEN et al., 2011a). Nesse contexto, no Brasil tem ocorrido migração para o uso de técnicas de mineração de dados, que permitem obter melhores resultados e uma maior compreensão das relações entre variáveis dependentes e independentes (PAHLAVAN RAD et al., 2014; DIAS et al., 2016; TENG et al., 2018).

Entre os algoritmos de mineração de dados, o uso de redes neurais artificiais (RNAs) tem apresentado bons resultados. As RNAs são modelos preditores compostos por unidades de processamento simples chamadas neurônios, organizadas em camadas e interligadas por conexões associadas a pesos que possuem a finalidade de ponderar as ligações entre as variáveis de entrada (dependentes e independentes). O resultado desta combinação é usado como argumento de uma função de ativação (*sigmoidal*) que dispara os sinais de ativação dos neurônios, indicando o caminho a ser seguido na classificação (AGATONOVIC-KUSTRIN; BERESFORD, 2000). As RNAs podem ser do tipo

Percetron (única camada) e *Multilayer Percetron* (multicamadas), sendo esta última aplicadas a problemas não lineares, sendo atualmente as mais aplicadas no MDS (ARRUDA et al., 2013; BAGHERI BODAGHABADI et al., 2015; HEUNG; HODÚL; SCHMIDT, 2017). Alguns autores têm argumentado que apesar dos bons resultados alcançados com o uso de RNAs, elas não permitem uma compreensão das relações entre as variáveis preditoras e a ocorrência dos solos, apontando essa característica com principal desvantagem na sua utilização (TEN CATEN et al., 2012; AITKENHEAD; COULL, 2016).

Nesse cenário tem ganhado destaque os minerados do tipo árvores de decisão (AD), que podem extrair padrões de grandes bases de dados e permitem uma maior compreensão das relações entre variáveis preditoras e a ocorrência das classes de solos (GIASSON et al., 2011; TEN CATEN et al., 2013; WOLSKI et al., 2017; MASSAWE et al., 2018). As ADs são métodos de classificação hierárquicos baseados na divisão recursiva binaria (dois nós), aplicada sucessivamente para dividir as classes da variável resposta (classes de solos) até que os subgrupos alcancem um tamanho mínimo ou que não ocorra mais ganho de informação adicional.

Dentre os algoritmos de AD, o *J48* tem apresentado bom desempenho na predição de ocorrência das classes de solos (GIASSON et al., 2011; WOLSKI et al., 2017). Esse algoritmo é uma implementação em *Java* do algoritmo *C4.5* (QUINLAN, 1993), e tem como critérios de classificação a divisão dos dados em cada nó de acordo com o ganho de informação das variáveis preditoras, baseada no conceito de entropia (YILDIRIM, 2015). Dessa forma os dados são divididos em nós e os caminhos por estes nós forma as regras de classificação que são aplicadas para predição. A complexidade dos modelos de AD depende do tamanho e número de regras geradas, sendo que modelos com grande número de regras podem dificultar a compreensão das relações entre as variáveis preditoras e as classes de solos (TAGHIZADEH-MEHRJARDI et al., 2015).

Outra técnica que tem sido aplicada recentemente com bons resultados no MDS é o *Random Forest* (Florestas aleatórias) (PAHLAVAN RAD et al., 2014; DIAS et al., 2016; CHAGAS et al., 2017; DORNIK; DRâGUŢ; URDEA, 2017; TENG et al., 2018). O *Random Forest* (RF) é uma técnica que usa modelos preditores formados pela combinação de conjuntos de árvores de decisão ou de regressão, com base no método de combinação Bagging de Breiman, pelo qual

são gerados e combinados vários modelos preditores de um algoritmo de classificação a partir dos quais são realizadas as predições (BREIMAN, 1996).

O uso da RF tem resultado em desempenho superior aos obtidos com outras técnicas e algoritmos de classificação utilizados no MDS. O uso de RF apresenta como desvantagem a complexidade dos modelos gerados, não permitindo a compreensão das relações entre variáveis independentes e dependentes (SUBBURAYALU; SLATER, 2013; DIAS et al., 2016; MASSAWE et al., 2018). Entretanto, apesar de alguns estudos argumentarem a necessidade da compreensão das regras geradas pelos modelos preditores, as metodologias testadas no MDS tem buscado selecionar as melhores técnicas ou algoritmos preditores, dando pouca atenção a interpretação das regras de classificação e às relações entre as variáveis envolvidas na predição (TEN CATEN et al., 2012; CARVALHO JUNIOR et al., 2014; DIAS et al., 2016; MASSAWE et al., 2018).

2.2.3. Amostragem utilizada no Mapeamento Digital de Solos

O esquema de amostragem é responsável por captar a variabilidade nos dados preditores, pela qual o modelo preditor vai discriminar as classes de solos, sendo que o número e forma de distribuição dos pontos amostrais podem apresentar diferentes eficiências na predição da ocorrência das classes de solos (ADHIKARI et al., 2014; ALVES; DEMATTÊ; BARROS, 2015; BAGATINI; GIASSON; TESKE, 2015; GIASSON et al., 2015; DIAS et al., 2016). Dentre os esquemas amostrais utilizados no MDS podemos destacar os esquemas de amostragem aleatório simples, estratificado e sistemáticos (ADHIKARI et al., 2014; ALVES; DEMATTÊ; BARROS, 2015; BAGATINI; GIASSON; TESKE, 2015; DIAS et al., 2016; TESKE; GIASSON; BAGATINI, 2015b).

O uso dos diferentes esquemas de amostragem está diretamente ligado à fonte de dados que será utilizado como variável resposta (perfis de solos georreferenciados ou mapas legados). A disponibilidade de mapas legados permite uma maior flexibilidade na escolha do esquema de amostragem, permitindo alterar o número e a forma de distribuição dos pontos amostrais, buscando alcançar o melhor desempenho dos algoritmos preditores.

Nesse contexto, o esquema de amostragem aleatória simples tem sido amplamente utilizado, apresentando maior concordância nos mapas preditos (BAGATINI; GIASSON; TESKE, 2015; TESKE; GIASSON; BAGATINI, 2015a).

Entretanto, esses estudos têm demostrado que esse esquema de amostragem favorece a melhor predição das classes de solos de maior proporção nas áreas de estudo, levando até mesma a total omissão na predição das classes de menor extensão.

Frente à dificuldade apresentada pela amostragem aleatória simples em amostrar de forma satisfatória para treinamento dos modelos preditores todas as classes presentes nos mapas de solos legados, estudos tem proposto como alternativa o uso de amostragem estratificada (BRUNGARD et al., 2015; TESKE; GIASSON; BAGATINI, 2015a; CHAGAS et al., 2017; BISWAS; ZHANG, 2018). O uso de esquemas de amostragem estratificado permite alocar pontos amostrais em todas as classes presentes nos mapas de referência, consequentemente, pode melhorar a predição das classes de menor extensão.

No MDS é constatada uma grande variabilidade na densidade de amostras utilizada entre os estudos, com essa densidade variando de 0,2 a 1100 amostras/km^2 (BAGATINI; GIASSON; TESKE, 2015; PELEGRINO et al., 2016). Essa variabilidade na densidade de amostras leva a formação de conjunto de dados insuficientes que não permitem bons ajustes dos modelos preditores ou excessivos os quais resulta em modelos superajustados aos dados de treinamento, contribuindo para os diferentes resultados de concordância observados na literatura e dificulta uma comparação direta da eficiência das metodologias testadas.

Mais recentemente tem ganhado destaque a amostragem pelo uso do hipercubo latino condicionado (cLHS) (MINASNY; MCBRATNEY, 2006). Este método gera uma amostragem aleatória estratificada que contempla a máxima variabilidade dos estratos que compõem as variáveis ambientais utilizadas no modelo preditor (CARVALHO JÚNIOR et al., 2014; BRUNGARD et al., 2015). O uso do cLHS tem sido a técnica de amostragem mais utilizada quando são necessárias coletas de dados a campo (perfis de solos georreferenciados) para treinamentos dos modelos preditores (ADHIKARI et al., 2014; CARVALHO JÚNIOR et al., 2014; PAHLAVAN-RAD et al., 2016; TENG et al., 2018).

Independentemente do esquema de alocação dos pontos amostrais, estudos tem demostrado que uma amostragem aproximada de 45 pontos.km^{-2} pode ser suficiente para obter acurácia geral superior a 70% na reprodutibilidade de mapas legados (TEN CATEN et al., 2013; BAGATINI; GIASSON; TESKE,

2015). Nesse contexto, novos estudos em outros locais com esta mesma densidade de amostras são necessários para reforçar esses resultados, buscando assim uma maior padronização da amostragem e permitindo uma comparação mais direta dos resultados obtidos em função deste componente do MDS.

2.2.4. Variáveis preditoras utilizadas no Mapeamento Digital de Solos

Na predição de ocorrência das classes de solos pode ser empregada uma vasta gama de variáveis ambientais que caracterizam os fatores de formação do solo. No entanto, a falta de dados em escala adequada sobre todos os fatores envolvidos no modelo *SCORPAN* tem limitado o número de fatores pedogenéticos utilizado na predição da ocorrência dos solos (TEN CATEN et al., 2012).

As principais fontes de variáveis preditoras para aplicação no MDS são os dados legados (geológicos, geomorfológicos, hídricos, topográficos e de solos) e, mais recentemente, as bases de dados obtidas por sensoriamento remoto (imagens espectrais de satélites e modelos digitais de elevação). Destas fontes de dados pode ser gerado um grande número de variáveis ambientais que apresentam relação direta ou indireta com os fatores de formação dos solos (MCBRATNEY et al., 2003; TEN CATEN et al., 2012; BEHRENS et al., 2014).

Atualmente podem ser identificadas em diversos estudos, mais de 100 variáveis ambientais em uso como preditoras no MDS. No entanto, com pouco conhecimento da importância da maioria destas variáveis para a predição das classes de solos, fato demostrado pela pequena quantidade de estudos realizados especificamente com esse objetivo (CHAGAS et al., 2010; ARRUDA et al., 2013; ADHIKARI et al., 2014; AFSHAR; AYOUBI; JAFARI, 2018). No Quadro 1 são apresentadas 40 variáveis preditoras que tem apresentado uma maior frequência de uso nos estudos de MDS, no entanto, cada estudo individualmente avaliou no máximo 18 destas variáveis.

Como pode ser constatado, há um grande número de variáveis que podem ser aplicadas no MDS, entretanto, os estudos têm se limitado a avaliar pequenos conjuntos destas variáveis, o que contribui para o pouco conhecimento da relevância da maioria das variáveis preditoras para a predição dos solos. O a avaliação de um pequeno número de variáveis associado a combinações em

subconjuntos, dificulta a identificação de novas variáveis relevantes para uso no MDS (GRINAND et al., 2008; CHAGAS et al., 2010; ABDEL-KADER, 2011; TEN CATEN et al., 2011b; MIRAKZEHI et al., 2018).

Quadro 1 - Variáveis ambientais utilizadas como preditoras em pelo menos três estudos de Mapeamento Digital de Solos.

Ambiente sombreado (C)	Formas do terreno (R)
Área de contribuição (R)	Geologia (P)
Comprimento do fluxo (R)	Geomorfologia (C/R)
Contribuição da declividade (R)	Índice de convergência (R)
Curvatura (R)	Índice de minerais de argila (S)
Curvatura de perfil (R)	Índice de oxido de ferro (S)
Curvatura máxima (R)	Índice de poder de escoamento (R)
Curvatura mínima (R)	Índice de transporte de sedimentos (R)
Curvatura planar (R)	Índice de umidade topográfica (R)
Curvatura total (R)	Índice de vegetação normalizado (0)
Declividade (R)	Índice topográfico composto (R)
Direção do escoamento (R)	Insolação direta (C)
Distância dos rios (R)	Multi-resolution ridge top flatness (MRRTF) (R)
Distância vertical da rede de canais (R)	Multi-resolution valley bottom flatness (MRVBF) (R)
Duração da radiação direta (C)	Orientação da declividade (R)
Elevação (R)	Orientação de vertentes (C)
Elevação normalizada (R)	Posição media da declividade
Elevação padronizada (R)	Profundidade do vale
Escoamento acumulado (R)	Radiação difusa (C)
Fator LS (R)	Radiação solar (C)

R – relevo, P – material de origem, C – Clima, O – organismos (ADHIKARI et al., 2014)

Dentre as variáveis ambientais mais utilizadas, ocorre uma predominância no uso de variáveis derivadas de modelos digitais de elevação (Quadro 1), que caracterizam principalmente o fator relevo (TEN CATEN et al., 2012; MINASNY; MCBRATNEY, 2016). A maior aplicação do fator relevo é justificada por este apresentar diretamente ou indiretamente relação com os outros fatores pedogenéticos do modelo *SCORPAN*, associado à grande disponibilidade de dados na forma de modelos digitais de elevação (MDE) em diferentes resoluções espaciais, que possibilita sua aplicação para caracterizar o ambiente em diversas escalas de detalhe. Outros fatores do modelo *SCORPAN* caracterizados por dados geológicos, geomorfológicos e climáticos, em muitos países, a exemplo do Brasil, estão disponíveis em pequena escala de detalhe, o que limitam sua aplicação em estudos com escalas mais detalhadas.

Além das variáveis presentes no Quadro 1, podemos destacar outras variáveis ambientais como índice de densidade de drenagem, rugosidade do terreno e curvas espectrais, que apresentam potencial para aplicação no MDS,

e que, no entanto, são pouco utilizadas (HORTON, 1955; DOBOS et al., 2000; ABDEL-KADER, 2011; ALVES; DEMATTÊ; BARROS, 2015). Diante do grande número de variáveis ambientais com potencial de uso no MDS, ainda não há consenso sobre quais variáveis são mais relevantes para predição da ocorrência das classes de solos, com os estudos limitando-se a testarem pequenos conjuntos de variáveis, buscando gerar modelos preditores simplificados, isto é, com menor grau de complexidade. Todavia, modelos preditores muito simplificados podem diminuir a qualidade das informações produzidas (KOHAVI; JOHN, 1997; LAL et al., 2006; HALL et al., 2009; BRUNGARD et al., 2015).

Nesse contexto, é importante a avalição de um maior número de variáveis ambientais, buscando identificar as mais relevantes para predição dos solos, assim, como a avaliação de métodos de seleção eficientes para identificação das variáveis relevantes, além de novas abordagens como avaliação em múltiplas escalas para identificação da escala de detalhe mais adequado para predição da ocorrência das classes dos solos (BEHRENS et al., 2010, 2018; MILLER et al., 2015). Paralelamente a isso são necessários estudos objetivando padronizar as densidades de amostragem e o uso de outras fontes de dados de referência, buscando diminuir a dependência na disponibilidade de mapas convencionais e permitindo uma melhor comparação dos resultados obtidos em áreas ou regiões distintas.

3. CAPITULO II – ESTUDO 1: SELEÇÃO DE VARIÁVEIS AMBIENTAIS PARA TREINAMENTO DE ALGORITMOS PREDITORES APLICADOS NO MAPEAMENTO DIGITAL DE SOLOS

3.1. INTRODUÇÃO

As principais variáveis ambientais preditoras utilizadas no Mapeamento Digital de Solos (MDS) são geradas a partir de modelos digitais de elevação (MDE), que podem ser empregados para caracterizar o ambiente nas mais variadas escalas de detalhe (TEN CATEN et al., 2012). A partir dos MDEs podem ser geradas diversas variáveis que caracterizam o relevo e apresentam direta ou indiretamente relação com outros fatores pedogenéticos do modelo *SCORPAN* (MCBRATNEY et al., 2003).

A concordância dos mapas obtidos pelo MDS têm apresentado valores de acurácia geral próximos a 60% e coeficiente Kappa de 0,52, e nem sempre tem sido possível predizer todas as classes de solos contidas nos dados de calibração dos modelos preditores (BEHRENS et al., 2010; COELHO; GIASSON, 2010; TEN CATEN et al., 2012). Esses valores de acurácia podem ser atribuídos à baixa correlação das variáveis preditoras com as classes de solos, sendo que ainda não existe consenso sobre quais variáveis devem ser utilizadas na predição de ocorrência dos solos, como pode ser constatado na variabilidade de combinações utilizadas nos estudos de MDS (HÖFIG; GIASSON; VENDRAME, 2014; TESKE; GIASSON; BAGATINI, 2015a; DIAS et al., 2016).

O desconhecimento da relevância das variáveis ambientais para discriminação das classes de solos tem levado à utilização de conjuntos de preditoras com número de variáveis insuficiente para a predição ou com a presença de variáveis redundantes que aumentam a complexidade e dificultam

a interpretação dos modelos de predição (GUYON; ELISSEEFF, 2003; BEHRENS et al., 2010; TEN CATEN et al., 2012; BRUNGARD et al., 2015).

Para atenuar o problema na complexidade e o baixo desempenho dos modelos preditores, pode-se optar pela aplicação de algoritmos de seleção visando à redução no número de variáveis preditoras (GUYON; ELISSEEFF, 2003; HENGL; GRUBER; SHRESTHA, 2003; COELHO; GIASSON, 2010; PAULA SANTANA et al., 2010; GIASSON et al., 2013). Os principais algoritmos de seleção aplicados na mineração de dados podem ser agrupados em *wrapper* (envelopados), filtros e *embedded* (integrados) (GUYON; ELISSEEFF, 2003; HALL et al., 2009). Os algoritmos do tipo *wrapper* realizam a seleção das variáveis preditoras mediante avaliação da relevância das mesmas através da indução de um modelo de predição. Estes métodos realizam a seleção subtraindo ou adicionando variáveis preditoras ao conjunto e estimando índices de desempenho do respectivo modelo preditor, até conseguir o menor subconjunto de preditoras com desempenho igual ou superior ao conjunto composto por todas as variáveis preditoras em estudo (GUYON; ELISSEEFF, 2003; HALL et al., 2009).

Os algoritmos do tipo filtro são aplicados independentemente do modelo de predição e tem como critério de seleção a avaliação de parâmetros como correlação, distância, ganha de informação e consistência das variáveis (HALL; SMITH, 1999; DASH; LIU; MOTODA, 2000; GUYON; ELISSEEFF, 2003). Este tipo de seleção tem sido utilizado em alguns estudos de MDS e seus resultados são aplicados a qualquer algoritmo de predição (GIASSON et al., 2013; PAES; PLASTINO; FREITAS, 2013; SUBBURAYALU; SLATER, 2013; SUBBURAYALU; JENHANI; SLATER, 2014; TAGHIZADEH-MEHRJARDI et al., 2016; VASU; LEE, 2016).

Os métodos do tipo *embedded* são integrados aos modelos de aprendizagem, sendo específicos para cada algoritmo de predição o que restringe sua aplicação para seleção das variáveis preditores (GUYON; ELISSEEFF, 2003; PAES; PLASTINO; FREITAS, 2013). Assim, como a seleção com algoritimos do tipo *wrapper* a os do tipo *embedded* apresentam seu máximo desempenho no classificador utilizado na seleção das variáveis preditoras. Dessa forma, a seleção do tipo *wrapper* é uma alternativa que poderá resultar na seleção de variáveis mais associadas com a ocorrência dos solos e,

consequentemente, modelos preditores mais acurados (BRUNGARD et al., 2015)

A aplicação de cada método de seleção resulta em diferentes subconjuntos de variáveis e, consequentemente, altera a capacidade preditiva do modelo de predição. Estudos têm demostrado que a seleção do tipo *wrapper* apresenta melhor resultado quando o modelo de predição é do tipo hierárquico (HALL et al., 2009; BRUNGARD et al., 2015). No MDS são utilizados principalmente modelos hierárquicos como método de classificação, no entanto, pode ser constatada na literatura o uso de algoritmos do tipo filtro para seleção de variáveis preditoras, uma vez que os procedimentos são mais rápidos e independem do modelo de predição (GIASSON et al., 2013; SILVA et al., 2013; SUBBURAYALU; SLATER, 2013; SUBBURAYALU; JENHANI; SLATER, 2014; TAGHIZADEH-MEHRJARDI et al., 2016). A aplicação de métodos do tipo filtro pode resultar em conjuntos de variáveis pouco correlacionadas com a ocorrência dos solos, acarretando nos baixos valores de acurácia observados na literatura.

Assim, o presente estudo teve por objetivo comparar três sistemas de seleção de variáveis preditoras, sendo dois algoritmos do tipo filtro e um sistema de seleção do tipo *wrapper*, e avaliar seus impactos no modelo preditor.

3.2. MATERIAL E MÉTODOS

Para a realização deste estudo foi utilizada a área da Bacia Hidrográfica do rio Lajeado Grande (Figura 1). A bacia contempla uma área de aproximadamente 533,8 km^2, inserida na Bacia Hidrográfica U030, na Região Hidrográfica do Alto Uruguai (FREITAS et al., 2012). O clima da região é subtropical úmido, tipo Cfa de Köppen, com precipitação pluvial média anual de 1.778 mm e temperatura média anual de 18.5°C. A geologia corresponde à Província do Paraná, caracterizada principalmente por derrames basálticos da formação Serra Geral (BAGATINI; GIASSON; TESKE, 2015).

A área possui um mapa de solos (Figura 1) na escala de 1:50.000 (KÄMPF; GIASSON; STRECK, 2004a) composto por 14 unidades de mapeamento de solos (UM) simples ou formadas por associações de classes taxonômicas de solos.

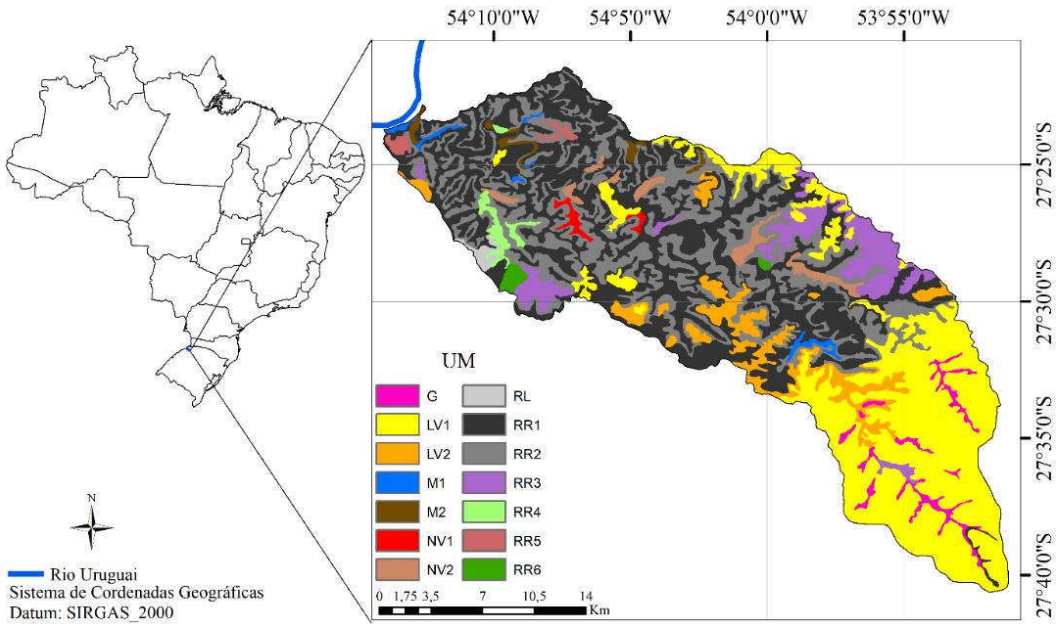


Figura 1 - Localização e mapa de solos da Bacia Hidrográfica do rio Lajeado Grande. UM – unidades de mapeamento de solos.

No Quadro 2 são descritas as composições das 14 unidades de mapeamento de solos que ocorrem na Bacia Hidrográfica do rio Lajeado Grande.

Quadro 2 - Unidades de mapeamento de solos que ocorrem na área da bacia do rio Lajeado Grande.

UM	Composição	Proporção	Inclusões	Área (%)
G	Gleissolos			1,61
LV1	Latossolo Vermelho distroférrico		RR, CX	27,21
LV2	Latossolo Vermelho + Neossolo Regolítico *	60/40	CX	5,84
M1	Chernossolo		RR, RF	0,67
M2	Chernossolo + Neossolo Regolítico*	60/40	CX, RF	0,73
NV1	Nitossolo Vermelho			0,45
NV2	Nitossolo Vermelho + Neossolo Regolítico*	60/40		1,68
RL	Neossolo Litólico + Neossolo Regolítico*		AR	0,41
RR1	Neossolo Regolítico		AR, RL, CX, M, LV	30,35
RR2	Neossolo Regolítico + Neossolo Litólico, relevo forte ondulado*	60/40	AR, CX	23,91
RR3	Neossolo Regolítico + Latossolo Vermelho*	60/40	CX	5,26
RR4	Neossolo Regolítico + Chernossolo*	60/40	CX, RF	0,86
RR5	Neossolo Regolítico + Cambissolo + Nitossolo Vermelho*	50/30/20		0,52
RR6	Neossolo Regolítico + AR *	70/30		0,49
Área total (ha)				53.388

*Associações; AR = afloramento de rocha; CX = Cambissolo; LV = Latossolo Vermelho; M = Chernossolo; RY = Neossolo Flúvico; RL = Neossolo Litólico; RL = Neossolo Litólico; RR = Neossolo Regolítico

Foram geradas 40 variáveis preditoras (Quadro 3) a partir do modelo digital de elevação (MDE) produzido com dados do sensor Aster/GDEM v2 (*Global Digital Elevation Models*), com resolução espacial de 30 metros, datado de 17/10/2011 e obtidos no Serviço Geológico Americano (TACHIKAWA et al., 2011).

Quadro 3 - Variáveis preditoras geradas a partir do modelo digital de elevação.

Abertura Negativa do terreno (ANT)	Ambiente Sombreado (HILLSAHD)
Abertura Positiva do terreno (APT)	Índice de convergência (INDCONVER)
Área de contribuição (ACT)	Índice de densidade de drenagem (IDD)
Área de contribuição/declividade (ACTDECL)	Índice de poder de escoamento (IPE)
Convexidade (CONV)	Índice de posição topográfico (IPT)
Curvatura (CURV)	Índice de umidade topográfico (IUT)
Curvatura de perfil (CURVPERF)	Índice rugosidade terreno (IRT)
Curvatura longitudinal (CURVLONG)	Insolação difusa (INSOLDIF)
Curvatura máxima (CURVMAX)	Insolação direta (INSOLDIR)
Curvatura mínima (CURVMIN)	Insolação total (INSOLTOT)
Declividade (DECL)	MDE suavizado (MDESUAV)
Declividade estandardizada (DECLSTAND)	Índice Multiresolução do Nivelamento do Topo da Crista (MRRTF)
Distância da rede de drenagem (DRD)	Índice Multiresolução de Nivelamento Inferior de Vale (MRVBF)
Distância euclidiana dos rios (DISTRIO)	Nível da base da rede de drenagem (NBRD)
Elevação (MDE)	Orientação de vertentes (ASPECT)
Elevação estandardizada (ALTPAD)	Posição média da declividade (PMDECL)
Elevação normalizada (DEMNORM)	Posição relativa da declividade (RSP)
Escoamento acumulado (ESCACUM)	Profundidade do vale (PROFVALE)
Feições morfométricas (FEIMORF)	Relação da radiação difusa e direta (RRDD)
Formas do terreno (GEOFORMA)	Seção cruzada da curvatura (SCC)

A variável formas do terreno (geoforma), foi derivada com o pacote de ferramentas LandMapR (MACMILLAN, 2003). O índice de densidade de drenagem ($km.km^{-2}$) foi obtido com a extensão *Raster Calculator* do ArcGIS 9.2, seguindo a metodologia proposta por Otto et al. (2017). As demais variáveis foram derivadas com o pacote RSAGA, versão 2.2.2 (BRENNING; BANGS; BECKER, 2018), integrado ao programa computacional R versão 3.3.1 (R CORE TEAM, 2018).

Na sequência, todas as variáveis preditoras foram amostradas juntamente com a variável resposta (mapa convencional de solos), seguindo um esquema de amostragem estratificada, com aproximadamente 30.000 pontos. A estratificação das amostras foi realizada com base no número e tamanho dos

polígonos de cada unidade de mapeamento de solo. Para isso, foram realizadas simulações com diferentes números de pontos em cada UM buscando alcançar um número mínimo de amostras para treinamento de todas as UMs no modelo preditor. Dessa forma foi definido o mínimo de 300 amostras nas UMs de menor extensão, com a distribuição de 3.000 pontos nas nove unidades com área inferior a 1000 hectares (Quadro 2), e os outros 27.000 pontos restantes foram estratificados aleatoriamente em todas as UMs do mapa convencional de solos.

No conjunto completo de dados amostrais (CJ40) foram aplicados três métodos de seleção e separados quatro subconjuntos de dados.

Subconjunto 1 – selecionado pela aplicação do algoritmo *Correlation-based Feature Selection* – CFS (HALL; SMITH, 1999). O algoritmo CFS realiza uma avaliação heurística baseada em correlação, objetivando encontrar subconjuntos que contenham variáveis altamente correlacionadas com a classe e não correlacionadas entre si, variáveis com forte intercorrelação são considerados redundantes e excluídas, esse processo foi realizado no programa computacional Weka 3.8.0 (HALL et al., 2009).

Subconjunto 2 – selecionado pela aplicação do algoritmo *Consistency Subset Eval* – CSE (LIU; SETIONO, 1996). O CSE utiliza a taxa de consistência de classe como medida de avaliação, nesse algoritmo o objetivo é a seleção da variável que dividam o conjunto de dados original em subconjuntos que contenham a maioria das classes, e utiliza o avaliador de consistência proposto por LIU & SETIONO (1996), esse processo também foi realizado no Weka 3.8.0 (HALL et al., 2009). O CFS e CSE tiveram como objetivo comparar a seleção do tipo filtro por correlação e por consistência dos dados, para isso foi mantido o método interno de busca *Best first D1-N5* nos respectivos algoritmos.

Subconjunto 3 – este subconjunto foi selecionado seguindo os fundamentos da seleção tipo *wrapper* (HALL; SMITH, 1999) . Foi montado um *script* em linguagem R, composto por modelos de predição com o algoritmo *J48,* com a seguinte configuração: *J48-C0.25-M2*. Para avaliar o desempenho de cada combinação das variáveis foi realizada a validação cruzada dos modelos preditores com a divisão dos dados em cinco blocos. Para seleção das variáveis foi utilizado o comando recursivo *while*, pelo qual o modelo de predição foi repetido com a eliminação de uma variável até se alcançar um número mínimo

de variáveis no modelo, que apresentavam uma acurácia geral igual ou superior ao conjunto com todas as variáveis (CJ40);

Subconjunto 4 – subconjunto composto pelas variáveis selecionadas simultaneamente nos subconjuntos 1, 2 e 3.

De posse dos conjuntos de variáveis preditoras (subconjuntos 1, 2, 3, 4 e CJ40), procedeu-se a avaliação de desempenho pela aplicação de quatro algoritmos de predição, *J48*, *REPTree* e *BFTree*, e o *Multilayer Perceptron*, já utilizados em outros estudos (COELHO; GIASSON, 2010; GIASSON et al., 2011; ARRUDA et al., 2013; TEN CATEN et al., 2013; CALDERANO FILHO et al., 2014; DIAS et al., 2016).

Estes algoritmos foram selecionados para comparação do desempenho das variáveis em modelos preditores com arquiteturas diferentes, sendo que os três primeiros apresentam arquitetura em árvore de decisão (AD) e o último em redes neurais artificiais (RNA). Para as ADs foi utilizado o mínimo de cinco instâncias por folha final. Todos os algoritmos tiveram validação cruzada por cinco blocos, o que corresponde a 24.000 pontos amostrais (0,45 pontos.ha^{-1}) para treinamento e 6.000 para validação em cada rodada de treinamento do modelo. Estes procedimentos foram realizados com o pacote de ferramentas Weka versão 3.8.0 (HALL et al., 2009) integrado ao programa computacional R.

Os resultados foram avaliados por matriz de erro (CONGALTON, 1991), por meio do coeficiente Kappa (Equação 1), acurácia do mapeador (Equação 2), acurácia geral (Equação 3) e pelos avaliadores erro médio absoluto (Equação 4), raiz quadrada do erro médio (Equação 5), pela área sob a curva precisão-Recall (PRC) obtida a partir do índice Recall (Equação 6) e do índice de Precisão (Equação 7) e coeficiente de correlação de Matthews (Equação 8) (SHI, 2007; Saito e Rehmsmeier, 2015).

Para calcular o coeficiente Kappa foi utilizado a Equação 1 (CONGALTON, 1991).

$$\text{Kappa} = \frac{N\sum_{i=1}^{j} X_{ii} - \sum_{i=1}^{j}(X_{i+} * X_{+i})}{N^2 - \sum_{i=1}^{j}(X_{i+} * X_{+i})} \quad (1)$$

Onde *j* é o número de linhas na matriz, x_{ii} é o número de observações na linha *i* e coluna *i*, x_{i+} e x_{+i} são os totais marginais da linha *i* e coluna *i*, respectivamente, e N é o número total de observações.

Os valores de acurácia do mapeador (AM) e acurácia geral (AG) foram calculados pelas Equações (2 e 3).

$$\mathrm{AM} = \frac{\mathrm{x_{ii}}}{\mathrm{x_{i+}}} \tag{2}$$

Onde x_{ii} são os elementos da diagonal da matriz de erro, e x_{i+} o somatório da linha para dada classe de solo.

$$\mathrm{AG} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{3}$$

Onde TP - são os verdadeiros positivos; TN – verdadeiros negativos; FP - falsos positivos; FN - falsos negativos.

Os valores de erro médio absoluto (MAE) e raiz quadrada do erro médio (RMSE) foram calculados pelas Equações (4 e 5).

$$\mathrm{MAE} = \frac{\sum(\mathrm{pi} - \mathrm{qi})^2}{\mathrm{M}} \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\mathrm{M}} \frac{\sum_{\mathrm{k}}(\mathrm{pi} - \mathrm{qi})^2}{\mathrm{k}}}{\mathrm{M}}} \tag{5}$$

Onde p*i*....p*k* são as probabilidades reais (0 a 1) de uma instância pertencente a uma classe em um problema com k classes, q*i*....q*k* são as probabilidades obtidas com aplicação do modelo preditor para uma instância de uma classe, e M é o número total de instâncias utilizados.

Os valores de Recall (REC), Precisão (PREC) e Coeficiente de correlação de Matthews (MCC) foram obtidos pelas Equações (6, 7 e 8).

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

$$\mathrm{PREC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{7}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) * (\mathrm{TP} + \mathrm{FN}) * (\mathrm{TN} + \mathrm{FP}) * (\mathrm{TN} + \mathrm{FN})}} \tag{8}$$

3.3. RESULTADOS E DISCUSSÃO

Os três métodos de seleção resultaram em diferentes combinações das variáveis preditoras. Isto ocorreu devido aos critérios utilizados por cada método para selecionar ou descartar as variáveis mais importantes para predição da ocorrência dos solos. Das 40 variáveis geradas a partir do MDE, apenas 22 foram relevantes na predição da ocorrência dos solos, tendo sido selecionadas para compor algum dos subconjuntos (Quadro 4) pelos métodos de seleção testados. Este resultado indica que as 18 variáveis não selecionadas apresentam baixa relação com a distribuição espacial dos solos ou são redundantes e, consequentemente, foram descartadas.

Quadro 4 - Variáveis preditoras selecionadas pelos métodos CFS (Subconjunto 1), CSE (Subconjunto 2), *wrapper* (Subconjunto 3) e selecionadas simultaneamente pelos três métodos (Subconjunto 4).

Conjuntos	Variáveis preditoras selecionadas
Subconjunto 1	Índice de densidade de drenagem (IDD); Nível da base da rede de drenagem (NBRD); Orientação de vertentes (ASPECT); Distância da rede de drenagem (DRD); Índice Multiresolução de Nivelamento Inferior de Vale (MRVBF); Área de contribuição/declividade (ACTDECL); Insolação difusa (INSOLDIF); MDE Suavizado (MDESUAV)
Subconjunto 2	Índice de densidade de drenagem (IDD); Nível da base da rede de drenagem (NBRD); Orientação de vertentes (ASPECT); Distância da rede de drenagem (DRD); Índice Multiresolução de Nivelamento Inferior de Vale (MRVBF); Área de contribuição/declividade (ACTDECL); Curvatura (CURV); Curvatura longitudinal (CURVLONG); Elevação (MDE); Índice de umidade topográfico (IUT); Profundidade do vale (PROFVALE); Relação da radiação difusa e direta (RRDD)
Subconjunto 3	Índice de densidade de drenagem (IDD); Nível da base da rede de drenagem (NBRD); Orientação de vertentes (ASPECT); Insolação difusa (INSOLDIF); MDE Suavizado (MDESUAV); Abertura Positiva do terreno (APT); Convexidade (CONV); Declividade (DECL); Distância euclidiana dos rios (DISTRIO); Índice de convergência (INDCONER); Posição média da declividade (PMDECL)
Subconjunto 4	Índice de densidade de drenagem (IDD); Nível da base da rede de drenagem (NBRD); Orientação de vertentes (ASPECT)

O subconjunto 3, selecionado pela aplicação do *wrapper*, resultou no modelo de predição com acurácia geral e coeficiente Kappa superiores aos valores obtidos nos subconjuntos 1 e 2, selecionados pelos filtros CFS e CSE (Figura 2). Esse comportamento foi observado para os quatro algoritmos, indicando que para esses índices não houve interação entre os métodos de seleção e os respectivos algoritmos de predição. O resultado evidencia que mesmo utilizando o *J48* para seleção recursiva das variáveis, o desempenho do

subconjunto de variáveis é mantido nos demais algoritmos de predição, reforçando o bom desempenho do *J48* constatado em outros estudos (COELHO; GIASSON, 2010; GIASSON et al., 2013; DIAS et al., 2016).

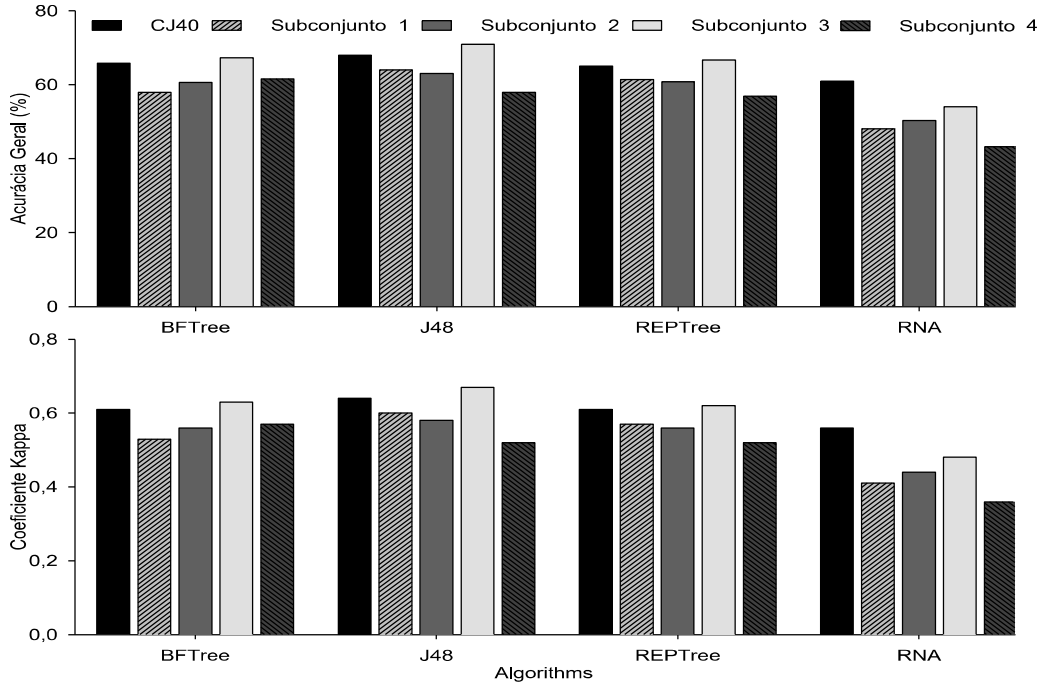


Figura 2 - Desempenho dos conjuntos de variáveis em quatro algoritmos de predição.

O uso das RNAs permitiu gera modelos preditores com número variando de 21 a 40 camadas (Tabela 1). Para as ADs, o algoritmo *BFTree* permitiu gera os modelos preditores com menor número de folhas (1048) e o algoritmo *J48* o os modelos com o maior número de folhas (2142). A diferença entre o tamanho dos modelos gerados pelas RNAs e ADs é explicada principalmente pela arquitetura adotada por estes algoritmos preditores (WITTEN; FRANK; HALL, 2011).

Tabela 1 - Tamanho dos modelos de predição gerados com os cinco conjuntos de variáveis preditoras.

Conjuntos	Nº de variáveis	Algoritmos			
		BFTree(1)	*J48*(1)	*REPTree*(1)	RNA(2)
CJ40	40	1048	1969	1579	40
Subconjunto 1	8	1376	1979	1487	23
Subconjunto 2	12	1243	2142	1647	26
Subconjunto 3	11	1162	1923	1601	25
Subconjunto 4	3	1269	1891	1693	21

(1) número de folhas; (2) número de camadas.

A arquitetura e o tamanho dos modelos preditores estão diretamente ligados à sua complexidade, sendo o ideal modelos com alta capacidade preditiva e uma complexidade que permita a compreensão e interpretação das relações entre variáveis preditoras e ocorrência dos solos (RUIZ; TEN CATEN; DALMOLIN, 2014). Nesse contexto, as RNAs aparentemente menores, apresentam maior complexidade e não permite um completo entendimento da natureza dos dados em análise (TEN CATEN et al., 2012).

Em nenhum dos métodos de seleção avaliados houve redução significativa no tamanho do modelo de predição em relação ao modelo obtido com todas as variáveis. Para o algoritmo *BFTree* e *J48*, os menores modelos preditores foram obtidos a partir do subconjunto 3, selecionado pelo método *wrapper*. Para o *REPTree* o menor modelo foi obtido com o subconjunto 1, selecionado pela aplicação do filtro CFS. Entretanto, vale salientar que apenas no *J48* a redução no tamanho do modelo preditor não resultou em redução nos valores de AG e coeficiente Kappa.

Dentre as variáveis preditoras testadas, apenas as variáveis orientação das vertentes (ASPECT), nível de base da rede de drenagem (NBRD) e índice de densidade de drenagem (IDD) foram selecionadas simultaneamente pelos três algoritmos de seleção para composição do subconjunto 4 (Figura 3).

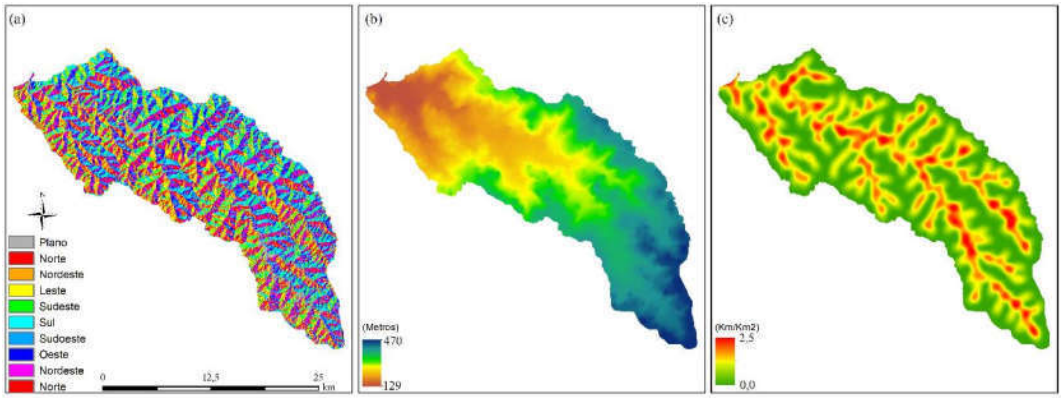


Figura 3 - Variáveis selecionadas respectivamente nos três métodos testados. Orientação das vertentes (a); Nível de base da rede de drenagem (b); Índice de densidade de drenagem (c).

A seleção simultânea destas variáveis é explicada pela sua forte associação com a distribuição espacial dos solos. A orientação das vertentes (Figura 3A) apresenta um efeito direto no microclima, alterando a disponibilidade de água e atividade biológica na pedogênese, consequentemente

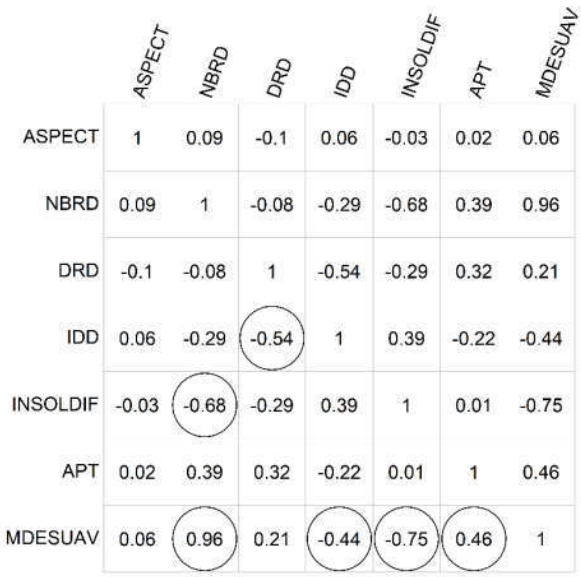
correlacionando-se com a distribuição espacial dos solos na paisagem (SCHAETZL; ANDERSON, 2005).

A variável nível de base da rede de drenagem (Figura 3B) é uma variável intermediária para o cálculo da distância vertical da rede drenagem. Essa variável é obtida relacionando-se a distância vertical até o nível de base da rede de drenagem local e é utilizada como indicativo da profundidade dos solos influenciados pelas águas subterrâneas (BOCK; KÖTHE, 2008).

O NBRD apresenta seus menores valores em ambientes de vales encaixados, indicando que nessas áreas a superfície ocorre mais próximo ao nível de base da rede de drenagem, auxiliando na separação dos solos que ocorrem nessas áreas dos solos que ocorrem nas áreas altimontanas (BOCK; KÖTHE, 2008). Na área de estudo ocorrem UMs formadas por associações com dominância de Latossolos nas áreas altimontanas e UMs com dominância de Neossolos Regolíticos nas encostas dos vales encaixados. Esses dois grupos de solos correspondem à aproximadamente 70% de toda a área de estudo, tornando a variável NBRD de grande importância para diferenciar estes grupos de solos.

O índice de densidade de drenagem (Figura 3C) apresenta relação com propriedades que influenciam a infiltração de água nos solos, como a profundidade, a textura e a drenagem, e é utilizado para diferenciar solos quanto à drenagem. Em áreas com alto IDD, teoricamente há maior limitação de infiltração e, consequentemente, solos menos profundos ou com limitações de percolação no horizonte B (DEMATTÊ; DEMÉTRIO, 1998; DOBOS et al., 2000), dessa forma o IDD contribui para separação de solos poucos profundos ou com limitações de percolação de água.

A aplicação do filtro CFS permitiu a maior eliminação de variáveis. Das 40 variáveis em estudo, apenas sete foram selecionadas para compor o subconjunto 1, dessa forma o CFS promoveu uma redução de 80% no número de variáveis. A matriz de correlação do subconjunto 1 (Figura 4), mostra que menos de 30% das correlações apresentam magnitude superior a 0,4, as quais são classificados como moderadas ou fortes (DANCEY; REIDY, 2006). O filtro CFS utiliza como critério de seleção a baixa correlação entre variáveis, e alta correlação destas com a variável resposta, fato que contribui para seleção de variáveis poucos correlacionadas (HALL et al., 2009).



	ASPECT	NBRD	DRD	IDD	INSOLDIF	APT	MDESUAV
ASPECT	1	0.09	-0.1	0.06	-0.03	0.02	0.06
NBRD	0.09	1	-0.08	-0.29	-0.68	0.39	0.96
DRD	-0.1	-0.08	1	-0.54	-0.29	0.32	0.21
IDD	0.06	-0.29	-0.54	1	0.39	-0.22	-0.44
INSOLDIF	-0.03	-0.68	-0.29	0.39	1	0.01	-0.75
APT	0.02	0.39	0.32	-0.22	0.01	1	0.46
MDESUAV	0.06	0.96	0.21	-0.44	-0.75	0.46	1

Figura 4 - Matriz de correlação das variáveis que compõe o subconjunto 1.

Na figura 4, pode ser observado em destaque, que o maior valor de correlação (0,96), ocorreu entre as variáveis NBRD e o MDESUAV, fato que é explicado pela forma com que o NBDR é calculado, a qual utiliza a distância horizontal da rede de drenagem e a distância vertical (elevação) da base da rede de drenagem, o que acarreta em forte correlação com a variável elevação (MDESUAV). O CFS apresentou alta eficiência na redução do número de preditoras, no entanto, reduziu o desempenho do modelo de predição.

A aplicação do filtro CSE resultou no subconjunto 2, com 12 variáveis selecionadas, o que representou uma redução de 70% no número de variáveis em relação ao conjunto com completo de variáveis (CJ40). Nesse subconjunto, menos de 20% das correlações são superiores a 0,4 (Figura 5) indicando que esse tipo de seleção é eficiente na remoção de variáveis com alta correlação. As variáveis PROFVALE e MDEPAD, DRD e MDEPAD apresentaram valores de correlação 0,63 e -0,71, respectivamente, sendo os valores de correlação mais fortes observados no subconjunto 2. O algoritmo CSE tem como critério de seleção a consistência de subconjuntos em relação a variável resposta, o que permitiria a ocorrência de maiores valores de correlação entre as variáveis selecionadas, no entanto, a aplicação desse filtro permitiu a maior eliminação de variáveis com forte correlação (HALL et al., 2009). Entretanto, assim como no CFS, o CSE também reduziu o desempenho do modelo de predição.

	ACTDECL	ASPECT	NBRD	DRD	CURV	CURVLONG	RRDD	IDD	MRVBF	MDEPAD	IUT	PROFVALE
ACTDECL	1	-0.03	-0.31	0	-0.19	-0.31	-0.23	0.02	-0.45	-0.25	-0.01	0.1
ASPECT	-0.03	1	0.09	-0.1	0	-0.03	0.03	0.06	0.02	0.01	0.01	0.01
NBRD	-0.31	0.09	1	-0.08	-0.01	0.14	0.25	-0.29	0.04	0.56	-0.03	-0.38
DRD	0	-0.1	-0.08	1	0.15	0.62	0.06	-0.54	-0.44	0.63	-0.2	-0.56
CURV	-0.19	0	-0.01	0.15	1	0.19	0.02	0.04	-0.09	0.2	-0.1	-0.11
CURVLONG	-0.31	-0.03	0.14	0.62	0.19	1	0.08	-0.32	-0.25	0.6	-0.17	-0.56
RRDD	-0.23	0.03	0.25	0.06	0.02	0.08	1	-0.14	0.03	0.18	-0.01	-0.16
IDD	0.02	0.06	-0.29	-0.54	0.04	-0.32	-0.14	1	0.35	-0.51	0.08	0.46
MRVBF	-0.45	0.02	0.04	-0.44	-0.09	-0.25	0.03	0.35	1	-0.37	0.16	0.52
MDEPAD	-0.25	0.01	0.56	0.63	0.2	0.6	0.18	-0.51	-0.37	1	-0.2	-0.71
IUT	-0.01	0.01	-0.03	-0.2	-0.1	-0.17	-0.01	0.08	0.16	-0.2	1	0.21
PROFVALE	0.1	0.01	-0.38	-0.56	-0.11	-0.56	-0.16	0.46	0.52	-0.71	0.21	1

Figura 5 - Matriz de correlação para os subconjuntos 2.

O subconjunto 3, selecionado pela aplicação da seleção *wrapper* foi composto por 11 variáveis. Nesse subconjunto 35% das correlações é maior que 0,4 (Figura 6), sendo que esse método foi o menos eficiente na eliminação de variáveis com forte correlação. Os maiores valores de correlação ocorrem entre as variáveis NBRD e MDESUAV (0,95) e PRDECL e INDCONVER (0,85). A seleção *wrapper* foi a única que não reduziu, exceto para a RNA, o desempenho do modelo de predição em relação ao conjunto com todas as 40 variáveis, sendo também observado uma ligeira redução no tamanho do modelo de predição nos algoritmos *BFTree*, *REPTree* e *J48*. Este resultado corrobora com os estudos de Hall e Holmes (2003) e Brungard et al., (2015), que também constataram melhor resultado na predição quando o método *wrapper* foi aplicado para seleção das variáveis preditoras.

A presença de variáveis com correlação classificadas como forte (DANCEY; REIDY, 2006) em todos os subconjuntos indica que as mesmas podem apresentar diferentes graus de importância para as classes de solos, justificando sua presença nos subconjuntos. Esse resultado indica que apenas a análise de correlação entre variáveis não é suficiente para seleção das preditoras mais relevantes para predição de ocorrência dos solos.

	ASPECT	NBRD	CONV	DISTRIO	IDD	INDCONVER	INSOLDIF	APT	PRDECL	DECL	MDESUAV
ASPECT	1	0.09	-0.05	0.01	0.06	-0.05	-0.03	0.02	-0.08	-0.05	0.06
NBRD	0.09	1	0.15	0.49	-0.29	0.12	-0.68	0.39	0.15	-0.27	0.96
CONV	-0.05	0.15	1	0.37	-0.62	0.75	-0.4	0.23	0.75	0.27	0.35
DISTRIO	0.01	0.49	0.37	1	-0.51	0.28	-0.42	0.25	0.36	-0.06	0.54
IDD	0.06	-0.29	-0.62	-0.51	1	-0.56	0.39	-0.22	-0.67	-0.09	-0.44
INDCONVER	-0.05	0.12	0.75	0.28	-0.56	1	-0.35	0.34	0.85	0.14	0.35
INSOLDIF	-0.03	-0.68	-0.4	-0.42	0.39	-0.35	1	0.01	-0.4	-0.32	-0.75
APT	0.02	0.39	0.23	0.25	-0.22	0.34	0.01	1	0.46	-0.56	0.46
PRDECL	-0.08	0.15	0.75	0.36	-0.67	0.85	-0.4	0.46	1	0.13	0.4
DECL	-0.05	-0.27	0.27	-0.06	-0.09	0.14	-0.32	-0.56	0.13	1	-0.19
MDESUAV	0.06	0.96	0.35	0.54	-0.44	0.35	-0.75	0.46	0.4	-0.19	1

Figura 6 - Matriz de correlação para os subconjuntos 3

Em todos os conjuntos de variáveis e algoritmos testados foi possível predizer as 14 unidades de mapeamento. A acurácia do mapeador (AM) seguiu o mesmo comportamento da AG, com os maiores valores observados no subconjunto 3 associado com o algoritmo *J48* (Tabela 2). Os menores valores de AM (<0,3) ocorrem nos algoritmos *BFTree* e RNA associados aos subconjuntos 1 e 2. A maior variação entre o valor mínimo e máximo da AM também ocorre nesses dois subconjuntos. A maior variação entre AM mínima e máxima nos subconjuntos 1 e 2 indica que o erro de predição pode ter se concentrado em poucas unidades de mapeamento.

Comportamento contrário é observado para o algoritmo *J48*, que apresentou no CJ40 e no subconjunto 3 a menor variação entre os valores de AM mínima (0,50) e máxima (0,90), indicando uma maior distribuição do erro entre as UMs, consequentemente, uma melhor predição das 14 unidades de mapeamento de solos. Apenas as unidades de mapeamento LV2, NV2 e RL não obtiveram sua máxima AM no *J48* associado ao subconjunto 3, entretanto, os valores de AM observados para essa combinação de algoritmo de predição e subconjunto foram superiores a 0,67, indicando boa predição também para essas três unidades de mapeamento de solos.

Tabela 2 - Acurácia do mapeador para os modelos de predição ajustados com os cinco conjuntos de variáveis preditoras.

Algoritmos	Unidade de mapeamento de solos													
	G	LV1	LV2	M1	M2	NV1	NV2	RL	RR1	RR2	RR3	RR4	RR5	RR6
CJ40														
BFTree	0,75	0,72	0,74	0,70	0,69	0,73	0,32	0,65	0,45	0,59	0,45	0,65	0,76	0,31
J48	0,79	0,79	0,64	0,79	0,82	0,83	0,65	0,89	0,50	0,55	0,65	0,75	0,85	0,79
REPTree	0,79	0,77	0,61	0,75	0,76	0,84	0,56	0,89	0,42	0,56	0,63	0,73	0,80	0,77
RNA	0,80	0,71	0,51	0,57	0,56	0,69	0,49	0,91	0,45	0,55	0,55	0,71	0,84	0,68
Subconjunto 1														
BFTree	0,69	0,74	0,79	0,73	0,66	0,78	0,43	0,64	0,24	0,56	0,25	0,63	0,67	0,51
J48	0,79	0,76	0,56	0,78	0,79	0,82	0,64	0,86	0,43	0,55	0,57	0,68	0,81	0,77
REPTree	0,80	0,75	0,56	0,68	0,78	0,83	0,59	0,85	0,34	0,55	0,53	0,68	0,79	0,72
RNA	0,63	0,65	0,39	0,21	0,44	0,70	0,20	0,59	0,38	0,56	0,40	0,46	0,44	0,20
Subconjunto 2														
BFTree	0,68	0,76	0,72	0,67	0,63	0,76	0,47	0,65	0,33	0,50	0,18	0,64	0,55	0,51
J48	0,77	0,76	0,54	0,77	0,77	0,80	0,59	0,83	0,43	0,53	0,56	0,69	0,77	0,78
REPTree	0,82	0,75	0,55	0,67	0,79	0,80	0,56	0,85	0,36	0,53	0,51	0,69	0,76	0,72
RNA	0,70	0,66	0,40	0,29	0,46	0,63	0,28	0,64	0,38	0,59	0,43	0,54	0,51	0,22
Subconjunto 3														
BFTree	0,79	0,74	0,73	0,73	0,67	0,72	0,29	0,68	0,46	0,58	0,46	0,67	0,75	0,34
J48	0,83	0,80	0,68	0,84	0,84	0,83	0,69	0,90	0,52	0,59	0,67	0,75	0,88	0,84
REPTree	0,81	0,78	0,64	0,76	0,80	0,84	0,64	0,90	0,43	0,56	0,65	0,74	0,80	0,82
RNA	0,75	0,69	0,48	0,33	0,41	0,62	0,34	0,77	0,40	0,59	0,40	0,57	0,71	0,43
Subconjunto 4														
BFTree	0,71	0,72	0,48	0,76	0,67	0,75	0,52	0,85	0,36	0,43	0,58	0,66	0,72	0,68
J48	0,65	0,73	0,47	0,75	0,75	0,73	0,62	0,82	0,38	0,41	0,51	0,67	0,73	0,76
REPTree	0,71	0,74	0,47	0,77	0,73	0,74	0,51	0,86	0,37	0,35	0,54	0,68	0,70	0,69
RNA	0,46	0,68	0,31	0,22	0,43	0,56	0,26	0,62	0,37	0,25	0,32	0,51	0,63	0,31

As unidades de mapeamento nas quais houve maior dificuldade na predição foram NV2, LV2, RR1, RR2 e RR3, as quais apresentaram os menores valores de AM. Estas unidades de mapeamento são compostas por Neossolos Regolíticos (RR1) ou destes em associação com outras classes, como Neossolos Litólicos e Latossolos Vermelhos. Dessa forma, essas UMs podem ocorrem de forma adjacentes ou em ambientes semelhantes, dificultando sua predição pelos algoritmos preditores. Höfig et al. (2014) relataram que unidades de mapeamento que ocupam posições muito semelhantes na paisagem podem impor maior dificuldades na discriminação pelos modelos preditores.

O desempenho dos subconjuntos e algoritmos também é respaldado pelos valores obtidos nos outros índices avaliados (Tabela 3). Com exceção do algoritmo *RNA*, podem ser observados que os valores de PRC, coeficiente

Kappa e AG apresentam uma variação inferior a 5% nos respectivos modelos preditores.

Tabela 3 - Desempenho dos conjuntos de variáveis preditoras em cada algoritmo de predição.

Conjuntos	Algoritmos	MAE	RMSE	Área PRC	CCM	Kappa	AG (%)
CJ40	*BFTree*	0,06	0,19	0,63	0,61	0,61	65,76
	J48	0,05	0,19	0,64	0,63	0,64	68,00
	REPTree	0,06	0,19	0,66	0,60	0,61	65,00
	RNA	0,07	0,20	0,61	0,56	0,56	60,97
Subconjunto 1	*BFTree*	0,07	0,20	0,58	0,52	0,53	57,95
	J48	0,07	0,21	0,56	0,52	0,60	64,00
	REPTree	0,07	0,20	0,62	0,56	0,57	61,41
	RNA	0,09	0,22	0,49	0,41	0,41	48,06
Subconjunto 2	*BFTree*	0,07	0,20	0,59	0,55	0,56	60,63
	J48	0,06	0,20	0,60	0,57	0,58	63,00
	REPTree	0,07	0,20	0,62	0,55	0,56	60,77
	RNA	0,09	0,21	0,51	0,44	0,44	50,33
Subconjunto 3	*BFTree*	0,06	0,19	0,65	0,63	0,63	67,32
	J48	0,05	0,18	0,68	0,67	0,67	71,00
	REPTree	0,06	0,19	0,68	0,62	0,62	66,66
	RNA	0,08	0,21	0,54	0,48	0,48	54,05
Subconjunto 4	*BFTree*	0,07	0,20	0,60	0,56	0,57	61,61
	J48	0,06	0,20	0,62	0,60	0,52	58,00
	REPTree	0,08	0,20	0,58	0,51	0,52	56,90
	RNA	0,10	0,22	0,42	0,36	0,36	43,25

AG – acurácia geral, RMSE - raiz quadrada do erro médio, MAE - erro médio absoluto, Área PRC - área sob a curva precisão-recall, CCM - coeficiente de correlação de Matthews

A pequena variabilidade nestes índices indica baixa aleatoriedade nas classificações realizadas por estes modelos, o que evidencia boa eficiência das variáveis utilizadas na predição da ocorrência das UMs. Como pode ser observado nos resultados para o subconjunto 4, apenas as três variáveis selecionadas simultaneamente nos três métodos, respondem por mais de 56% da concordância obtidos nos modelos de árvore de decisão e 43% no modelo com *RNA*, reforçando a relevância destas variáveis na predição da ocorrência dos solos.

O valor do erro médio absoluto (MAE) e da raiz quadrada do erro médio (RMSE) apresentam valores relativamente baixos, com os maiores valores iguais a 0,10 e 0,22, respectivamente, ocorrendo nos modelos preditores gerados pelas RNAs. Os valores observados para o coeficiente de Matthews (CCM) indicam boa correlação dos dados preditos com o mapa de referência,

com o maior valor (0,67) no algoritmo *J48* associado ao subconjunto 3. Para o CCM, apenas o algoritmo *RNA* apresentou valores abaixo de 0,5, no entanto, vale ressaltar que o CCM é melhor utilizado quando se trabalha com número de amostras da variável resposta balanceadas, sendo que no presente estudo o número de amostras por classe apresenta variação de 300 a 6000 amostras entre as classes de menor e maior extensão (RL e RR1), assim os valores observados podem ter sido subestimados (SAITO; REHMSMEIER, 2015).

Como visto os três métodos de seleção resultaram em diferentes combinações das variáveis preditoras e, consequentemente, diferentes desempenhos na predição da ocorrência das UMs, sendo que a seleção do tipo *wrapper* apresentou desempenho ligeiramente superior aos demais métodos em todos os algoritmos testados, concordando com outros estudos (HALL; SMITH, 1999; BRUNGARD et al., 2015).

Todos os métodos de seleção resultaram em redução no número de variáveis, reforçando a necessidade do pré-processamento dos dados para maximizar o desempenho dos algoritmos preditores utilizados no Mapeamento Digital de Solos, concordando com outros estudos (GIASSON et al., 2013; PAES; PLASTINO; FREITAS, 2013; SUBBURAYALU; SLATER, 2013; SUBBURAYALU; JENHANI; SLATER, 2014; TAGHIZADEH-MEHRJARDI et al., 2016; VASU; LEE, 2016). A seleção *wrapper* por realizar a avaliação direcionada a um modelo de predição da relevância de cada variável de forma independente permitiu a seleção de variáveis altamente correlacionadas com a variável resposta e o máximo desempenho dos algoritmos utilizados para classificação.

3.4. CONCLUSÕES

- O uso da seleção *wrapper* permitiu obter os melhores valores de desempenho para o modelo preditor nos algoritmos *BFTree*, *REPTree*, *J48* e na RNA *Multilayer Perceptron*.
- A aplicação de cada método de seleção resultou em subconjuntos com 30% das 40 variáveis preditores testadas, e permitiu a predição das 14 unidades de mapeamento.

- A aplicação dos algoritmos de seleção do tipo filtro *Correlation-based Feature Selection (CFS)* e *Consistency Subset Eval* (CSE) reduziram os valores de acurácia e coeficiente Kappa em relação ao conjunto com todas as variáveis.
- Apenas as variáveis: orientação das vertentes, nível de base da rede de drenagem e índice de densidade de drenagem foram às únicas selecionadas simultaneamente pelos três métodos testados.
- O modelo preditor gerado com o algoritmo *J48* apresentou maior concordância na validação, confirmando os resultados obtidos por outros estudos que indicam esse algoritmo como o mais eficiente na predição de ocorrência das classes de solos.

4. CAPITULO III – ESTUDO 2: AVALIAÇÃO DE VARIÁVEIS PREDITORAS GERADAS DE MODELOS DIGITAIS DE ELEVAÇÃO SUAVIZADOS NA PREDIÇÃO DE OCORRÊNCIA DE SOLOS

4.1. INTRODUÇÃO

A qualidade dos mapas de classes de solos gerados pelo MDS tem apresentado grande variação, com valores de acurácia geral variando de 30% a 95% e média inferior a 60%. O mesmo comportamento é observado para o coeficiente Kappa, que apresenta valor médio inferior a 0,52 (TEN CATEN et al., 2012; CALDERANO FILHO et al., 2014; PELEGRINO et al., 2016).

Estes valores refletem a baixa eficiência das metodologias testadas e contribui para o não uso do MDS como método de levantamentos de solos. Dentre os fatores que contribuem para a baixa eficiência das metodologias testadas está o desempenho dos conjuntos de variáveis preditoras utilizadas (MCBRATNEY et al., 2003; BEHRENS et al., 2010; GIASSON et al., 2011; TEN CATEN et al., 2011a; BRUNGARD et al., 2015; TESKE; GIASSON; BAGATINI, 2015a).

As variáveis ambientais preditoras utilizadas no MDS são responsáveis por diferenciar os padrões de ocorrência das classes de solos na paisagem, no entanto, sua relevância para predição da ocorrência dos solos tem sido pouco estudada. Atualmente, são utilizadas mais de 100 variáveis ambientais como preditoras, com os estudos individualmente se limitando a avaliar conjuntos que variam de 3 a 18 destas variáveis (ARRUDA et al., 2013; AFSHAR; AYOUBI; JAFARI, 2018), sendo predominante a utilização de variáveis geradas de modelos digitais de elevação (MDE). Geralmente são testados pequenos subconjuntos, objetivando reduzir o número de variáveis nos modelos preditores, permitindo assim uma maior compreensão das relações entre a variável resposta e as preditoras.

Além do baixo número de variáveis testadas por estudo, estas podem não apresentar escala adequada para representar a ocorrência dos solos, contribuindo para os valores de concordância observados na literatura (BEHRENS et al., 2005; GRINAND et al., 2008). Nesse contexto, a aplicação de técnicas para suavizar as variáveis preditoras poderá aumentar a correlação com a distribuição dos solos, permitindo selecionar variáveis no nível de detalhe mais adequado para predição da ocorrência dos solos (BEHRENS et al., 2005; GRINAND et al., 2008).

Dentre as técnicas que podem ser aplicadas nas variáveis, podemos destacar o filtro de média (tipo de filtro passa baixa), utilizados geralmente como ferramenta de pré-processamento de MDEs para remoção de ruídos e suavização da elevação. O filtro de média é uma técnica baseada na convolução de janelas moveis (máscaras), que através de um conjunto de pixels calcula novos valores para os dados de uma matriz. A suavização promovida pela aplicação do filtro de média permite alterar o nível de detalhe das informações contidas no MDE, modificando as formas de representação do terreno e facilita na identificação de feições do terreno (BATES; METCALFE, 2006).

Behrens et al.(2005, 2010) e Grinand et al.(2008) constataram que as variáveis do terreno em diferentes tamanhos de filtro apresentam variação no desempenho na predição da ocorrência dos solos, sendo que as classes de solos podem necessitar da mesma variável preditora em diferentes escalas de detalhe. Esses estudos indicaram ganho de 10% nos valores de concordância em mapas preditos com variáveis preditoras que receberam aplicação do filtro de média.

Diante do exposto, é possível pressupor que a aplicação do filtro de média no MDE altera sua escala, assim como das demais variáveis obtidas a partir da elevação e aumenta a correlação com a distribuição espacial dos solos e o desempenho dos modelos preditores utilizados no MDS. Nesse contexto, este estudo foi realizado para avaliar a aplicação do filtro de média no MDE e o desempenho de variáveis preditoras geradas dos MDEs filtrados na predição de ocorrência dos solos utilizando dois esquemas de alocação de pontos amostrais.

4.2. MATERIAL E MÉTODOS

Foram utilizados dados de solos e de elevação das Bacias dos rios Santo Cristo (SC) e Lajeado Grande (LG), ambas localizadas na unidade hidrográfica U030, região noroeste do Rio Grande do Sul. Às áreas possuem mapas de solos na escala 1:50.000 (KÄMPF; GIASSON; STRECK, 2004b, 2004a), compostos por unidades de mapeamento de solos (Quadro 5). A geologia das duas áreas corresponde à Província do Paraná, caracterizada principalmente por derrames basálticos da formação Serra Geral. O clima da região é o subtropical úmido, tipo Cfa de Köppen, com precipitação pluvial média anual de 1.778 mm e temperatura média anual de 18.5°C (FREITAS et al., 2012).

Quadro 5 - Unidades de mapeamento de solos que ocorrem nas áreas das bacias dos rios Lajeado Grande e Santo Cristo.

UM	Composição	Proporção	Inclusões	Área (%)
	Bacia hidrográfica do rio Lajeado Grande (LG)			
G	Gleissolos			1,61
LV1	Latossolo Vermelho distroférrico		RR, CX	27,21
LV2	Latossolo Vermelho + Neossolo Regolítico *	60/40	CX	5,84
M1	Chernossolo		RR, RF	0,67
M2	Chernossolo + Neossolo Regolítico*	60/40	CX, RF	0,73
NV1	Nitossolo Vermelho			0,45
NV2	Nitossolo Vermelho + Neossolo Regolítico*	60/40		1,68
RL	Neossolo Litólico + Neossolo Regolítico*		AR	0,41
RR1	Neossolo Regolítico		AR, RL, CX, M, LV	30,35
RR2	Neossolo Regolítico + Neossolo Litólico, relevo forte ondulado*	60/40	AR, CX	23,91
RR3	Neossolo Regolítico + Latossolo Vermelho*	60/40	CX	5,26
RR4	Neossolo Regolítico + Chernossolo*	60/40	CX, RF	0,86
RR5	Neossolo Regolítico + Cambissolo + Nitossolo Vermelho*	50/30/20		0,52
RR6	Neossolo Regolítico + AR *	70/30		0,49
Área total (ha)				53.388
	Bacia hidrográfica do rio Santo Cristo (SC)			
G1	Gleissolos			2,46
LV1	Latossolo Vermelho distroférrico		RR, CX	38,13
LV2	Latossolo Vermelho + Neossolo Regolítico*	60/40	CX	7,92
M1	Chernossolo Háplico		CX	0,20
RL	Neossolo Litólico			0,01
RR1	Neossolo Regolítico + Cambissolo Háplico*	60/40	AR, RL, CX	34,79
RR2	Neossolo Regolítico + Neossolo Litólico, relevo forte ondulado**	50/50	AR, CX	8,24
RR3	Neossolo Regolítico + Latossolo Vermelho*	60/40	CX	7,84
RR4	Neossolo Regolítico + Neossolo Litólico*	70/30		0,03
RR5	Neossolos Regolíticos + Cambissolo Háplico + Latossolo Vermelho*	50/30/20		0,38
Área total (ha)				90.073

*Associações; **Complexos; AR = afloramento de rocha; CX = Cambissolo; LV = Latossolo Vermelho; M = Chernossolo; RY = Neossolo Flúvico; RL = Neossolo Litólico; RL = Neossolo Litólico; RR = Neossolo Regolítico

Para avaliação do modelo digital de elevação (MDE) em múltiplos níveis de detalhe optou-se pelo uso do MDE com suavização pela aplicação do filtro de média (GRINAND et al., 2008; BEHRENS et al., 2010). Foram aplicados seis tamanhos (máscaras) de filtro (5x5, 10x10, 15x15, 20x20, 25x25 e 30x30 pixels) em duas formas (sequencial e não sequencial) como pode ser observado na Figura 6.

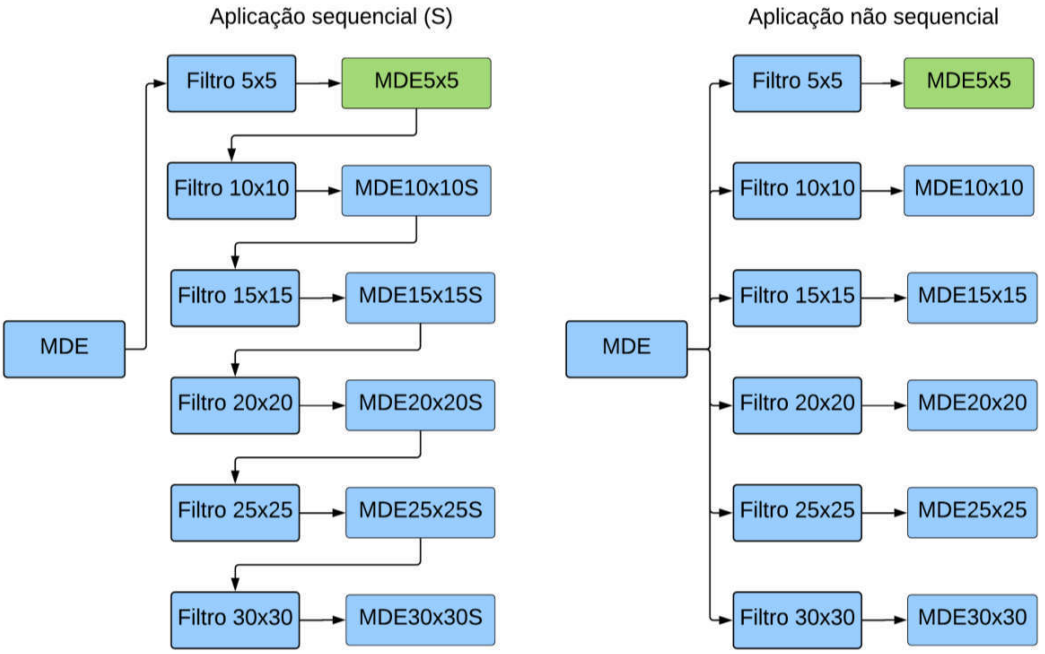


Figura 7 - Tamanho e forma de aplicação do filtro de média no modelo digital de elevação e os respectivos modelos digitais de elevação (MDE) obtidos.

O filtro de média utilizado foi o *Simple Filter/Smooth/Circle* (CONRAD et al., 2015). Para esse procedimento foi utilizado o programa SAGA GIS integrado ao programa R (R CORE TEAM, 2018). O MDE utilizado foi produzido a partir de dados do sensor Aster/GDEM v2 (*Global Digital Elevation Models*), com resolução espacial de 30 m, datado de 17/10/2011 e obtido no Serviço Geológico Americano (TACHIKAWA et al., 2011).

A aplicação do filtro de média no MDE deu origem a 11 modelos digitais de elevação suavizados (Figura 6), que juntamente com o MDE sem filtro totalizou 12 MDEs, a partir dos quais foram geradas oito variáveis preditoras (nível da base da rede de drenagem, orientação de vertentes, insolação difusa, abertura positiva do terreno, convexidade, declividade, índice de convergência e posição média da declividade). Além das oito variáveis listadas acima, foram geradas as variáveis índice de densidade de drenagem (OTTO et al., 2017) e a distância

euclidiana dos rios a partir da rede hidrográfica na escala 1:50.000 (HASENACK; WEBER, 2010) que tambem foram filtradas com os respectivos tamanhos e formas aplicadas aos MDEs. As variáveis utilizadas tiveram como critério os resultados obtidos no Estudo 1 (Capitulo II), no qual as 11 variáveis listadas foram selecionadas com mais eficientes na predição das classes de solos.

No total foram gerados 12 conjuntos de variáveis preditoras, as quais foram amostradas utilizando dois esquemas de amostragem:

Amostragem aleatória simples (AA), no qual utilizando a ferramenta *Create Random Points* (ArcGIS), foram criados *shapefiles* com pontos alocados completamente aleatório em cada uma das bacias em estudo;

Amostragem aleatória estratificada (AE), no qual a seleção dos pontos amostrais foi realizada nos pixels dos respectivos mapas de solos pelo comando *sample.split* no programa computacional R (R CORE TEAM, 2018), com as amostras distribuídas de forma aleatória em todos os polígonos dos respectivos mapas de solos.

Para os dois esquemas de amostragem foi adotado a densidade de 40 pontos.km^{-2} (TEN CATEN et al., 2013; BAGATINI; GIASSON; TESKE, 2015). A combinação dos três fatores em estudo (tamanho do filtro, forma de aplicação do filtro e esquema de amostragem) resultou em 24 conjuntos de dados preditores para cada bacia (Quadro 6).

Quadro 6 - Conjuntos de variáveis preditoras utilizadas na predição da ocorrência dos solos nas bacias Lajeado Grande e Santo Cristo.

MDE de origem das variáveis	Conjuntos de variáveis preditoras	
	Amostragem Aleatória (AA)	Amostragem Estratificada (AE)
MDE	AAF0	AEF0
MDE5x5	AAF5	AEF5
MDE10x10	AAF10	AEF10
MDE15x15	AAF15	AEF15
MDE20x20	AAF20	AEF20
MDE25x25	AAF25	AEF25
MDE30x30	AAF30	AEF30
MDE10x10S	AAF10S	AEF10S
MDE15x15S	AAF15S	AEF15S
MDE20x20S	AAF20S	AEF20S
MDE25x25S	AAF25S	AEF25S
MDE30x30S	AAF30S	AEF30S

O efeito dos tamanhos e formas de aplicação do filtro de média nos MDEs foram avaliados pelas alterações nos valores mínimos, máximos e médios dos

respectivos MDEs. As alterações no MDE foram analisadas na área total das bacias e dentro das unidades de mapeamento de solos das respectivas áreas de estudo.

Para avaliação dos conjuntos de variáveis preditoras foram preditos mapas de solos com a aplicação do classificador hierárquico *J48* (J48 -C 0.25 -M 10) (QUINLAN, 1993), utilizando o pacote *RWeka* (HORNIK et al., 2016). O uso do *J48* teve como base os resultados obtidos no Estudo 1 e outros trabalhos publicados (COELHO; GIASSON, 2010; GIASSON et al., 2011, 2013b). As concordâncias entre os mapas preditos e os mapas de referência foram avaliadas utilizando a matriz de erro, por meio do coeficiente Kappa, Acurácia Geral (AG) e Acurácia do Mapeador (AM) (CONGALTON, 1991).

4.3. RESULTADOS E DISCUSSÃO

4.3.1. Análise descritiva do modelo digital de elevação antes e após aplicação do filtro de média

A aplicação do filtro de média alterou os valores da elevação, reduzindo principalmente a amplitude dos dados. Os valores mínimos da elevação nas bacias Lajeado Grande e Santo Cristo aumentaram, respectivamente, em 39,8 e 46,2 metros entre o MDE sem filtro e o MDE30x30S (Tabela 4). Comportamento contrário ocorreu para os valores máximos da elevação, onde houve uma redução de 48,9 m na bacia Lajeado Grande, e 41,5 m na bacia Santo Cristo entre o MDE sem filtro e o MDE30x30S. Para os valores da média não foram observadas alterações significativas.

Esses valores de redução estão de acordo com o objetivo da aplicação do filtro, que foi promover uma suavização das informações contidas nos MDEs, permitindo alterar o nível de representação dos atributos do terreno gerados a partir destes MDEs. Essas alterações permitem identificar o nível de detalhe das variáveis preditoras mais adequado para representar a distribuição espacial das classes de solos na paisagem, resultando em maior desempenho dos modelos preditores (BEHRENS et al., 2010, 2018).

Tabela 4 – Análise descritiva dos modelos digitais de elevação antes e após a aplicação dos filtros de média para suavização.

Modelo digital de elevação	Bacia hidrográfica do rio Lajeado Grande (LG)			Bacia hidrográfica do rio Santo Cristo (SC)		
	Mínimo (m)	Máximo (m)	Média (m)	Mínimo (m)	Máximo (m)	Média (m)
MDE	132	520,4	336,8	100,1	430,8	257,2
MDE5x5	132,7	508,3	336,7	103,7	419,8	257,2
MDE10x10	137,5	499	336,7	107,6	410,7	257,2
MDE15x15	140	491,9	336,6	114,3	404,8	257,2
MDE20x20	148,9	487,6	336,5	120,9	399,8	257,2
MDE25x25	153,9	483,8	336,4	124,2	395,8	257,2
MDE30x30	159,1	479,8	336,3	126,9	392,7	257,3
MDE10x10S	138,8	497,1	336,6	108,9	409,1	257,2
MDE15x15S	146,4	489,4	336,5	118,5	401,5	257,2
MDE20x20S	156	482,8	336,4	127,5	396	257,2
MDE25x25S	164,5	476,6	336,3	136,4	392,1	257,3
MDE30x30S	171,8	471,5	336,1	146,3	389,3	257,5

Como pode ser observado nos perfis topográficos das bacias Lajeado Grande (Figura 8) e do Santo Cristo (Figura 9) as maiores modificações ocorrem a curtas distâncias, sendo que quanto maior o tamanho do filtro aplicado maior foi à suavização dos dados, como pode ser visto no perfil topográfico onde é visualizada a remoção de alguns picos de elevação.

Figura 8 - Perfil topográfico dos modelos digitais de elevação sem e com aplicação do filtro de média sequencial (S) para a bacia do rio Santo Cristo

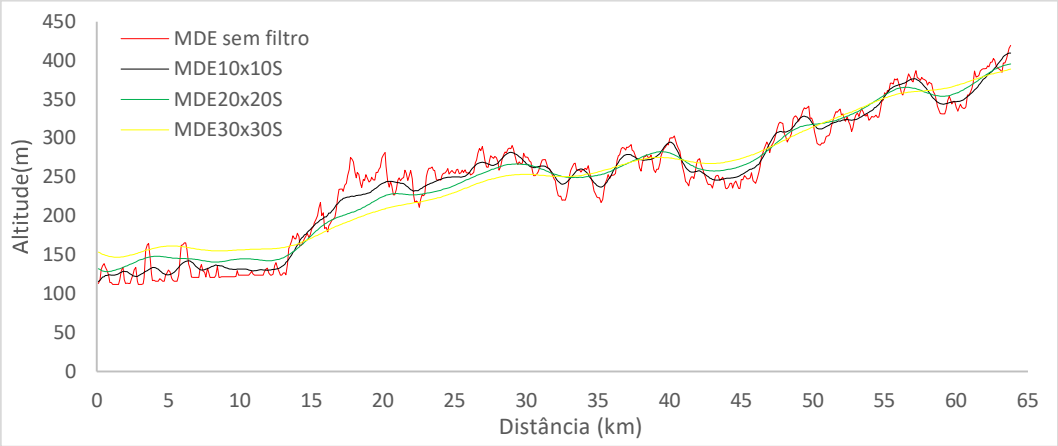


Figura 9 - Perfil topográfico dos modelos digitais de elevação sem e com aplicação do filtro de média sequencial (S) para a bacia do rio Lajeado Grande.

Assim como para os parâmetros dos MDEs delimitados pelas respectivas bacias, também foram constatadas alterações nos dados quando delimitados pelas unidades de mapeamento de solos (UMs). A aplicação do filtro de forma sequencial promoveu maior alteração na elevação dentro das UMs, no entanto, essas modificações são pouco visíveis quando analisadas apenas no gráfico (Figura 10), motivo pelo qual apenas os dados para o filtro sequencial aplicado ao MDE sem filtro, MDE15x15S, MDE30x30S são apresentados na Figura 10.

A aplicação do filtro de média promoveu redução na amplitude da elevação na maioria das UMs, sendo que para algumas ocorreu redução ou completa eliminação de pontos com valores atípicos (*outliers)* para as UMs (Figura 10). Esse mesmo comportamento também foi observado em menor escala para os dados que receberam filtragem não sequencial. A diferença que ocorre entre as formas de aplicação do filtro de média se deve a maior generalização promovida pelo filtro aplicado de forma sequencial, uma vez que os MDEs já foram alterados pelo filtro aplicado com menor tamanho.

Na bacia Lajeado Grande houve redução significativa de *outliers* nas unidades G, LV1, LV2, NV1 e RR3 e uma redução mais acentuada na amplitude dos dados nas unidades RL, RR4 e RR6 (Figura 10) sendo que essas alterações foram mais significativas a partir do MDE15x15S. Para a bacia Santo Cristo também foi observado comportamento semelhante ao da LG, com redução na amplitude da elevação, principalmente nas unidades LV1, RR1, RR3, RR4 e RR5, sendo que a partir do MDE20x20S apenas a unidade LV1 permaneceu com *outliers*.

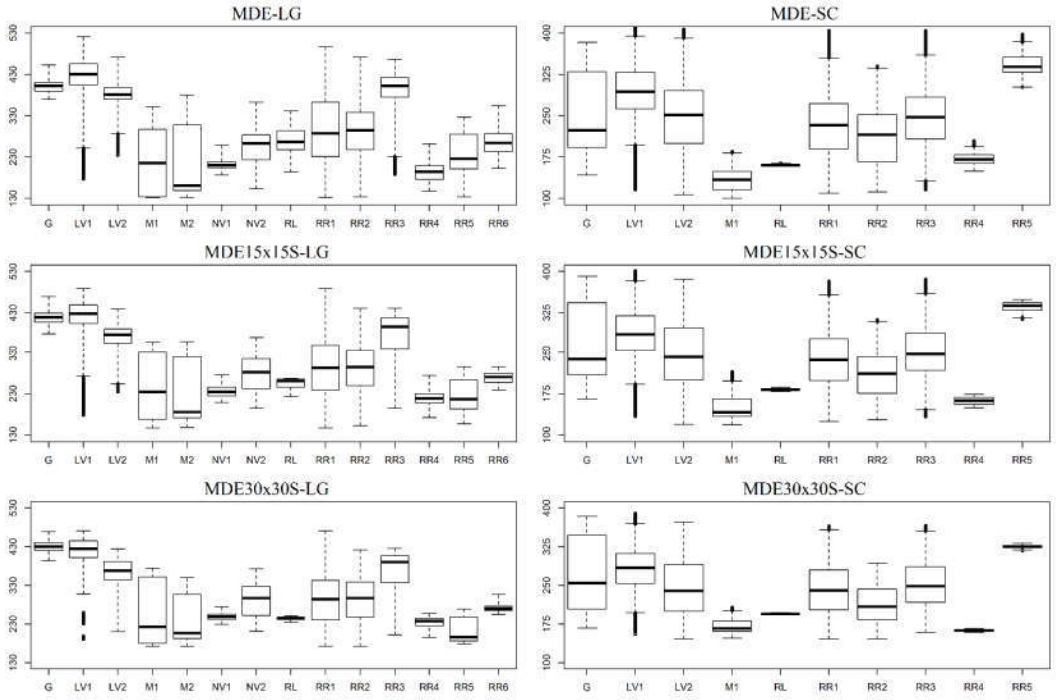


Figura 10 - Gráficos para os MDE suavizados com filtro sequencial para cada unidade de mapeamento constante no mapa convencional de solos das bacias Lajeado Grande (LG) e Santo Cristo (SC).

A redução na amplitude dos dados e a remoção dos *outliers* aumentou a homogeneidade das variáveis preditoras dentro das unidades de mapeamento de solos, criando uma identidade espacial mais consistente das variáveis preditoras com a distribuição dos solos na paisagem e facilitando sua predição (KERRY; OLIVER, 2011; BEHRENS et al., 2014). Para os valores médios da elevação delimitados pelas respectivas UMs não houve grandes alterações, indicando que os valores médios de elevação associado a cada UM não foram alterados pela aplicação do filtro de média.

4.3.2. Desempenho na predição da ocorrência dos solos dos conjuntos de variáveis preditoras geradas a partir dos modelos digitais de elevação com e sem suavização pela aplicação do filtro de média.

Os conjuntos de variáveis preditoras obtidas a partir dos MDEs resultaram em diferentes desempenhos na predição da ocorrência dos solos. De modo geral, a aplicação do filtro no MDE resultou em ganho na concordância dos mapas preditos (Tabela 5). Na bacia LG, a acurácia geral alcançou 77% nos

conjuntos AEF20S, AEF25S, AEF30S e AAF20S, representando um ganho de 14,9% em relação aos mapas preditos com os conjuntos AAF0 e AEF0, derivados do MDE sem filtro.

Tabela 5 – Acurácia geral (reprodutibilidade) dos mapas obtidos a partir dos respectivos conjuntos de variáveis preditoras.

Esquema de amostragem	Conjuntos de dados	Bacia do rio Lajeado Grande (LG)		Bacia do rio Santo Cristo (SC)	
		Acurácia Geral	Coeficiente Kappa	Acurácia Geral	Coeficiente Kappa
Estratificada	AEF0	67	0,57	63	0,47
	AEF5	72	0,63	70	0,57
	AEF10	73	0,65	72	0,6
	AEF15	73	0,65	73	0,62
	AEF20	74	0,67	75	0,65
	AEF25	76	0,69	76	0,66
	AEF30	76	0,68	76	0,67
	AEF10S	75	0,67	75	0,65
	AEF15S	76	0,69	76	0,66
	AEF20S	77	0,70	77	0,68
	AEF25S	77	0,70	78	0,68
	AEF30S	77	0,70	78	0,69
Aleatória	AAF0	67	0,56	63	0,47
	AAF5	71	0,63	70	0,57
	AAF10	73	0,65	71	0,59
	AAF15	74	0,66	73	0,62
	AAF20	75	0,67	74	0,64
	AAF25	75	0,67	75	0,66
	AAF30	75	0,68	76	0,66
	AAF10S	74	0,66	75	0,65
	AAF15S	75	0,67	77	0,67
	AAF20S	77	0,69	77	0,68
	AAF25S	76	0,69	77	0,68
	AAF30S	77	0,70	78	0,69

Nas Figuras 11 e 12 são aparesentados os mapas obtidos com o conjunto de dados AEF20S, os quais atigiram os maximos valores de concordância (acurácia geral e acurácia do mapeador) com os mapas convencionais das bacias Lajeado Grande (Figura 11) e Santo Cristo (Figura 12).

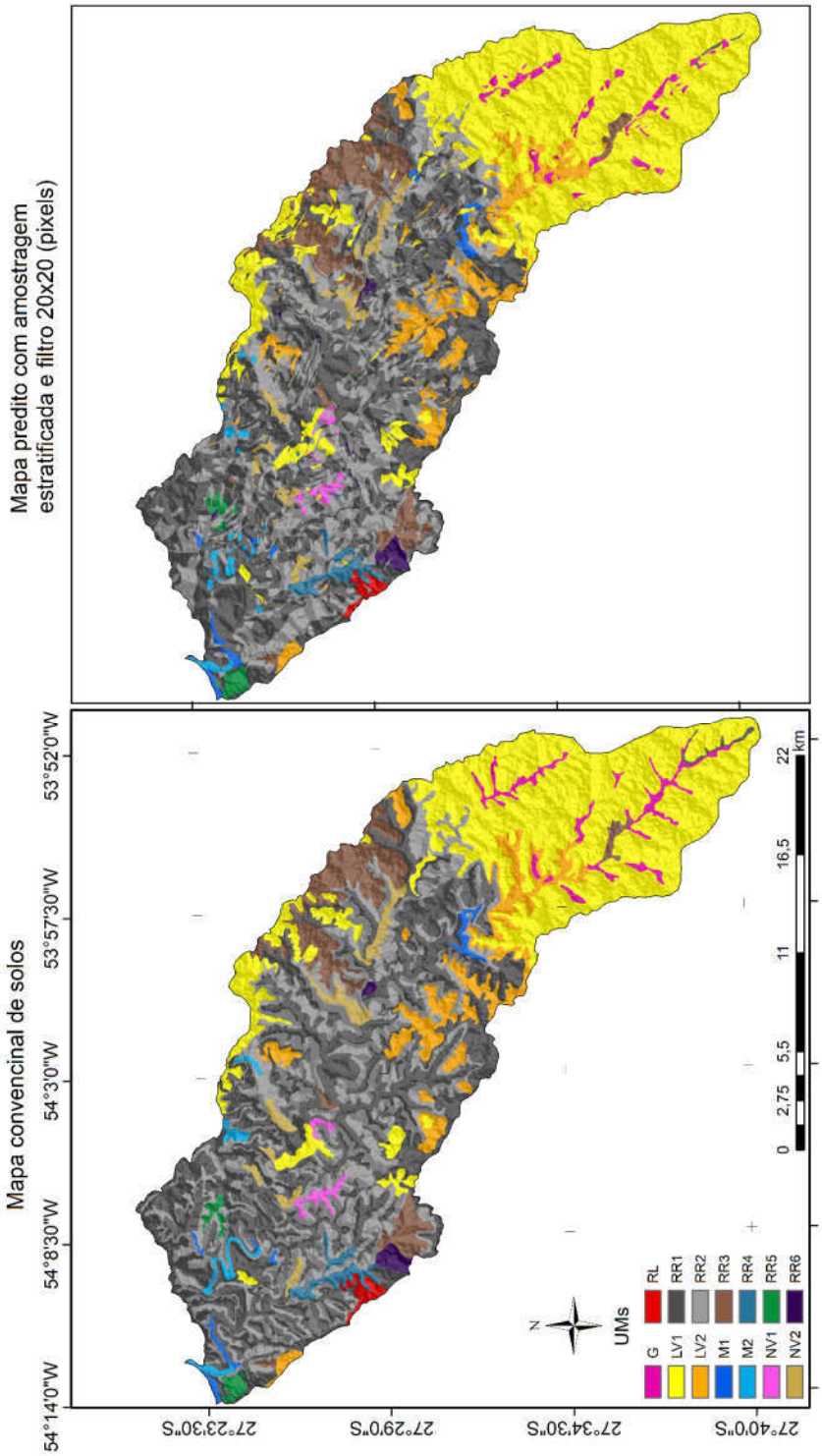


Figura 11 - Mapa convencional de solos e mapa predito com o conjunto de variáveis AEF20S da bacia Lajeado Grande.

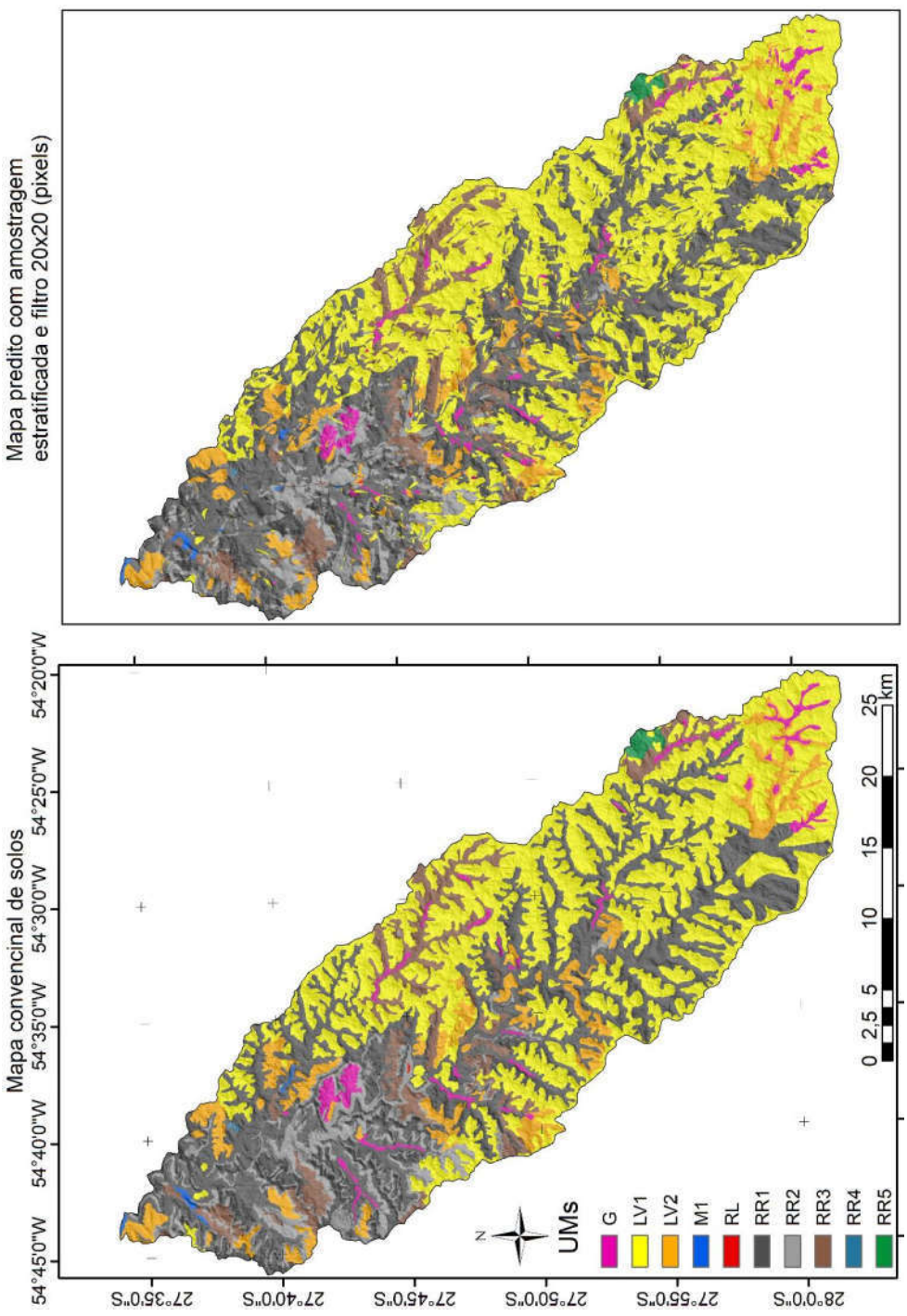


Figura 12 - Mapa convencional de solos e mapa predito com o conjunto de variáveis AEF20S da bacia Santo Cristo.

Na bacia do Santo Cristo, a acurácia geral alcançou 78% nos conjuntos AEF25S, AE30S e AAF30S, no entanto, 1% apenas superior ao obtido nos conjuntos AAF15S e AEF25S, indicando bom desempenho destes conjuntos também. O valor máximo de AG obtido na SC representa um ganho de 23,8% em relação aos mapas preditos com os conjuntos AAF0 e AEF0.

De modo geral, em ambas as bacias o filtro de média aplicado de forma sequencial resultou em AG ligeiramente superior à aplicação não sequencial. Os valores de AG obtidos nos conjuntos AAF0 e AEF0 foram semelhantes à média obtida em outros estudos no MDS (BEHRENS et al., 2010; COELHO; GIASSON, 2010; TEN CATEN et al., 2013; HÖFIG; GIASSON; VENDRAME, 2014; TESKE; GIASSON; BAGATINI, 2015a), confirmando a boa capacidade das variáveis preditoras utilizadas na predição de ocorrencia dos solos, mesmo sem aplicação dos filtros de média.

O aumento no tamanho do filtro, promoveu um ganho mais acentuado de AG até o filtro 15x15, sendo que a partir deste, o aumento foi inferior a 2%, indicando uma consolidação nos valores de AG. Esse comportamento foi observado para as duas bacias, evidenciando que no presente estudo, as variáveis preditoras geradas dos MDEs filtrados com tamanho superiores a 15x15 foram as que mais se correlacionaram com a distribuição das UMs na paisagem, indicando que a partir do 15x15 foi atingida a resolução do mapa original de solos utilizado com referência, e concordando com os resultados obtidos em estudos de Behrens et al. (2010, 2014), nos quais as variáveis preditoras com suavização apresentaram maior capacidade na predição de ocorrência das classes de solos.

O coeficiente Kappa apresentou comportamento semelhante à AG, com maiores valores nos mapas preditos com as variáveis geradas dos MDEs que receberam aplicação do filtro de média de forma sequencial. Na bacia LG, o maior valor do coeficiente Kappa (0,70) foi alcançado nos conjuntos AEF20S, AEF25S, AEF30S e AAF30S (Tabela 5). Para a bacia SC, o valor máximo do coeficiente Kappa (0,69) foi obtido com os conjuntos AEF30S e AAF30S (Tabela 5).

Nos mapas que utilizaram os conjuntos obtidos a partir do filtro 20x20 houveram pequenos aumentos no coeficiente Kappa, com a relação AG/Kappa reduzindo com o aumento no tamanho do filtro. Este resultado indica redução

nas classificações aleatórias com o aumento no tamanho de filtro aplicado ao MDE, evidenciando maior consistência nas classificações realizadas pelo modelo preditor.

Para os diferentes esquemas de amostragem, não foram observadas diferenças na AG e coeficiente Kappa, no entanto, esse é um resultado possível, principalmente por estes índices de concordância serem bastante influenciados em dados com classes desbalanceadas, onde poucas classes são predominantes (CHAWLA et al., 2002). A predição correta das classes dominantes em dados desbalanceados resulta em AG próximo a proporcionalidade dessas classes no total a ser predito. Portanto, a definição do melhor resultado em dados desbalanceados não pode ser definida apenas com base nestes índices, sendo necessária a avaliação de outros índices, como acurácia do mapeador.

4.3.3. Acurácia do mapeador dos mapas preditos

Como observado para acurácia geral e coeficiente Kappa, os filtros aplicados de forma sequencial também resultaram em mapas com acurácia do mapeador (AM) ligeiramente superiores aos valores obtidos nos mapas preditos com as variáveis geradas dos MDEs com filtros não sequenciais. Para a maioria das UMs, os máximos valores de AM foram obtidos nos mapas preditos com os conjuntos que receberam filtro com tamanho igual ou superior ao 15x15 (Tabela 6), entretanto, com pouca variação entre eles, indicando estabilização a partir deste tamanho de filtro, comportamento semelhante ao apresentado pela acurácia geral.

O aumento no tamanho do filtro resultou em ganho na AM nos mapas preditos para a bacia Lajeado Grande, sendo que apenas na unidade G houve redução (Tabela 6), indicando que para a predição dessa UM as variáveis geradas do MDE com maior nível de detalhe apresentaram maior correlação com sua distribuição espacial. Esse comportamento pode ser atribuído ao tamanho e formato dos polígonos dessa unidade de mapeamento de solos, que apresentam pequenas áreas com formatos alongados que ocorrem geralmente paralelamente a rede de drenagem, característica que reduziu sua correlação com as variáveis preditoras à medida que o MDE foi suavizado (BEHRENS et al., 2010).

A aplicação do filtro 20x20 sequencial (AEF20S e AAF20S) resultou em ganho médio na AM de 5%, 42%, 41%, 22%, 118%, 117%, 228%, 15%, 89%, 135%, 133% e 424% nas unidades LV1, LV2, M1, M2, NV1, NV2, RL, RR1, RR3, RR4, RR5 e RR6 respectivamente, e redução média aproximada de 23% e 6% nas unidades G e RR2 respectivamente. Esses resultados reforçam que os conjuntos AEF0 e AAF0 derivados do MDE no seu nível original de detalhe não foram adequados para predição das UMs, confirmando a necessidade de pré-processamento para adequar as variáveis à escala de ocorrência espacial dos solos, concordando com outros estudos (BEHRENS et al., 2005; GRINAND et al., 2008).

Tabela 6 - Acurácias do mapeador dos mapas preditos com os respectivos conjuntos de variáveis preditoras para a bacia do rio Lajeado Grande.

Conjuntos de dados	Unidades de mapeamento de solos														Média
	G	LV1	LV2	M1	M2	NV1	NV2	RL	RR1	RR2	RR3	RR4	RR5	RR6	
AEF0	57	87	50	53	47	27	40	32	65	66	41	33	36	13	46
AEF5	51	89	62	64	59	33	50	57	69	68	56	46	46	31	56
AEF10	48	90	68	59	59	39	52	61	70	65	71	60	51	63	61
AEF15	45	91	70	60	59	50	57	66	73	61	73	55	74	73	65
AEF20	45	91	71	63	60	56	63	73	72	62	78	73	75	80	69
AEF25	36	93	72	56	70	43	63	80	75	64	77	66	61	81	67
AEF30	46	92	72	55	61	63	62	77	75	63	81	69	71	78	69
AEF10S	49	91	69	70	56	47	56	68	73	65	74	66	66	74	66
AEF15S	52	92	73	68	68	58	65	86	75	65	79	70	63	82	71
AEF20S	39	93	70	61	55	66	78	87	77	65	82	61	79	84	71
AEF25S	47	92	76	59	75	74	74	74	78	60	83	60	81	85	73
AEF30S	42	93	73	55	75	62	75	79	77	61	82	71	73	83	71
AAF0	49	88	53	37	44	32	26	18	64	67	42	29	28	21	42
AAF5	51	89	61	42	52	30	46	35	68	71	55	48	55	35	53
AAF10	46	90	67	51	48	42	51	53	69	67	70	61	60	63	60
AAF15	46	90	72	44	46	48	61	73	72	63	73	55	84	70	64
AAF20	45	92	71	57	45	62	56	70	74	63	73	70	76	78	66
AAF25	40	91	72	41	48	65	59	82	72	65	79	58	80	89	67
AAF30	46	91	71	47	46	78	65	64	74	63	81	76	64	75	67
AAF10S	48	91	69	45	48	52	50	62	72	67	72	62	68	75	63
AAF15S	48	93	69	58	58	66	63	81	68	68	80	65	67	76	69
AAF20S	37	93	75	60	51	73	72	77	76	63	79	72	72	91	71
AAF25S	34	93	69	55	56	58	71	70	75	65	78	67	80	85	68
AAF30S	40	92	75	54	47	71	71	82	79	61	81	73	83	91	72
Área (ha)	861	14529	3116	360	390	242	896	220	16202	12767	2810	457	278	259	-
Área (%)	1,61	27,21	5,84	0,67	0,73	0,45	1,68	0,41	30,35	23,91	5,31	0,86	0,52	0,49	-

Na bacia LG, o maior valor médio de AM (73%) foi obtido com o conjunto AEF25S (amostragem estratificada), sendo que para esse mapa apenas as

unidades G e M1 obtiveram AM inferior a 60% (Tabela 6). Para a amostragem aleatória, o maior valor médio de AM (72%) foi obtido com o conjunto AAF30S, sendo que nesse mapa as unidades G, M1 e M2 obtiveram AM inferior a 60%. Os máximos valores médios indicam maior distribuição do erro entre as unidades de mapeamento de solos, consequentemente, resultando em um mapa com maior número de UMs individualizadas.

Quando comparados os dois esquemas de amostragem no mesmo tamanho de filtro, adotando como critério o primeiro mapa a atingir a máxima acurácia geral (AEF20S e AAF20S), foi possível constatar que o uso da amostragem aleatória resultou em melhor predição nas unidades LV2, NV1, RR4 e RR6, sendo que outras dez UMs restantes foram melhor preditas quando utilizada a amostragem estratificada. Nesse contexto, foi possível afirmar que a alocação de pontos pela amostragem estratificada em todos os polígonos do mapa de solos aumentou o desempenho do modelo de predição, resultando em melhor predição para a maioria das UMs e, principalmente, nas de menor extensão.

Na bacia SC**,** o comportamento da AM foi semelhante ao observado para a LG, com os menores valores médios da AM (46% e 35%) observados nos mapas preditos com os conjuntos AEF0 e AAF0, e os maiores valores médios (70% e 65%) obtidos nos conjuntos AEF25 e AAF25S (Tabela 7). Na bacia SC ocorreu maior variação nos valores de AM, com total ausência na predição de algumas unidades de mapeamento de solos.

No mapa predito com o conjunto AEF0 apenas três unidades de mapeamento (LV1, RL, RR1) obtiveram AM superior a 45% e no conjunto AAF0 apenas duas unidades (LV1 e RR1). Esses resultados indicam uma maior complexidade na predição da ocorrência dos solos na bacia SC e a baixa capacidade do MDE sem aplicação do filtro em diferenciar as UMs, sendo que apenas as UMs de maior extensão obtiveram boa predição.

Para o mapa obtido com o conjunto AEF25S apenas as unidades RL e RR4 apresentam valor de AM inferior a 59% (Tabela 7), indicando boa predição na maioria das unidades de mapeamento de solos. Quando foi utilizada a amostragem aleatória (AAF25S), todas as unidades obtiveram AM superior a 57%, no entanto, com maiores erros nas UMs de menor extensão e o total erro de predição na unidade RL (Tabela 7).

Quando comparando os esquemas de amostragem nos mapas obtidos com os conjuntos AEF25S e AAF25S, observa-se que a amostragem estratificada reduziu a acurácia do mapeador apenas nas unidades RR3, RR4 e RR5, com essas UMs representando apenas 8% de toda a área da bacia do Santo Cristo. Esses resultados reforçam a baixa eficiência da amostragem aleatória na predição de UMs de menor extensão, como observado em outros estudos (COELHO; GIASSON, 2010; TEN CATEN et al., 2011c, 2011a; HÖFIG; GIASSON; VENDRAME, 2014; BAGATINI; GIASSON; TESKE, 2015).

Tabela 7 - Acurácias do mapeador obtidas nos mapas preditos em cada conjunto de variáveis preditoras para a bacia do rio Santo Cristo.

Conj. Preditoras	Unidades de mapeamento de solos										Média
	G	LV1	LV2	M1	RL	RR1	RR2	RR3	RR4	RR5	
AEF0	38	80	33	40	92	67	44	27	0	43	46
AEF5	51	82	56	62	91	71	53	42	11	75	59
AEF10	53	82	61	45	0	72	53	55	25	74	52
AEF15	52	82	66	52	26	74	56	59	30	81	58
AEF20	58	82	69	59	43	76	57	66	38	81	63
AEF25	56	84	72	56	78	75	60	67	77	77	70
AEF30	59	83	73	51	48	76	60	68	40	88	65
AEF10S	62	83	72	55	84	75	58	64	42	80	67
AEF15S	59	84	71	60	40	77	57	66	37	86	64
AEF20S	59	85	74	50	50	77	62	71	33	92	65
AEF25S	65	85	74	78	52	78	59	68	41	87	69
AEF30S	63	85	77	71	0	76	63	75	40	83	63
AAF0	41	79	36	30	0	67	43	28	0	31	35
AAF5	51	81	55	40	94	71	55	45	48	75	62
AAF10	53	80	62	47	57	71	56	58	66	75	62
AAF15	59	82	66	44	0	75	52	61	0	77	52
AAF20	56	81	69	44	56	76	55	63	0	89	59
AAF25	59	83	72	65	0	75	59	69	0	78	56
AAF30	59	83	73	52	45	76	56	70	26	84	62
AAF10S	61	84	70	33	29	74	56	67	67	73	61
AAF15S	63	83	74	48	0	77	58	71	60	89	62
AAF20S	63	84	73	43	0	77	60	72	34	80	59
AAF25S	65	84	73	58	0	77	59	70	69	92	65
AAF30S	64	85	75	63	0	78	58	73	56	84	64
Área (ha)	2213	34347	7134	179	12	31334	7422	7064	29	340	-
Área (%)	2,46	38,13	7,92	0,20	0,01	34,79	8,24	7,84	0,03	0,38	-

A análise visual dos mapas preditos não permite distinção entre os sistemas de amostragem ou as formas de aplicação do filtro de média, no entanto, é possível ver de forma clara as modificações promovidas pelos diferentes tamanhos de filtros utilizados, sendo essa diferença visual mais

contrastante nos mapas preditos com os conjuntos AEF0, AEF20S e AEF30S, motivo pelo qual somente estes mapas estão apresentados na Figura 13.

Para ambas as bacias, os mapas preditos com o conjunto AEF0 (Figura 13B e 13F), apresentam dominância das unidades LV1 e RR1, com pouca distinção entre as unidades RR1 e RR2. A partir do filtro 15x15, com a maior suavização dos dados preditores, ocorre uma maior continuidade espacial das manchas de solos das UM de menor extensão, contribuindo para um maior nível de distinção entre as classes de solos.

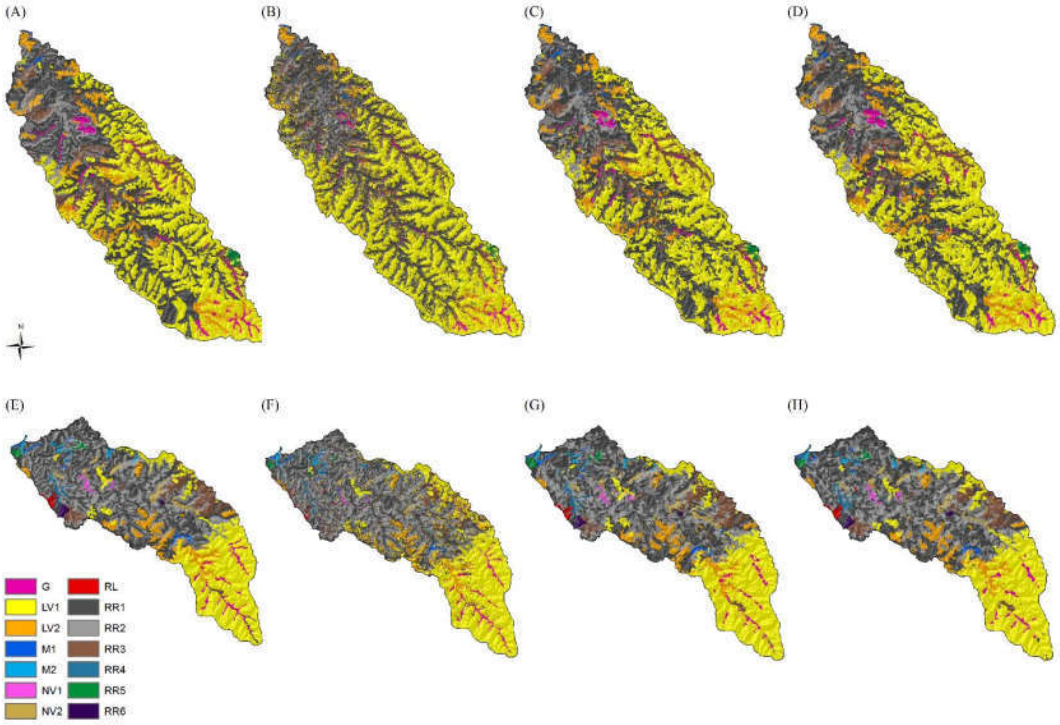


Figura 13 - Mapa convencional de solos das bacias do rio Santo Cristo (A) e Lajeado Grande (E) e os mapas preditos com os conjuntos AEF0 (B e F), AEF20S (C e G), AEF30S (D e H).

Na bacia SC, as unidades G, LV2, RR2 e RR6 dos mapas preditos com os conjuntos AEF20S e AEF30S (Figura 13C e 13D) apresentam maior semelhança visual com o mapa convencional (Figura 13A). Para a bacia LG, as unidades LV2, M1, RL e RR6 dos mapas preditos com os conjuntos AEF20S e AEF30S (Figura 13G e 13H) apresentam maior semelhança com o mapa convencional (Figura 13E). A análise visual confirma a maior concordância observada nos valores de acurácia geral, acurácia do mapeador e coeficiente Kappa e é

atribuída a maior aproximação entre a escala de detalhe das variáveis com as relações solo-paisagem e da escala do mapa convencional de referência.

Como constatado, tanto nos índices de concordância como na análise visual dos mapas, o conjunto filtro 20x20 foi o mais adequando para predição da ocorrência dos solos na bacia LG, e o filtro 25x25 foi o mais adequando na bacia SC, resultando em predição com boa concordância entre os mapas preditos e o mapa de referência. Estes resultados indicam que para ambas às áreas de estudo este nível de detalhe nas preditoras foi mais adequado para discriminação das unidades de mapeamento de solos e reforça o pressuposto da necessidade de adequação das variáveis preditoras à escala de ocorrência das classes de solos na paisagem (BEHRENS et al., 2014; MILLER et al., 2015).

Os esquemas de amostragem não alteraram os índices de concordância do coeficiente Kappa e da acurácia geral, no entanto, foi possível identificar maior capacidade da amostragem estratificada em predizer unidades de menor extensão, sem prejuízo na predição das unidades de maior extensão. Esse comportamento da amostragem estratificada foi de grande importância para as duas áreas e, principalmente, para a bacia do Lajeado Grande, onde às áreas de menor extensão podem representam centenas de hectares.

4.4. CONCLUSÕES

- Os diferentes tamanhos do filtro de média promoveram alteração no nível de detalhe do modelo digital de elevação;
- Os máximos valores de concordância dos mapas preditos foram obtidos nos conjuntos de variáveis que receberam aplicação dos tamanhos de filtro 20x20, 25x25 e 30x30, aplicados de forma sequencial;
- Os sistemas de amostragem não apresentaram diferenças quanto à acurácia geral e coeficiente Kappa.
- A aplicação do filtro de média sequencial com tamanhos 20x20, 25x25 e 30x30 promoveu ganho médio de 14,9 e 23,2% na acurácia geral dos mapas preditos das bacias Lajeado Grande e Santo Cristo quando foi utilizada a amostragem estratificada e ganho médio de 14,4% e 22,7% quando foi utilizada a amostragem aleatória.

5. CAPITULO IV – ESTUDO 3: PREDIÇÃO DE CLASSES DE SOLOS COM DADOS COLETADOS EM PIXELS DELIMITADOS POR *BUFFERS* EM PERFIS DE SOLO GEORREFERENCIADOS

5.1. INTRODUÇÃO

No mapeamento digital de classes de solos, diversos estudos têm testado as mais variadas abordagens buscando, principalmente, a reprodução de mapas legados, objetivando criar modelos matemáticos capazes de reproduzir o modelo criado pelo pedólogo na delimitação das classes de solos. Estes modelos preditores podem ser extrapolados para áreas fisiograficamente semelhantes que não disponham de mapas de solos (GIASSON et al., 2006; BEHRENS et al., 2010; ARRUDA et al., 2013; DIAS et al., 2016).

No entanto, essa abordagem exige a existência de mapas legados para calibração dos modelos de predição, sendo inviável para regiões onde não há disponibilidade dessas informações. Uma alternativa para a ausência de mapas legados de solos é a predição da ocorrência dos solos a partir de informações coletadas em pontos de perfis georreferenciados (HENGL et al., 2007; ALVES; DEMATTÊ; BARROS, 2015; ARRUDA et al., 2016). Dessa forma, pode-se realizar o mapeamento de áreas que já foram amostradas, além da confecção de bancos de dados estruturados e a extrapolação para áreas fisiograficamente semelhantes.

O uso dessa alternativa pode ser comprometido em áreas com pouca disponibilidade de perfis georreferenciados, pois um pequeno número de perfis pode ser insuficiente para uma boa calibração dos modelos preditores (TESKE; GIASSON; BAGATINI, 2015b), havendo a necessidade de um pedólogo experiente na delimitação das classes de solos, o que poderia levar a realização de procedimentos semelhantes aos adotados nos levantamentos tradicionais de

solos, demandando mais tempo e recursos (TESKE; GIASSON; BAGATINI, 2015b).

A baixa disponibilidade de perfis georreferenciados pode ser contornada com a utilização de estratégias de amostragem mais representativas nas áreas de localização de cada perfil, permitindo a coleta de pontos amostrais suficientes para o uma boa calibração dos modelos preditivos. Uma estratégia é a delimitação de *buffers* em torno de cada perfil, para obter um maior número de amostras nas áreas vizinhas aos perfis de solos georreferenciados (CHAGAS et al., 2010; DIAS et al., 2016). O *buffer* é uma técnica empregada na delimitação de áreas de interesse para estudos em diversos ramos da ciência, como por exemplo, na delimitação de áreas para amostragem de solos no campo (CARVALHO JÚNIOR et al., 2014).

O uso do *buffer* permite a estrapolação da classificação do perfil de solo para as áreas adjacentes a este, permitindo a coleta de um maior número de pontos amostrais o que pode levar a uma maior eficiência dos modelos preditores aplicados para predição de ocorrência das classes de solos. Nesse contexto, o objetivo deste estudo foi avaliar o desempenho na predição de ocorrência de solos com amostras coletadas em pixels de perfis de solos georreferenciados e em pixels coletados em *buffers* com raio de 50, 100, 150, 200 e 250 m, dos perfis de solos nas bacias dos rios Lajeado Grande e Santo Cristo.

5.2. MATERIAL E MÉTODOS

Foram utilizados dados de solos, hidrográficos e do modelo digital de elevação (MDE) das Bacias dos rios Santo Cristo (SC) e Lajeado Grande (LG), ambas inseridas na Bacia Hidrográfica U030, localizada na região noroeste do Estado do Rio Grande do Sul. A geologia das duas áreas corresponde à Província do Paraná, caracterizada principalmente por derrames basálticos da formação Serra Geral (FREITAS et al., 2012).

Às áreas correspondem a 900,7 km^2 na Bacia do Santo Cristo e 533,3 km^2 na Bacia do Lajeado Grande, e possuem mapas de solos na escala 1:50.000 (KÄMPF; GIASSON; STRECK, 2004a, 2004b). No Quadro 7 está descrita a composição das unidades de mapeamento de solos (UM) que ocorre nos mapas das respectivas áreas de estudo.

Quadro 7 - Unidades de mapeamento de solos que ocorrem nas áreas das bacias dos rios Lajeado Grande e Santo Cristo.

UM	Composição	Proporção	Inclusões	Área (%)
	Bacia hidrográfica do rio Lajeado Grande (LG)			
G	Gleissolos			1,61
LV1	Latossolo Vermelho distroférrico		RR, CX	27,21
LV2	Latossolo Vermelho + Neossolo Regolítico *	60/40	CX	5,84
M1	Chernossolo		RR, RF	0,67
M2	Chernossolo + Neossolo Regolítico*	60/40	CX, RF	0,73
NV1	Nitossolo Vermelho			0,45
NV2	Nitossolo Vermelho + Neossolo Regolítico*	60/40		1,68
RL	Neossolo Litólico + Neossolo Regolítico*		AR	0,41
RR1	Neossolo Regolítico		AR, RL, CX, M, LV	30,35
RR2	Neossolo Regolítico + Neossolo Litólico, relevo forte ondulado*	60/40	AR, CX	23,91
RR3	Neossolo Regolítico + Latossolo Vermelho*	60/40	CX	5,26
RR4	Neossolo Regolítico + Chernossolo*	60/40	CX, RF	0,86
RR5	Neossolo Regolítico + Cambissolo + Nitossolo Vermelho*	50/30/20		0,52
RR6	Neossolo Regolítico + AR *	70/30		0,49
Área total (ha)				53.388
	Bacia hidrográfica do rio Santo Cristo (SC)			
G1	Gleissolos			2,46
LV1	Latossolo Vermelho distroférrico		RR, CX	38,13
LV2	Latossolo Vermelho + Neossolo Regolítico*	60/40	CX	7,92
M1	Chernossolo Háplico		CX	0,20
RL	Neossolo Litólico			0,01
RR1	Neossolo Regolítico + Cambissolo Háplico*	60/40	AR, RL, CX	34,79
RR2	Neossolo Regolítico + Neossolo Litólico, relevo forte ondulado**	50/50	AR, CX	8,24
RR3	Neossolo Regolítico + Latossolo Vermelho*	60/40	CX	7,84
RR4	Neossolo Regolítico + Neossolo Litólico*	70/30		0,03
RR5	Neossolos Regolíticos + Cambissolo Háplico + Latossolo Vermelho*	50/30/20		0,38
Área total (ha)				90.073

*Associações; **Complexos; AR = afloramento de rocha; CX = Cambissolo; LV = Latossolo Vermelho; M = Chernossolo; RY = Neossolo Flúvico; RL = Neossolo Litólico; RL = Neossolo Litólico; RR = Neossolo Regolítico

O MDE utilizado foi produzido a partir de dados do sensor *Aster/GDEM* v2 (*Global Digital Elevation Models*), com resolução espacial de 30 metros, obtidos no Serviço Geológico Americano (TACHIKAWA et al., 2011), a partir do qual foram derivadas oito variáveis preditoras (nível da base da rede de drenagem, orientação de vertentes, insolação difusa, abertura positiva do terreno, convexidade, declividade, índice de convergência e posição média da declividade). Além destas, foram derivadas as variáveis índice de densidade de drenagem (OTTO et al., 2017) e a distância euclidiana dos rios a partir da rede hidrográfica na escala 1:50.000 (HASENACK; WEBER, 2010). As variáveis

selecionadas para o estudo tiveram como critério os resultados obtidos no Estudo 1 (Capitulo II), no qual o conjunto de variáveis listado apresentou a maior eficiência na predição de ocorrência dos solos.

Para delimitação dos *buffers* e coleta dos pixels amostrais foram utilizados pontos de observação georreferenciados (perfis de solos e amostras extras) com os solos classificados no nível de subordem (EMBRAPA, 2013), sendo 169 na bacia LG e 196 na SC. Para coleta dos dados para treinamento foram eliminados pontos vizinhos com distância inferior a 500 m, de forma a selecionar amostras para todas as classes descritas nos levantamentos de solos (KÄMPF; GIASSON; STRECK, 2004b, 2004a), restando 142 pontos na LG e 157 na SC, distribuídos em oito classes taxonômicas de solos (Figura 14).

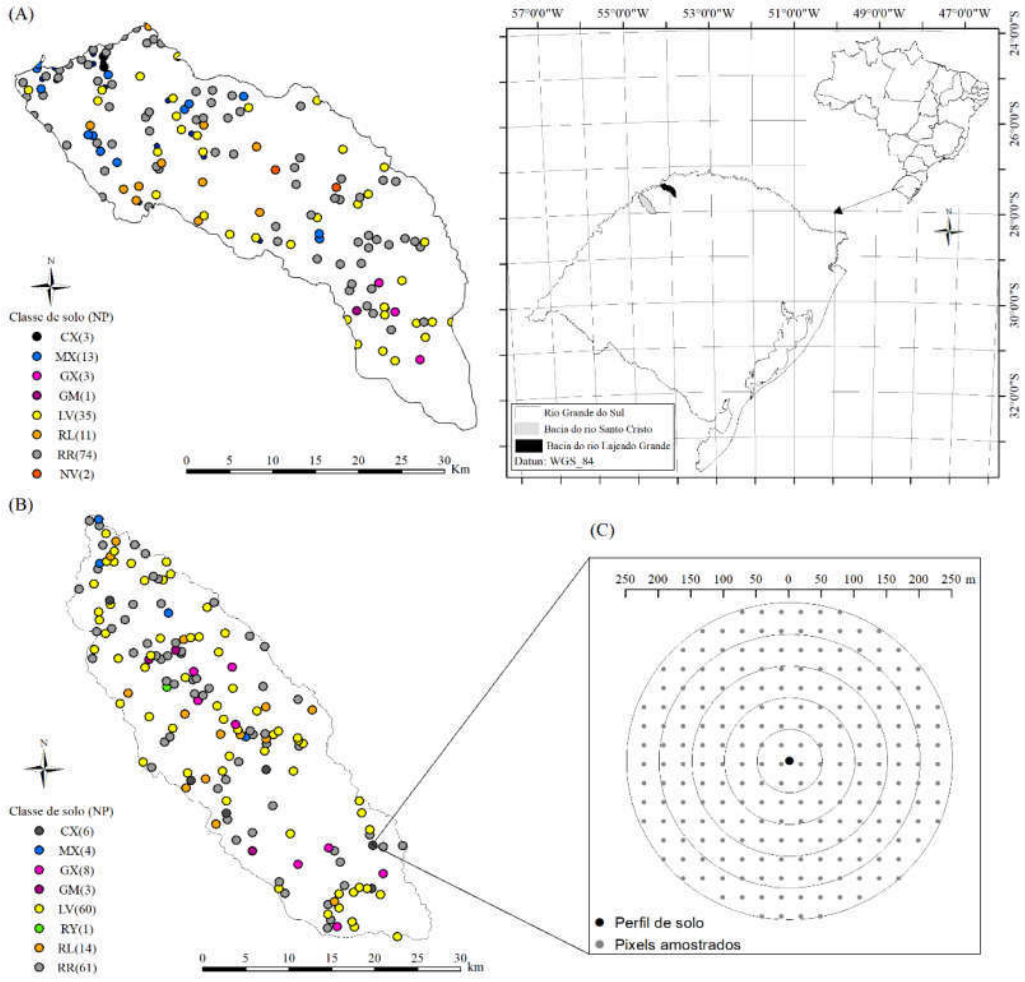


Figura 14 - Mapa com o número e distribuição dos perfis de solos georreferenciados das bacias Lajeado Grande (A) e Santo Cristo (B), e o esquema dos *buffers* (C) utilizados para coleta dos pixels amostrais. NP - número de perfis de solos.

Foram testados cinco *buffers* com raios de 50, 100, 150, 200, 250 m, em torno dos perfis georreferenciados (Figura 14C), além dos pontos sem *buffer* (BF00), totalizando cinco bases de pontos amostrais (BF50, BF100, BF150, BF200 e BF250). Na amostragem das áreas delimitadas pelos *buffers*, foram considerados todos os pixels dentro do raio de alcance dos respectivos *buffers* (Figura 14). Com esses dados foram preditos mapas de ocorrência de classes taxonômicas de solos (CT00, CT50, CT100, CT150, CT200 e CT250) a partir da classificação dos perfis de solos no nível de subordem (EMBRAPA, 2013), e mapas de unidades de mapeamento de solos (UM00, UM50, UM100, UM150, UM200 e UM250) tendo como referência os mapas convencionais disponíveis para às áreas de estudo. Para a predição foi utilizando o classificador *Random Forest* do pacote *RandomForest* (BREIMAN et al., 2018) para ambas as áreas de estudo.

Para validação dos mapas preditos foram removidos todos os agrupamentos de pixels com menos de cinco hectares. As concordâncias foram avaliadas quanto a exatidão em relação aos perfis de solos e a reprodutibilidade dos mapas convencionais disponíveis para as áreas de estudo. Na concordância com os perfis (exatidão) foram utilizados todos os perfis das bacias LG (169) e da SC (196), sendo que os mapas preditos de CT foram comparados de forma direta com os perfis, e os mapas preditos de UM, avaliados com base na composição de cada UM, como pode ser visto no Quadro 2.

Quadro 8 - Classes taxonômicas consideradas como corretas para avaliação da reprodutibilidade dos mapas de unidades de mapeamento de solos com os perfis de solos.

CT	Unidades de mapeamento de solos													
	G	LV1	LV2	M1	M2	NV1	NV2	RL	RR1	RR2	RR3	RR4	RR5	RR6
CX													X	
MX				X	X									
GX	X													
GM	X													
LV		X	X								X		X	
RY														
RL								X	X [(1)]	X		X		
RR			X		X		X	X	X	X	X	X	X	X
NV						X	X							

CT – Classe taxonômica de solo, CX = Cambissolos, LV = Latossolo Vermelho, MX = Chernossolo Háplico, RY= Neossolo Flúvico, RL = Neossolo Litólico, RR = Neossolo Regolítico, NV – Nitossolo Vermelho, [(1)] associação ocorre apenas no mapa da bacia do rio Santo Cristo.

Na avaliação de concordância dos mapas preditos de ocorrência de classes taxonômicas com os mapas convencionais das respectivas áreas, foram considerados como acertos a predição de uma ou mais das classes que compõe as respectivas UM (Quadro 8). Para avaliação da reprodutibilidade do mapa de UM, este foi comparado diretamente com mapa convencional de solos das áreas de estudo. De acordo com as formas de avaliação descritas acima foram geradas matrizes de erro, das quais foram calculados os valores de acurácia geral (AG) e acurácia do mapeador (AM) (CONGALTON, 1991).

5.3. RESULTADOS E DISCUSSÃO

A utilização de *buffers* com diferentes raios permitiu um aumento expressivo no número médio de pixels coletados por perfil de solo. Na bacia Lajeado Grande o valor médio aumentou para 209 pixels (30x30m) por perfil de solo no *buffer* de 250 m, e para 239 na bacia do rio Santo Cristo (Tabela 8).

Tabela 8 - Número médio de pixels coletados por perfil de solo nos cinco *buffers* testados.

Áreas de estudo	Total de perfis	BF00	BF50	BF100	BF150	BF200	BF250
Bacia do rio Lajeado Grande	141	1	9	54	75	135	209
Bacia do rio Santo Cristo	157	1	9	39	86	153	239

BF100, BF150, BF200, BF250 – *buffers* de 50, 100, 150, 200 e 250 metros, respectivamente.

Entretanto, o aumento no número de pixels não alterou de maneira significativa a proporcionalidade de pontos amostrais por unidade de mapeamento de solo (UM) ou nas classes taxonômicas de solos (CT), como pode ser observado na Figura 15. Para os pixels coletados nos mapas de UM, as unidades LV1, LV2, RR1 e RR2 representam aproximadamente 70 e 80% dos pontos amostrais nas bacias Lajeado Grande (Figura 15A) e Santo Cristo (Figura 15B), respectivamente. De modo geral, o número de amostras por UM está de acordo com a proporção observada nos mapas convencionais de solos. Para as unidades RL e RR6 não houve disponibilidade de perfis descritos localizados nas suas respectivas áreas, portanto estas não foram amostradas.

Nas classes taxonômicas de solos, a maior proporcionalidade de amostras ocorre nos Latossolos Vermelhos e Neossolos Regolíticos, que correspondem mais de 80% em ambas às áreas de estudo (Figura 15C e 15D). Com exceção da unidade RR5 na bacia Lajeado Grande e a RR4 e RR5 na bacia Santo Cristo,

foi possível predizer com uma área mínima de cinco hectares todas as demais classes que foram amostradas.

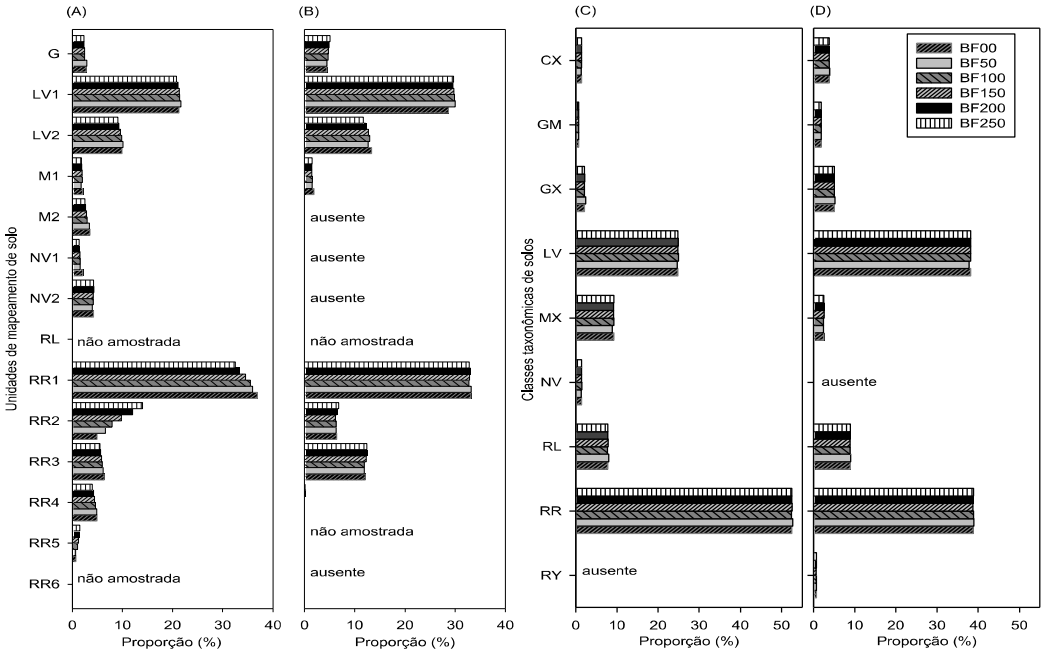


Figura 15 - Proporção de pixels coletados em cada buffer (BF) nas unidades de mapeamento e classes taxonômicas de solos nas bacias do Lajeado Grande (A e C) e Santo Cristo (B e D). CX – Cambissolo Háplico, MX - Chernossolo Háplico, GX – Gleissolo Háplico, GM – Gleissolo Melânico, LV – Latossolo Vermelho, RL – Neossolo Litólico, RR - Neossolo Regolítico e NV – Nitossolo Vermelho.

Na avaliação de exatidão dos mapas preditos de ocorrência de UM com os perfis de solos não houve expressiva alteração na acurácia geral (AG) com os diferentes raios de *buffers*. Os mapas preditos de UM com maior concordância (acurácia geral e do acurácia do mapeador) são apresentados nas Figuras 16 e 17. Entre os mapas UM00 e UM250 ocorreu uma pequena alteração de 3% e 6% nas bacias do Lajeado Grande e Santo Cristo, respectivamente (Tabela 9).

Na bacia Lajeado Grande, as acurácias do mapeador nas classes de maior extensão (LV1, LV2, RR1 e RR2) obtiveram valores superiores a 67% no mapa UM250. Para a bacia Santo Cristo as unidades LV1, LV2, RR1 e RR2 obtiveram respectivamente 65%, 88%, 49% e 64% de concordância no mapa UM200 (Tabela 9), sendo que apenas as unidades G e M1 apresentaram redução de concordância com o aumento no raio do *buffer*, indicando melhor predição com dados coletados apenas nos pixels dos perfis de solos.

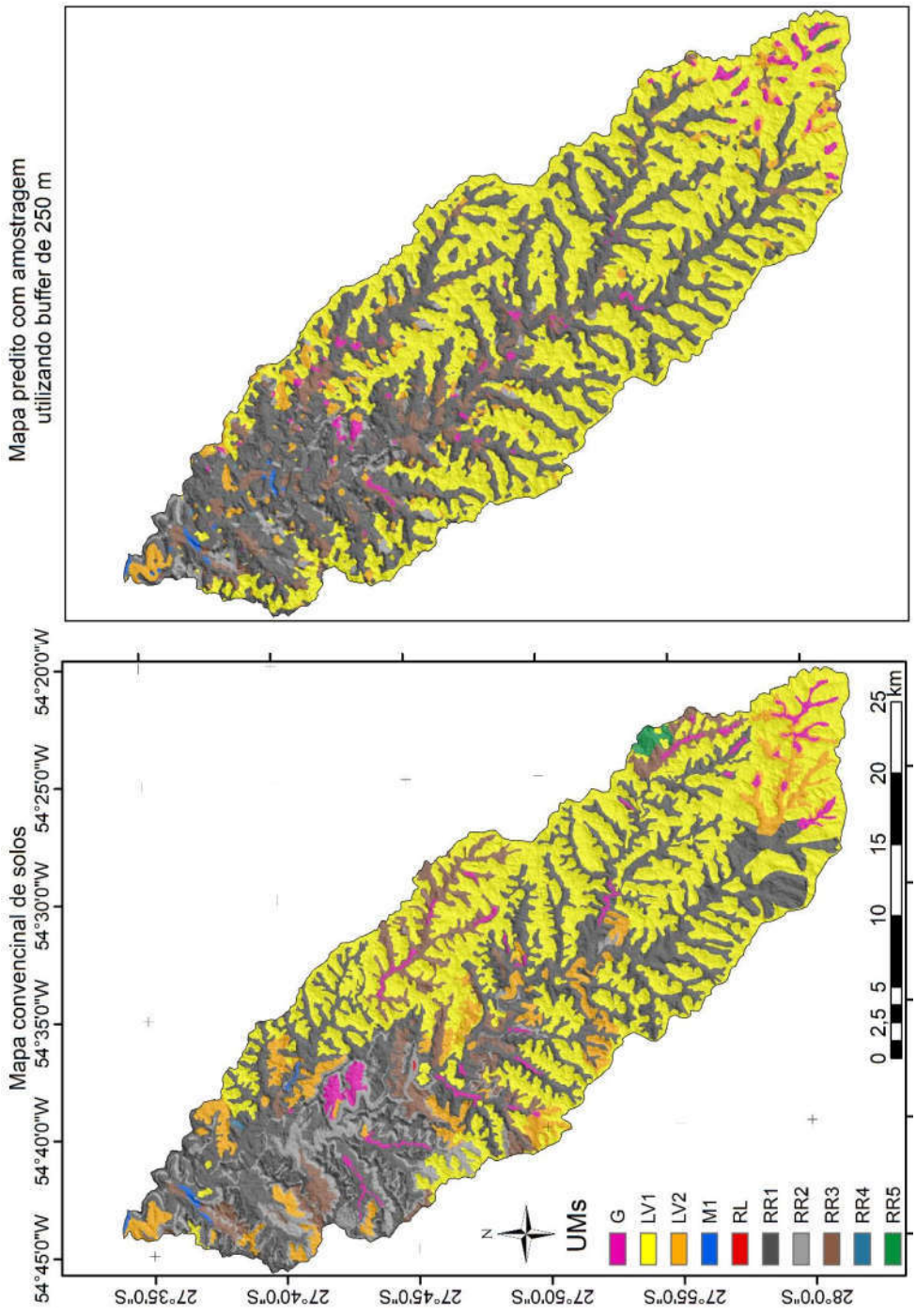


Figura 16 - Mapa convencional de solos e mapa predito de Unidades de Mapeamento com dados coletados no *buffer* de 250 m na área da bacia Santo Cristo. *Legenda igual para os dois mapas.

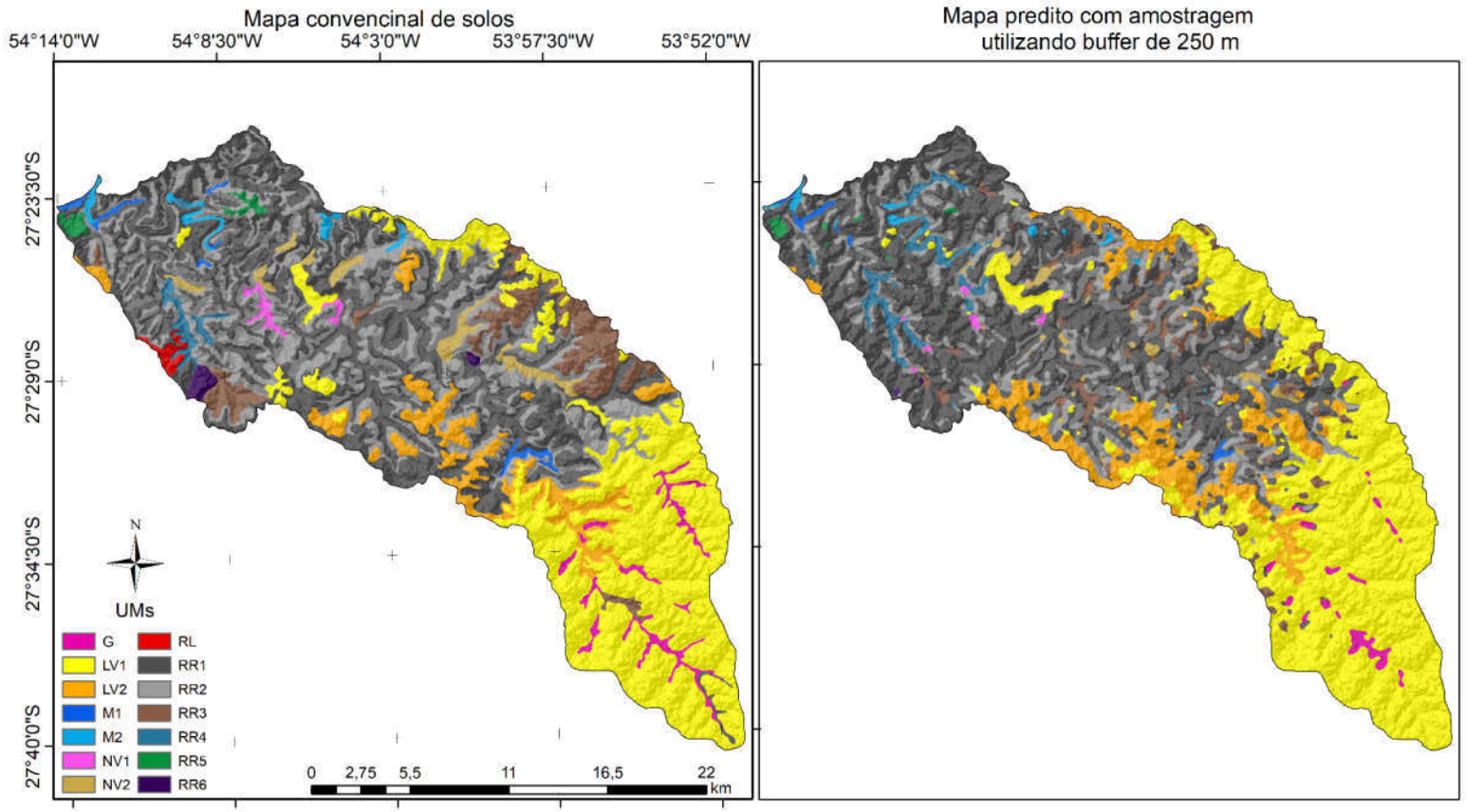


Figura 17 - Mapa convencional de solos e mapa predito de Unidades de Mapeamento com dados coletados no *buffer* de 250 m na área da bacia Lajeado Grande. *Legenda igual para os dois mapas.

Tabela 9 – Exatidão (%) dos mapas preditos de ocorrência das unidades de mapeamento de solos com os perfis de solos.

Mapas	Unidades de mapeamento de solos														Acurácia Geral
	G	LV1	LV2	M1	M2	NV1	NV2	RL	RR1	RR2	RR3	RR4	RR5	RR6	
Bacia hidrográfica do rio Lajeado Grande															
UM00	100	57	80	100	100	20	43	0	68	86	80	80	0	0	69
UM50	83	62	77	100	100	17	57	0	70	83	78	73	0	0	69
UM100	83	61	77	100	100	17	57	0	72	71	75	70	0	0	69
UM150	100	64	77	100	100	17	57	0	75	67	78	70	0	0	71
UM200	100	66	81	100	100	17	57	0	74	67	78	64	0	0	71
UM250	100	68	81	100	100	17	57	0	72	67	78	78	0	0	72
Área	1,6	27,2	5,8	0,7	0,7	0,5	1,7	0,4	30,3	23,9	5,3	0,9	0,5	0,5	533,7 (km^2)
Bacia hidrográfica do rio Santo Cristo															
UM00	67	59	79	100	-	-	-	0	50	50	71	0	0	-	57
UM50	63	58	84	50	-	-	-	0	46	71	80	0	0	-	59
UM100	50	62	83	75	-	-	-	0	50	64	83	0	0	-	62
UM150	50	64	88	75	-	-	-	0	48	64	83	0	0	-	62
UM200	50	65	88	75	-	-	-	0	49	64	80	0	0	-	63
UM250	50	65	88	75	-	-	-	0	51	54	82	0	0	-	63
Área	2,46	38,13	7,92	0,2	-	-	-	0,01	34,79	8,24	7,84	0,03	0,38	-	900,7 (km^2)

(-) não ocorre no mapa convencional.

Os valores de exatidão com os perfis de solos para a bacia Lajeado Grande são superiores ao valor médio de 60% encontrado em outros estudos com esse tipo de avaliação (CHAGAS et al., 2010; SILVEIRA et al., 2012; CHAGAS; OLIVEIRA; FERNANDES, 2013; ARRUDA et al., 2016). Para a bacia Santo Cristo os valores de AG estão próximos ao valor médio encontrado na literatura, mesmo quando a predição foi realizada apenas utilizando os pixels dos perfis de solos.

Os resultados indicam que o uso apenas dos perfis geram resultados semelhantes a outras técnicas de amostragem aplicadas no MDS, diferentes dos resultados encontrados por Teske, Giasson, Bagatini (2015b), no qual o mapa predito com os dados coletados nos pixels dos perfis não apresentou continuidade espacial nas manchas de solos, necessitando da delimitação manual dos polígonos de solos, sendo que no presente estudo não foi necessário essa etapa na produção do mapa de solos o que representa economia de tempo para produção destes mapas.

Na avaliação de exatidão dos mapas preditos de classes taxonômicas (Tabela 10) ocorreu pequenas alterações na acurácia geral da bacia Lajeado Grande, com os maiores valores de AG (91%) obtidos nos mapas CT100, CT150, CT200 e CT250. Para os mapas de maior concordância dessa área, apenas a classe NV apresentou concordância inferior a 65% (Figura 18 e 19).

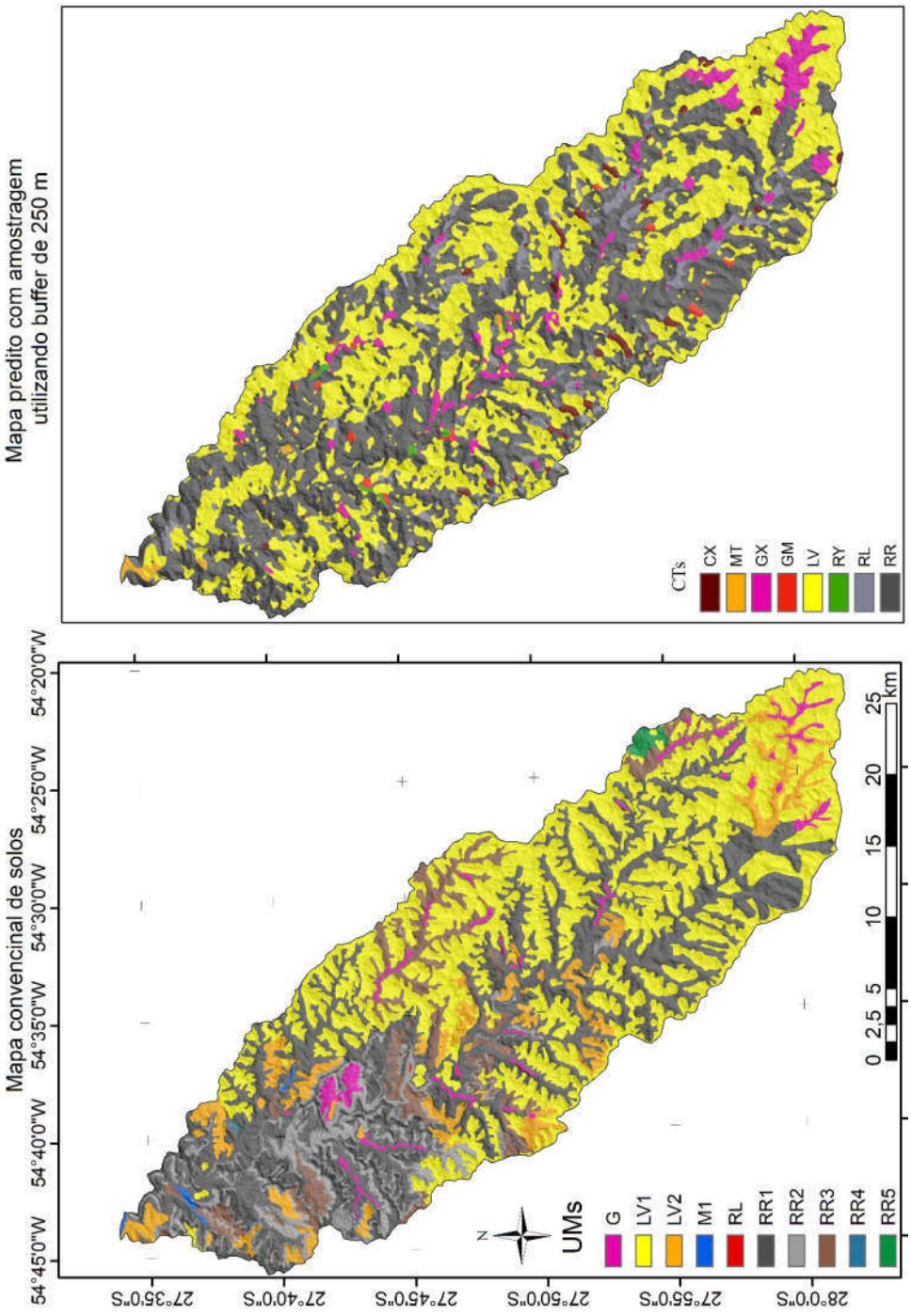


Figura 18 - Mapa convencional de solos e mapa predito de classes taxonômicas de solos com dados coletados no *buffer* de 250 m na área da bacia Santo Cristo.

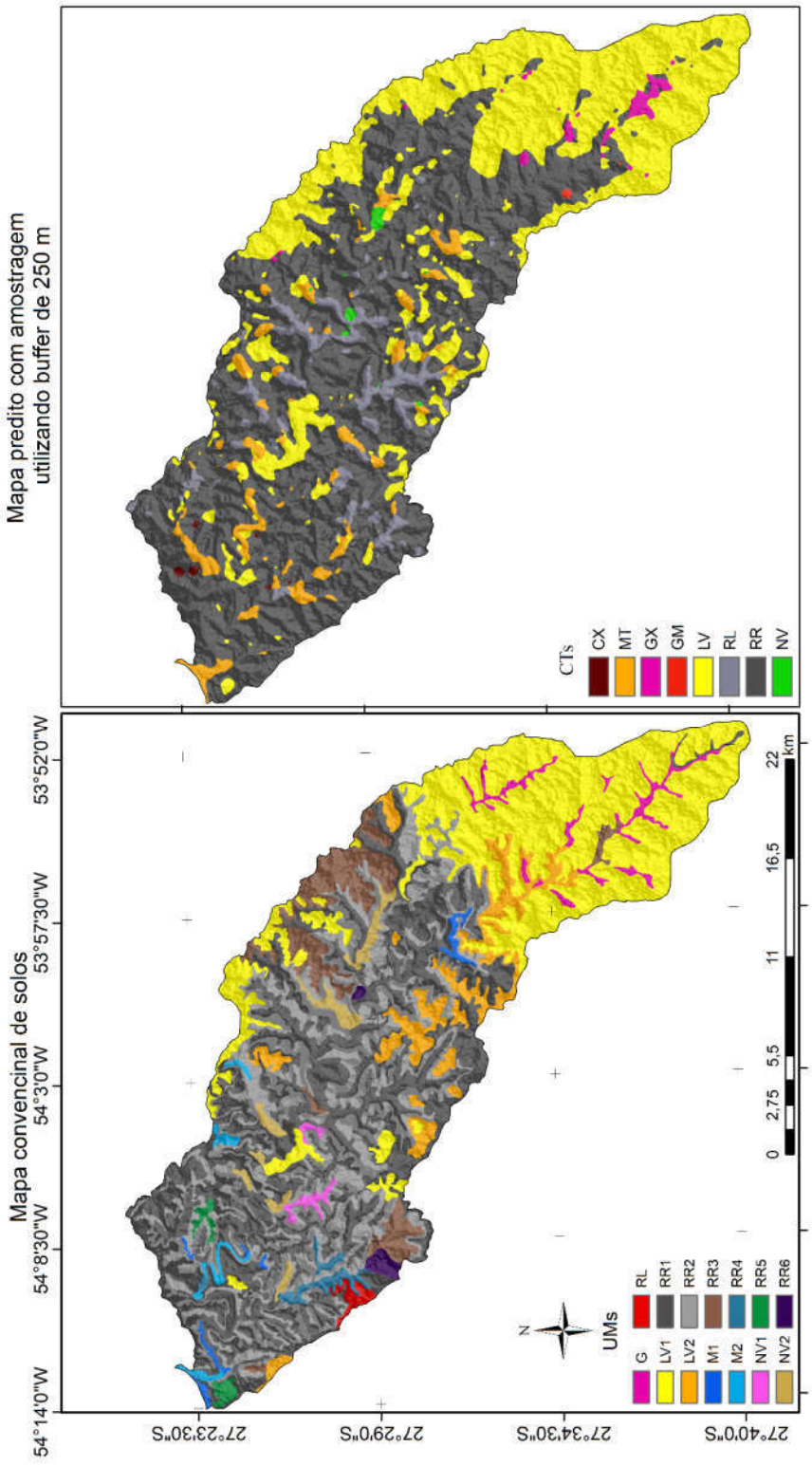


Figura 19 - Mapa convencional de solos e mapa predito de classes taxonômicas de solos com dados coletados no *buffer* de 250 m na área da bacia Lajeado Grande.

Tabela 10 - Exatidão (%) dos mapas preditos de classes taxonômicas de solos com os perfis de solos nas bacias dos rios Lajeado Grande e Santo Cristo.

Mapas	Unidades de mapeamento de solos									Acurácia Geral
	CX	MX	GX	GM	LV	RY	RL	RR	NV	
Bacia hidrográfica do Lajeado Grande										
CT00	0	94	100	100	82	-	24	100	0	82
CT50	33	88	100	100	95	-	59	99	20	89
CT100	67	88	100	100	97	-	65	98	40	91
CT150	67	88	100	100	100	-	65	98	40	91
CT200	67	88	100	100	100	-	65	96	40	91
CT250	67	88	100	100	100	-	65	96	40	91
Nº de Perfis	3	16	3	2	38	0	17	85	5	168
Bacia hidrográfica do Santo Cristo										
CT00	14	67	100	50	95	0	10	91	-	77
CT50	64	67	100	50	93	100	55	95	-	86
CT100	64	89	100	100	96	100	70	92	-	89
CT150	64	89	100	100	93	100	70	94	-	89
CT200	64	89	100	100	93	100	70	91	-	88
CT250	64	89	100	100	91	100	70	90	-	86
Nº de Perfis	14	9	9	4	66	1	19	74	-	196

AG – Acurácia geral, CX – Cambissolo Háplico, MT - Chernossolo Argilúvico, GX – Gleissolo Háplico, GM – Gleissolo Melânico, LV – Latossolo Vermelho, RL – Neossolo Litólico, RR - Neossolo Regolítico e NV – Nitossolo Vermelho.

Na bacia Santo Cristo os valores de AG variaram de 77% a 89%, representado um ganho de 15,6% entre o mapa de menor concordância (CT00) e os de maior concordância (CT100 e CT150). Para essa área todas as classes apresentaram acurácia do mapeador igual ou superior a 64%, sendo que as duas com maior número de perfis (LV e RR) apresentaram concordância superior a 90%.

O aumento no número de pixels coletados nos *buffers* promoveu ganho de acurácia do mapeador para a maioria das classes, entretanto, ocorreu uma perda de aproximadamente 4% nas classes LV e RR (Tabela 10). Essa redução de concordância pode ser atribuída a melhor predição das classes que disponibilizavam de menor número de perfis de solos e que obtiveram aumento no número de pontos amostrais com a utilização do *buffer*, consequentemente, inserção de erro nas classes com maior proporção de perfis.

Esses resultados concordam com os observados na predição de solos com amostragem estratificada, que favorece classes de menor extensão, e acarretam em erros nas classes de maior extensão (TESKE; GIASSON; BAGATINI, 2015b). Na validação com os perfis de solos para ambas às áreas de estudo, os valores

AG obtidos podem ser considerados satisfatórios, uma vez que estão de acordo com outros estudos que utilizaram perfis de solos para predição de ocorrência dos solos no MDS (HÄRING et al., 2012; BAGHERI BODAGHABADI et al., 2015; VASQUES et al., 2015; DIAS et al., 2016).

Na avaliação de reprodutibilidade dos mapas de UM, a acurácia geral apresentou variação de 50% a 59% na bacia Lajeado Grande e 55% a 57% na bacia do Santo Cristo. As unidades LV1 e RR1 predominantes em ambas às áreas apresentaram concordância superior a 70% (Tabela 11), concordando com a proporcionalidade de pontos amostrais disponíveis para as respectivas UM. Na bacia Lajeado Grande, apenas na unidade G ocorreu redução de concordância com o aumento do número de pixels amostrados, sendo que a maior concordância foi obtida no mapa UM50, o aumento no raio do *buffer* pode ter permitido coleta de dados com maior heterogeneidade dificultando a predição desta UM.

Tabela 11 - Concordância de reprodutibilidade dos mapas preditos de unidades de mapeamento com os mapas convencionais de solos das bacias dos rios Lajeado Grande e Santo Cristo.

Mapas	Unidades de mapeamento de solos														AG
	G	LV1	LV2	M1	M2	NV1	NV2	RL	RR1	RR2	RR3	RR4	RR5	RR6	
Bacia hidrográfica do Lajeado Grande															
UM00	47	84	43	26	28	20	6	0	73	4	3	41	0	0	50
UM50	63	80	52	25	37	14	10	0	70	3	11	41	3	0	50
UM100	54	84	49	20	29	24	14	0	65	21	9	39	6	0	53
UM150	46	85	49	31	28	26	22	0	64	28	9	47	18	2	55
UM200	38	86	49	33	33	28	25	0	63	37	11	54	26	7	57
UM250	35	86	47	37	42	31	28	0	62	46	14	53	33	9	59
Área (%)	1,61	27,21	5,84	0,67	0,73	0,45	1,68	0,41	30,35	23,91	5,26	0,86	0,52	0,49	533,3(km^2)
Bacia hidrográfica do Santo Cristo															
UM00	11	78	11	9	-	-	-	0	69	1	6	0	0	-	55
UM50	10	80	16	17	-	-	-	0	63	6	7	0	0	-	55
UM100	15	77	20	41	-	-	-	0	63	7	8	0	0	-	55
UM150	17	76	21	36	-	-	-	0	65	9	12	0	0	-	55
UM200	28	76	19	54	-	-	-	0	65	14	14	19	0	-	56
UM250	30	77	19	60	-	-	-	0	64	18	16	23	0	-	57
Área (%)	2,46	38,13	7,92	0,20	-	-	-	0,01	34,79	8,24	7,84	0,03	0,38	-	900,7(km^2)

AG- Acurácia Geral.

De modo geral, a reprodutibilidade do mapa convencional com os dados coletados apenas nos pixels dos perfis de solos, alcançou valores próximos ao valor médio de 60% observado na literatura em estudos com outras estratégias

de amostragem utilizada no MDS (HÖFIG; GIASSON; VENDRAME, 2014; GIASSON et al., 2015; BAGATINI; GIASSON; TESKE, 2016; DIAS et al., 2016; PELEGRINO et al., 2016). Vale destacar que em ambas às áreas de estudo, apenas duas unidades representam mais de 75% da composição dos respectivos mapas de solos, sendo estas as mais importantes na predição e as quais apresentaram maior concordância com os mapas legados utilizados na validação.

Na avaliação de concordância dos mapas preditos de classes taxonômicas de solo com os mapas de unidades de mapeamento de solo, constatou-se uma ligeira redução nos valores de AG (Tabela 12). Para a bacia do Lajeado Grande, o maior valor de AG (86%) foi obtido no mapa CT00, predito com os dados coletados nos pixels dos perfis de solo.

Tabela 12 - Concordância de reprodutibilidade dos mapas preditos de classes taxonômicas de solos com os mapas convencionais de solos das bacias dos rios Lajeado Grande e Santo Cristo.

Mapas	Unidades de mapeamento de solo														
	G	LV1	LV2	M1	M2	NV1	NV2	RL	RR1	RR2	RR3	RR4	RR5	RR6	AG
Bacia do rio Lajeado Grande															
CT00	32	72	98	43	98	0	73	100	88	97	98	99	94	100	86
CT50	45	72	98	51	98	0	62	100	83	94	96	97	91	96	83
CT100	42	72	99	47	97	0	53	100	80	91	95	96	78	95	81
CT150	37	73	98	47	97	0	55	100	77	89	95	95	75	93	80
CT200	32	73	97	47	96	0	54	100	75	88	95	95	74	92	79
CT250	28	73	97	46	96	0	57	100	75	87	93	94	72	91	79
Área (%)	1,61	27,21	5,84	0,67	0,73	0,45	1,68	0,41	30,35	23,91	5,26	0,86	0,52	0,49	533,3(km^2)
Bacia do rio Santo Cristo															
CT00	40	62	91	54	-	-	-	0	61	81	89	100	45	-	67
CT50	28	62	94	27	-	-	-	57	62	82	92	100	30	-	68
CT100	27	64	94	29	-	-	-	60	59	78	92	100	28	-	67
CT150	29	60	91	25	-	-	-	59	59	78	89	100	16	-	65
CT200	31	59	90	28	-	-	-	100	56	74	88	90	15	-	63
CT250	31	57	90	31	-	-	-	100	56	71	87	93	15	-	62
Área (%)	2,46	38,13	7,92	0,20	-	-	-	0,01	34,79	8,24	7,84	0,03	0,38	-	900,7(km^2)

AG - Acurácia Geral. Área – áreas em porcentagem correspondente a cada unidade de mapeamento no mapa convencional de solos, (-) Unidade de mapeamento ausente no mapa convencional.

Na bacia Santo Cristo, o maior valor de AG (68%) foi obtido no mapa CT50, sendo que nestes mapas também são observados os maiores valores de acurácia do mapeador nas UMs de maiores extensões de área (Tabela 12). Com exceção das unidades LV1 e RR1, que são unidades simples, os valores

observados nas outras UMs de maiores extensões são explicados por estas serem compostas por duas ou mais classes taxonômicas de solos (CT), para as quais são considerados corretos os acertos em qualquer uma das CTs da composição.

As unidades de mapeamento LV1, LV2, RR1 e RR2 são compostas por Latossolos Vermelhos (LV) e Neossolos Regolíticos (RR) ou associação destes, e ocorrem predominantemente em ambas às áreas de estudo, classes estas que também possuem o maior número de perfis de solos, consequentemente, maior concordância de acertos em todas as avaliações realizadas. As classes taxonômicas LV e RR representam mais de 85% da composição dos mapas preditos (Figura 20), concordando com os mapas convencionais de solos, uma vez que estas CTs estão presentes na composição da maioria das UMs.

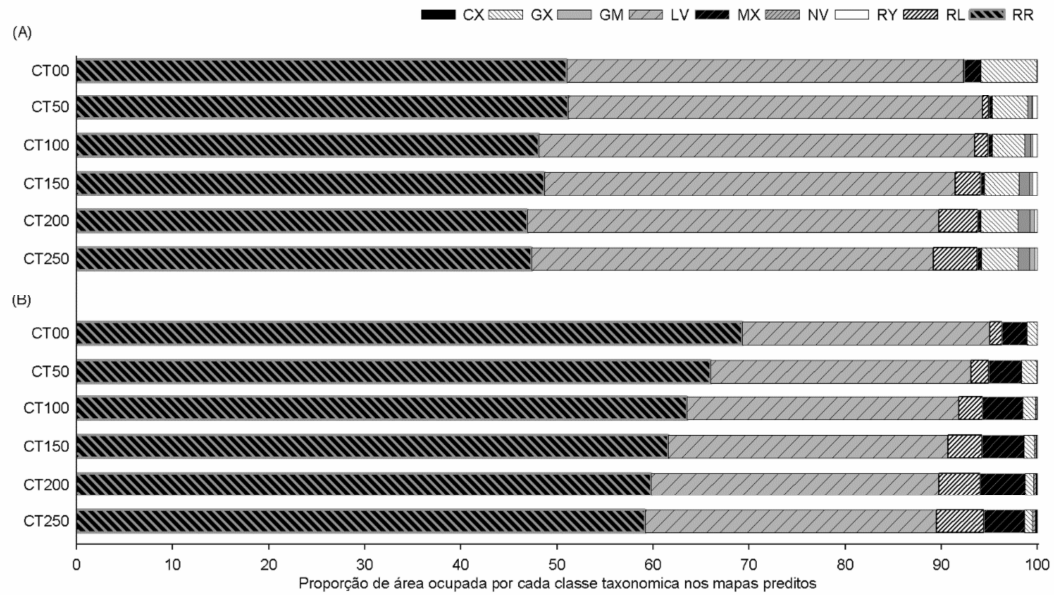


Figura 20 - Proporção de área de cada classe taxonômica nos mapas preditos das bacias dos rios Santo Cristo (A) e Lajeado Grande (B). CX – Cambissolo Háplico, MX - Chernossolo Háplico, GX – Gleissolo Háplico, GM – Gleissolo Melânico, LV – Latossolo Vermelho, RL – Neossolo Litólico, RR - Neossolo Regolítico e NV – Nitossolo Vermelho.

De modo geral, o aumento no número de pixels amostrados promovido pelo *buffer* levou à redução na proporção das CTs predominantes e a um ligeiro aumento na predição das classes com menor proporção, indicando redução na superestimação das classes de maior extensão, como pode ser observado nos tratamentos CT00 e CT250 (Figura 20).

Quando foram utilizados os pixels coletados nos *buffers*, houve redução na superestimação das UMs de maiores extensões, com a composição do mapa UM250 ficando mais próximo do mapa convencional, sendo que esse efeito foi mais expressivo na bacia do Lajeado Grande e, principalmente, nas unidades LV2 e RR2 (Figura 21).

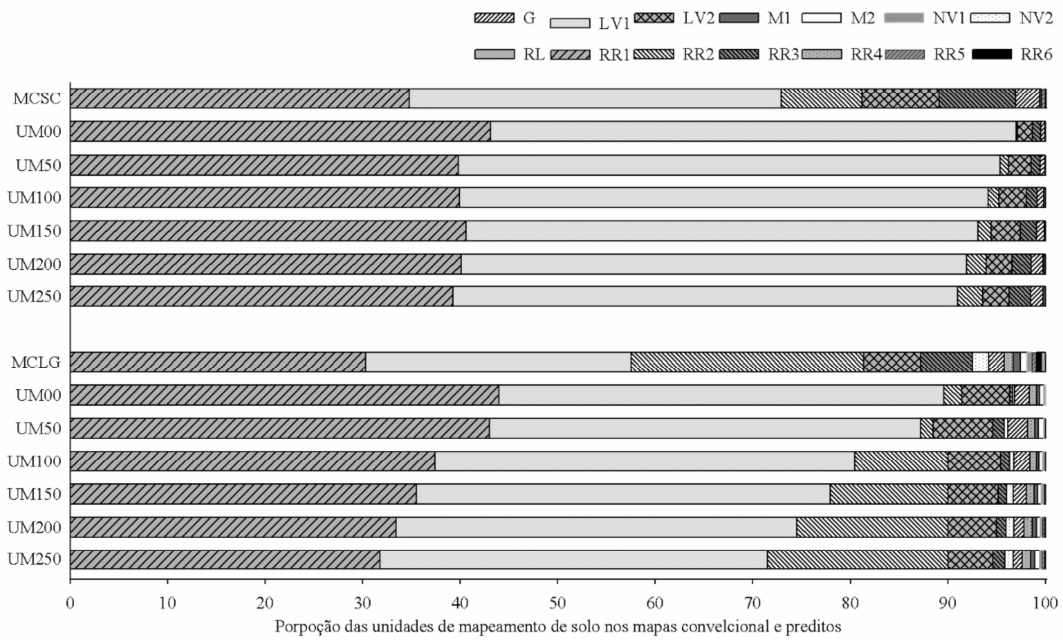


Figura 21 - Proporção de área de cada unidade de mapeamento de solo nos mapas preditos das bacias dos rios Santo Cristo (A) e Lajeado Grande (B). MCSC – mapa convencional da bacia Santo Cristo, MCLG - mapa convencional da bacia Lajeado Grande

As alterações nos mapas preditos, promovidas pela utilização de pixels amostrais coletados nos *buffers* também podem ser constatadas através da análise visual dos mapas preditos. As principais alterações ocorreram entre os mapas CT00, CT150 e CT250 para as bacias Santo Cristo (Figuras 22B, 22C e 22D) e Lajeado Grande (Figuras 22G, 22H e 22I), nestes mapas é possível visualizar a evolução na individualização de áreas correspondentes as classes de menor ocorrência, como a RL, GX e CX.

As classes Latossolos e Neossolos ocupam posição bem definidas na paisagem, com os Latossolos ocorrendo principalmente nas posições superiores e planas e os Neossolos ocorrendo nos vales encaixados, características estas que permitem um aprendizado mais eficiente do classificador para as respectivas CTs, consequentemente, resultando nos maiores valores de concordância constatados.

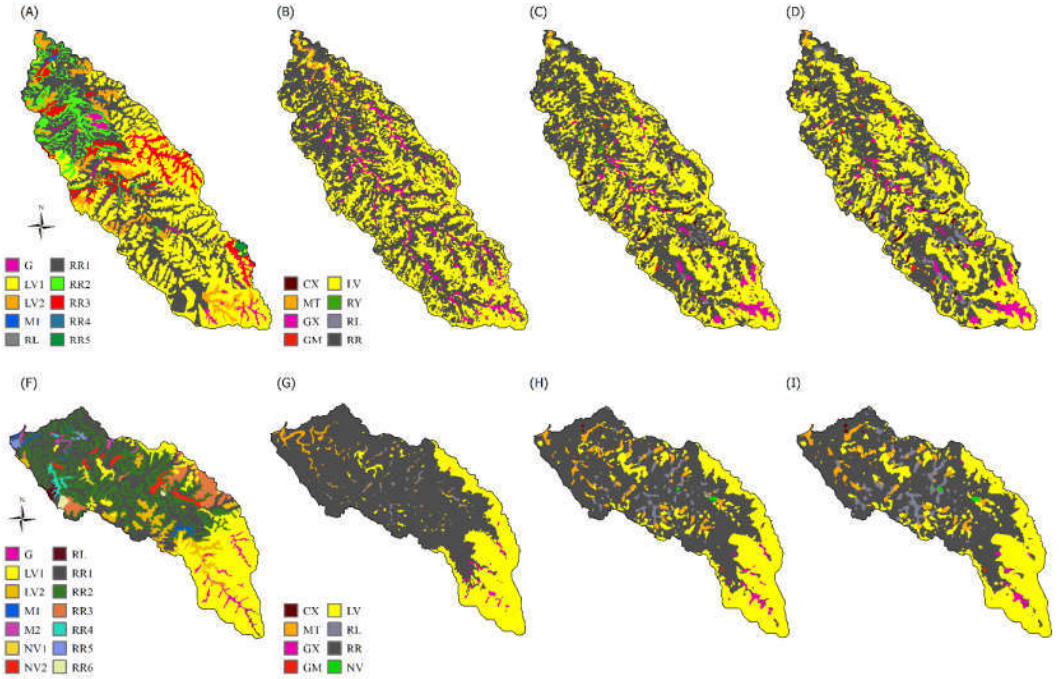


Figura 22 - Mapas convencional de unidades de mapeamento solos das bacias dos rios Santo Cristo (A) e Lajeado Grande (F) e mapas de classes taxonômicas preditos com os pontos coletados nos *buffers* de 50, 150 e 250m das bacias Santo Cristo (B, C e D) e Lajeado Grande (G, H e I). CX – Cambissolo Háplico, MT - Chernossolo Argilúvico, GX – Gleissolo Háplico, GM – Gleissolo Melânico, LV – Latossolo Vermelho, RL – Neossolo Litólico, RR - Neossolo Regolítico e NV – Nitossolo Vermelho.

Nos mapas preditos de unidades de mapeamento de solo também é possível constatar as alterações promovidas pela amostragem com os pixels coletados nos diferentes raios do *buffer*, com essa diferença sendo mais expressiva entre os mapas UM00, UM150 e UM250 das bacias Santo Cristo (Figuras 23B, 23C e 23D) e Lajeado grande (Figuras 23G, 23H e 23I).

Na bacia do Santo Cristo, as principais diferenças ocorrem com a maior individualização de áreas das unidades G, M1, RR2 e RR3, já na bacia do Lajeado Grande as maiores diferenças ocorrem nas áreas de LV1, RR2, RR3 e RR4. De modo geral, a utilização dos pontos coletados nos *buffers* permitiu ganho de predição nas UMs de menor extensão e bom desempenho na predição das unidades de maior extensão.

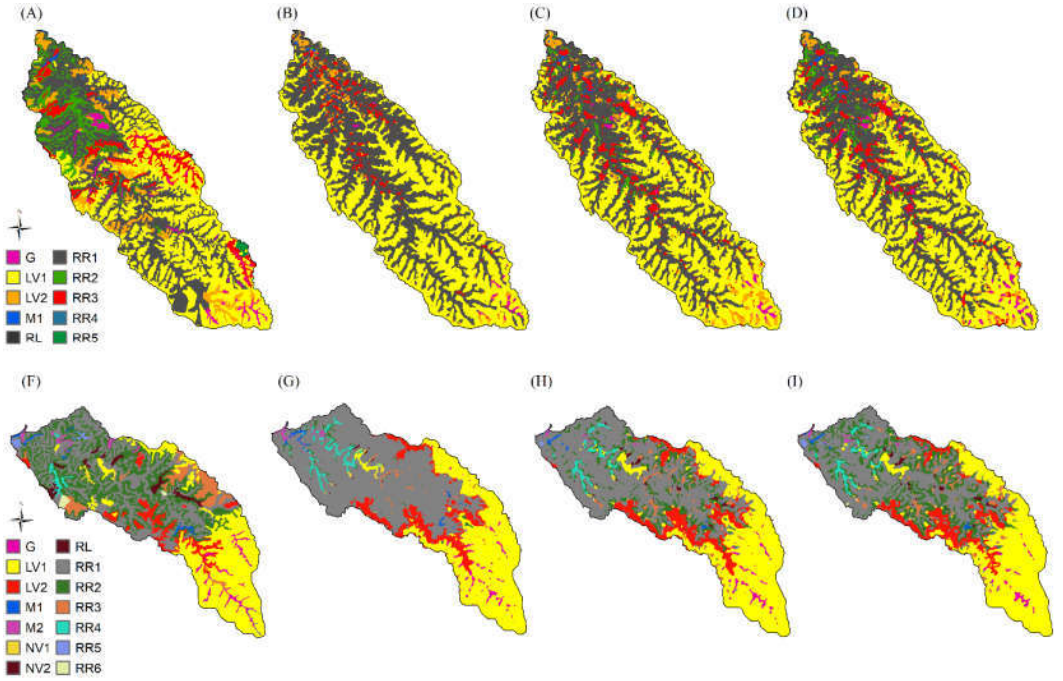


Figura 23 - Mapas convencional de unidades de mapeamento solos das bacias dos rios Santo Cristo (A) e Lajeado Grande (F) e mapas de unidades de mapeamento preditos com os pontos coletados nos *buffers* de 50, 150 e 250m das bacias Santo Cristo (B, C e D) e Lajeado Grande (G, H e I).

Para ambas às áreas de estudos, as classes predominantes foram preditas de forma satisfatória em todos os mapas. Estes resultados indicam que mesmo com a utilização apenas dos pixels dos perfis de solos associados às variáveis preditoras e ao classificador utilizado (*RanomForest*) foram eficientes na predição da ocorrência dos solos.

Os resultados obtidos indicam que a utilização do *buffer* associado a locais com disponibilidades de perfis georreferenciados permite a predição de ocorrência dos solos, sem a necessidade de delineamento manual dos polígonos de solos como exigido pelo mapeamento tradicional (DEMATTÊ; RIZZO; BOTTEON, 2015; TESKE; GIASSON; BAGATINI, 2015b). Assim, a aplicação do MDS em áreas com disponibilidade de perfis de solos georreferenciados poderá levar a economia de tempo em relação ao mapeamento tradicional, com boa qualidade nos mapas gerados, sendo que os mesmos podem auxiliar os pedólogos na produção de mapas mais acurados.

5.4. CONCLUSÕES

- A utilização dos pixels amostrais coletados nos *buffers* não alterou de forma expressiva a acurácia geral dos mapas preditos na bacia do rio Lajeado Grande, mas permitiu um ganho de 15,6% de concordância na bacia do rio Santo Cristo na validação com os perfis de solos.
- Na avaliação de reprodutibilidade dos mapas convencionais de unidades de mapeamento de solos, os maiores valores de concordância foram obtidos nos mapas preditos com os dados coletados no *buffer* de 250 metros.
- Nos mapas preditos de classes taxonômicas e nos mapas preditos de unidades de mapeamento de solos, as classes Latossolos Vermelhos e Neossolos Regolíticos e as unidades LV1 e RR2 foram preditas como predominantes, concordando com as proporções observadas nos perfis de solos e nos mapas convencionais disponíveis para ambas às áreas de estudo.

6. CAPÍTULO V – CONSIDERAÇÕES GERAIS

Com base nos resultados óbitos constata-se que a seleção de variáveis eficientes na discriminação das classes de solos, assim com a identificação do nível de detalhe mais adequado destas variáveis para representar a ocorrência dos solos, permite melhorar o desempenho dos modelos preditores e a acurácia dos mapas preditos. Nesse contexto, destaca-se a seleção *wrapper* e o filtro de média como técnicas que podem ser aplicadas para aumentar a eficiência das metodologias utilizadas no Mapeamento Digital de Solos.

Os resultados obtidos no terceiro estudo permitem concluir que o uso do *buffer* associado a perfis de solos georreferenciados permite a predição de todas as classes taxonômicas de solos presentes nos dados de treinamento, sendo que o aumento nas áreas dos *buffers* melhora a predição das classes que apresentam menor proporção nos dados de treinamento.

Assim, com base nos resultados obtidos nos três estudos concluímos que as metodologias testadas aumentam o desempenho das técnicas utilizadas na predição de ocorrência dos solos e apresentam potencial para uso no Mapeamento Digital de Solos em áreas com disponibilidade de dados de solos na forma de mapas legados ou perfis de solos georreferenciados, entretanto, recomenda-se mais estudos com as presentes metodologias em outras áreas, buscando confirmar sua eficiência, bem como avaliar a capacidade de estrapolação dos modelos preditores para áreas fisiograficamente semelhantes a áreas que já possuem dados legados.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ABDEL-KADER, F. H. Digital soil mapping at pilot sites in the northwest coast of Egypt: A multinomial logistic regression approach. **Egyptian Journal of Remote Sensing and Space Science**, [S.l], v. 14, n. 1, p. 29–40, 2011.

ADHIKARI, K. et al. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. **Geoderma**, Amsterdam, v. 214–215, n. 2014, p. 101–113, 2014.

AFSHAR, F. A.; AYOUBI, S.; JAFARI, A. The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. **Geoderma**, Amsterdam, v. 315, p. 36–48, 2018.

AGATONOVIC-KUSTRIN, S.; BERESFORD, R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. **Journal of Pharmaceutical and Biomedical Analysis**, Amsterdam, v. 22, n. 5, p. 717–727, 2000.

AITKENHEAD, M. J.; COULL, M. C. Mapping soil carbon stocks across Scotland using a neural network model. **Geoderma**, Amsterdam, v. 262, p. 187–198, 2016.

ALVES, M. R.; DEMATTÊ, J. A. M.; BARROS, P. P. S. Multiple geotechnological tools applied to digital mapping of tropical soils. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 39, n. 5, p. 1261–1274, 2015.

ARRUDA, G. P. et al. Mapeamento Digital de solos por redes neurais artificias com base na relação solo-paisagem. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 37, n. 1, p. 327–338, 2013.

ARRUDA, G. P. et al. Digital soil mapping using reference area and artificial neural networks. **Scientia Agricola**, Piracicaba, v. 73, n. 3, p. 266–273, 2016.

BAGATINI, T.; GIASSON, E.; TESKE, R. Seleção de densidade de amostragem com base em dados de áreas já mapeadas para treinamento de modelos de árvore de decisão no mapeamento digital de solos. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 39, n. 4, p. 960–967, 2015.

BAGATINI, T.; GIASSON, E.; TESKE, R. Expansão de mapas pedológicos para áreas fisiograficamente semelhantes por meio de mapeamento digital de solos. **Pesquisa Agropecuária Brasileira**, Brasília, v. 51, n. 9, p. 1317–1325, 2016.

BAGHERI BODAGHABADI, M. et al. Digital soil mapping using artificial neural networks and terrain-related attributes. **Pedosphere**, Beijing, v. 25, n. 4, p. 580–591, 2015.

BATES, S.C., METCALFE, R.A. **Identifying depression-focussed groundwater recharge areas within hummocky terrain landscapes using automated digital landform classification techniques**. Peterborough, Ontario: Watershed Science Centre, Trent University, 2006. (WSC Report**,** n.02-2006)

BEHRENS, T. et al. Digital soil mapping using artificial neural networks. **Journal of Plant Nutrition and Soil Science**, Weinheim, v. 168, n. 1, p. 21–33, 2005.

BEHRENS, T. et al. Multi-scale digital terrain analysis and feature selection for digital soil mapping. **Geoderma**, Amsterdam, v. 155, n. 3–4, p. 175–185, 2010.

BEHRENS, T. et al. Hyper-scale digital soil mapping and soil formation analysis. **Geoderma**, Amsterdam, v. 213, p. 578–588, 2014.

BEHRENS, T. et al. Multiscale contextual spatial modelling with the Gaussian scale space. **Geoderma**, Amsterdam, v. 310, p. 128–137, 2018.

BISWAS, A.; ZHANG, Y. Sampling designs for validating digital soil maps: A Review. **Pedosphere**, Beijing, v. 28, n. 1, p. 1–15, 2018.

BOCK M, KÖTHE R. Predicting the depth of hydromorphic soil characteristics influenced by ground water. In: BÖHNER, J., BLASCHKE, T., MONTANARELLA, L. (Eds.). **SAGA-seconds Out**. v 2. Hamburg: Universität Hamburg, Institut für

Geographie. 2008. p. 113

BREIMAN, L. Bagging Predictors. **Machine Learning**, Dordrecht, v. 24, n. 2, p. 123–140, 1996.

BREIMAN, L. et al. **Breiman and Cutler's Random Forests for Classification and Regression**. [S.l.]: CRAN, 2018.

BRENNING, A.; BANGS, D.; BECKER, M. **RSAGA: SAGA Geoprocessing and Terrain Analysis**. [S.l.]: CRAN, 2018.

BRUNGARD, C. W. et al. Machine learning for predicting soil classes in three semi-arid landscapes. **Geoderma**, Amsterdam, v. 239, p. 68–83, 2015.

CALDERANO FILHO, B. et al. Artificial neural networks applied for soil class prediction in mountainous landscape of the Serra do Mar. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 38, n. 6, p. 1681–1693, 2014.

CARDOSO, C. A. et al. Caracterização morfométrica da bacia hidrográfica do rio Debossan, Nova Friburgo, RJ. **Revista Árvore**, Viçosa, v. 30, n. 2, p. 241–248, 2006.

CARVALHO, C. C. N.; NUNES, F. C.; ANTUNES, A. M. H. Histórico do levantamento de solos no Brasil: Da industrialização brasileira à era da informação. **Revista Brasileira de Cartográfia**, Rio de Janeiro, v. 65, p. 997–1013, 2013.

CARVALHO JUNIOR, W. et al. Evaluation of statistical and geostatistical models of digital soil properties mapping in tropical mountain regions. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 38, n. 1, p. 706–717, 2014.

CARVALHO JÚNIOR, W. et al. Método do hipercubo latino condicionado para a amostragem de solos na presença de covariáveis ambientais visando o mapeamento digital de solos. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 38, n. 2, p. 386–396, 2014.

CHAGAS, C. DA S. et al. Atributos topográficos e dados do Landsat7 no mapeamento digital de solos com uso de redes neurais. **Pesquisa Agropecuária Brasileira**, Brasília v. 45, n. 5, p. 497–507, 2010.

CHAGAS, C. DA S. et al. Data mining methods applied to map soil units on tropical hillslopes in Rio de Janeiro, Brazil. **Geoderma Regional**, Amsterdam, v. 9, p. 47–55, 2017.

CHAGAS, C. DA S.; OLIVEIRA, C. A.; FERNANDES, E. I. Comparison Between Artificial Neural Networks and Maximum Likelihood Classification in Digital Soil Mapping. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 37, n. 2, p. 339–351, 2013.

CHAGAS, C. S.; CARVALHO JÚNIOR, W.; BHERING, S. B. Integração de dados do quickbird e atributos do terreno no mapeamento digital de solos por redes neurais artificiais. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 35, n. 3, p. 693–704, 2011.

CHAWLA, N. V et al. SMOTE: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, San Francisco, v. 16, p. 321–357, 2002.

COELHO, F. F.; GIASSON, E. Comparação de métodos para mapeamento digital de solos com utilização de sistema de informação geográfica. **Ciência Rural**, Santa Maria, v. 40, n. 10, p. 2099–2106, 2010.

CONGALTON, R. G. A review of assessing the accuracy of classifications of remotely sensed data. **Remote Sensing of Environment**, New York, v. 37, n. 1, p. 35–46, 1991.

DALMOLIN, R. S. D. et al. Relação entre as características e o uso das informações de levantamentos de solos de diferentes escalas. **Ciência Rural**, Santa Maria, v. 34, n. 5, p. 1479–1486, 2004.

DALMOLIN, R. S. D.; TEN CATEN, A. Mapeamento Digital : nova abordagem em levantamento de solos. **Investigación Agraria**, San Lorenzo, v. 17, n. 2, p. 77–86, 2015.

DANCEY, C. P.; REIDY, J. **Estatística sem matemática para psicologia**. 3. ed. Porto Alegre: Artmed, 2006.

DASH, M.; LIU, H.; MOTODA, H. Consistency Based Feature Selection. In:

PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2000, Kyoto. **Anais**... Kyoto: PAKDD, 2000.

DEMATTÊ, J. A. M.; DEMÉTRIO, V. A. Caracterização de solos por padrões de drenagem e sua relação com índices de intemperismo. **Pesquisa Agropecuária Brasileira**, Brasília, v. 33, n. 1, p. 87–95, 1998.

DEMATTÊ, J. A. M.; RIZZO, R.; BOTTEON, V. W. Pedological mapping through integration of digital terrain models spectral sensing and photopedology. **Revista Ciência Agronomica**, Fortaleza, v. 46, n. 4, p. 669–678, 2015.

DIAS, L. M. DA S. et al. Predição de classes de solo por mineração de dados em área da bacia sedimentar do São Francisco. **Pesquisa Agropecuária Brasileira**, Brasília, v. 51, n. 9, p. 1396–1404, 2016.

DOBOS, E. et al. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. **Geoderma**, Amsterdam, v. 97, n. 3–4, p. 367–391, 2000.

DORNIK, A.; DRÂGUŢ, L.; URDEA, P. Classification of soil types using geographic object-based image analysis and Random Forest. **Pedosphere**, Beijing, v. 0160, p. 1-21, 2017.

EMBRAPA. **Sistema brasileiro de classificação de solos**. 2. ed. Rio de Janeiro: Embrapa Solos, 2006.

EMBRAPA. **Programa Nacional de Solos do Brasil (PronaSolos)**. Rio de Janeiro: Embrapa Solos, 2016.

FIGUEIREDO, S. R. et al. Uso de regressões logísticas múltiplas para mapeamento digital de solos no Planalto Médio do RS. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 32, n. esp., p. 2779–2785, 2008.

FLACH, W.; CORR, E. A. Levantamento de solos no brasil: métodos, práticas e dificuldades. **Geographia Meridionalis**, Pelotas, v.3, n.3, p. 420–431, 2017.

FRANCO, Â. M. P. et al. Delineamento das unidades de mapeamento de solos utilizando o google Earth. **Geociências**, São Paulo, v. 34, n. 4, p. 861–871, 2015.

FREITAS, M. et al. Avaliação do potencial hidrogeológico, vulnerabilidade intrínseca e hidroquímica do Sistema Aquifero Serra Geral no Noroeste do Estado do Rio Grande do Sul. **Revista Brasileira de Recursos Hídricos**, Porto Alegre, v. 17, n. 2, p. 31–41, 2012.

GEUS. **Danmarks Kvartære Jordartskort (Denmark's Quaternary Soil Maps), 1:25,000**, 1998. 1 CD-ROM

GIASSON, E. et al. Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. **Scientia Agricola**, Piracicaba, v. 63, n. 3, p. 262–268, 2006.

GIASSON, E. et al. Decision trees for digital soil mapping on subtropical basaltic steeplands. **Scientia Agricola**, Piracicaba, v. 68, n. 2, p. 167–174, 2011.

GIASSON, E. et al. 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. **Ciência Rural**, Santa Maria, v. 43, n. 11, p. 1967–1973, 2013.

GIASSON, E. et al. Instance selection in digital soil mapping: a study case in Rio Grande do Sul, Brazil. **Ciência Rural**, Santa Maria, v. 45, n. 9, p. 1592–1598, 2015.

GRINAND, C. et al. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. **Geoderma**, Amsterdam, v. 143, n. 1–2, p. 180–190, 2008.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research**, Cambridge, v. 3, n. 3, p. 1157–1182, 2003.

HALL, M. A. et al. The WEKA data mining software: an update. **SIGKDD Explorations**, New York, v. 11, n. 1, p. 10–18, 2009.

HALL, M. A.; HOLMES, G. Benchmarking attribute selection techniques for discrete class data mining. **IEEE Transactions on Knowledge and Data Engineering**, New York, v. 15, n. 6, p. 1437–1447, 2003.

HALL, M.; SMITH, L. A. Feature Selection for Machine Learning : Comparing a Correlation-based Filter Approach to the Wrapper CFS : Correlation-based Feature. In: PROCEEDINGS OF THE TWELFTH INTERNATIONAL FLORIDA ARTIFICIAL INTELLIGENCE RESEARCH SOCIETY CONFERENCE, 1999, Orlando. **Anais**...Orlando, EUA: AAAI Press, 1999.

HÄRING, T. et al. Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. **Geoderma**, Amsterdam, v. 185–186, p. 37–47, 2012.

HARTEMINK, A. E.; MCBRATNEY, A. A soil science renaissance. **Geoderma**, Amsterdam, v. 148, n. 2, p. 123–129, 2008.

HASENACK, H.; WEBER, E. **Base cartográfica vetorial contínua do Rio Grande do Sul - escala 1:50.000**. Porto Alegre: UFRGS Centro de Ecologia, 2010.

HENGL, T. et al. Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. **Geoderma**, Amsterdam, v. 140, n. 4, p. 417–427, 2007.

HEUNG, B.; HODÚL, M.; SCHMIDT, M. G. Comparing the use of training data derived from legacy soil pits and soil survey polygons for mapping soil classes. **Geoderma**, Amsterdam, v. 290, p. 51–68, 2017.

HÖFIG, P.; GIASSON, E.; VENDRAME, P. R. S. Mapeamento digital de solos com base na extrapolação de mapas entre áreas fisiograficamente semelhantes. **Pesquisa Agropecuária Brasileira**, Brasília, v. 49, n. 12, p. 958–966, 2014.

HORNIK, K. et al. **R/Weka Interface Description.** [S.l.]: CRAN, 2016.

HORTON, R. E. Erosional development of streams and their drainage basins, hydrophyrical approach to quantitative morphology. **Journal of the Japanese Forestry Society**, Tokio, v. 37, n. 6, p. 257–262, 1955.

KÄMPF, N.; GIASSON, E.; STRECK, E. V. **Levantamento pedológico e análise qualitativa do potencial de uso dos solos para o descarte de dejetos suínos da microbacia do Lajeado Grande [Relatório]**. Porto Alegre: Secretaria do

Meio Ambiente do Rio Grande do Sul, 2004a.

KÄMPF, N.; GIASSON, E.; STRECK, E. V. **Levantamento semidetalhado dos solos da microbacia do rio Santo Cristo [Relatório]**. Porto Alegre: Secretaria do Meio Ambiente do Rio Grande do Sul, 2004b.

KERRY, R.; OLIVER, M. A. Soil geomorphology: Identifying relations between the scale of spatial variation and soil processes using the variogram. **Geomorphology**, Amsterdam, v. 130, n. 1–2, p. 40–54, 2011.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial Intelligence**, Amsterdam, v. 97, n. 2, p. 273–324, 1997.

LAGACHERIE, P.; MCBRATNEY, A. B.; VOLTZ, M. Digital Soil Mapping. In: **Digital Soil Mapping: An Introductory Perspective**. Amisterdã: Elsevier, 2006.

LAL, T. N. et al. Embedded Methods. In: GUYON, I. et al. (Eds.). **Studies in Fuzziness and Soft Computing**. Heidelberg: Springer, 2006.

LIU, H.; SETIONO, R. **A probabilistic approach to feature selection - a filter solution**. International Conference on Machine Learning, 13., 1996, Bari.. **Proceedings**...Bari: Morgan Kaufmann Publishers Inc, 1996.

MACMILLAN, R. A. **LandMapper LandMapR Software Toolkit- C ++ Version (2003) Users ManualLandMapper Environmental Solutions Inc.** Edmonton: LandMapper Environmental Solutions Inc., 2003.

MASSAWE, B. H. J. et al. Mapping numerically classified soil taxa in Kilombero Valley, Tanzania using machine learning. **Geoderma**, Amsterdam, v. 311, p. 143–148, 2018.

MCBRATNEY, A. B. et al. On digital soil mapping. **Geoderma**, Amsterdam, v. 117, n. 1–2, p. 3–52, 2003.

MENDONÇA-SANTOS, M. DE L.; SANTOS, H. G. **Mapeamento Digital de Classes e Atributos de Solos métodos, paradigmas e novas técnicas**. Rio de Janeiro: Embrapa Solos 2003. (Documento 55)

MILLER, B. A. et al. Impact of multi-scale predictor selection for modeling soil

properties. **Geoderma**, Amsterdam, v. 239, p. 97–106, 2015.

MINASNY, B.; MCBRATNEY, A. B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. **Computers and Geosciences**, Oxford, v. 32, n. 9, p. 1378–1388, 2006.

MINASNY, B.; MCBRATNEY, A. B. Digital soil mapping: A brief history and some lessons. **Geoderma**, Amsterdam, v. 264, n. Aug., p. 301–311, 2016.

MIRAKZEHI, K. et al. Digital soil mapping of deltaic soils: A case of study from Hirmand (Helmand) river delta. **Geoderma**, Amsterdam, v. 313, n. Oct. 2017, p. 233–240, 2018.

ODGERS, N. P.; MCBRATNEY, A. B.; MINASNY, B. Bottom-up digital soil mapping. II. Soil series classes. **Geoderma**, Amsterdam, v. 163, n. 1–2, p. 30–37, 2011.

OTTO J. C. et al.GIS Applications in Geomorphology. In: COVA T. J. et al. (Eds). **Comprehensive Geographic Information Systems**. Amisterdã: Elsevier Inc., 2017.

PAES, B. C.; PLASTINO, A.; FREITAS, A. A. Seleção de Atributos Aplicada à Classificação Hierárquica. In: SYMPOSIUM ON KNOWLEDGE DISCOVERY, MINING AND LEARNING, 2013, São Carlos. **Anais**...São Carlos: KDMile'13, 2013.

PAHLAVAN-RAD, M. R. et al. Legacy soil maps as a covariate in digital soil mapping: A case study from Northern Iran. **Geoderma**, Amsterdam, v. 279, p. 141–148, 2016.

PAHLAVAN RAD, M. R. et al. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. **Geoderma**, Amsterdam, v. 232–234, p. 97–106, 2014.

PAULA, S. H. M. et al. Unidades pedoambientais da região de santa tereza, estado do tocantins. **Pesquisa Agropecuária Tropical**, Goiânia, v. 40, n. 1, p. 8–19, 2010.

PELEGRINO, M. H. P. et al. Mapping soils in two watersheds using legacy data

and extrapolation for similar surrounding areas. **Ciência e Agrotecnologia**, Lavras, v. 40, n. 5, p. 534–546, 2016.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Francisco: Morgan Kaufmann Publishers Inc., 1993.

R CORE TEAM. **R: A Language and Environment for Statistical Computing.** Vienna, Austria, 2018. Disponível em: <https://www.r-project.org/>. Acesso em: nov. 2018.

RUIZ, L. F. C.; TEN CATEN, A.; DALMOLIN, R. S. D. Árvore de decisão e a densidade mínima de amostras no mapeamento da cobertura da Terra. **Ciência Rural**, Santa Maria, v. 44, n. 6, p. 1001–1007, 2014.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. **Plos One**, San Francisco, v. 10, n. 3, p. 1–21, 2015.

SANTOS, H. G. et al. Distribuição Espacial dos Níveis de Levantamento de Solos no Brasil. In: CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO, 34., 2013, Florianópolis. **Anais**... Florianópolis: Sociedade Brasileira de Ciência do Solo, 2013.

SCHAETZL, R. J.; ANDERSON, S. **Soils: Genesis and Geomorphology**. New York: Cambridge University Press, 2005.

SILVA, C. C. et al. Mapeamento pedológico digital da folha botucatu (SF-22-Z-B-VI-3): Treinamento de dados em mapa tradicional e validação de campo. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 37, n. 4, p. 846–857, 2013.

SILVEIRA, C. T. et al. Pedometria apoiada em atributos topográficos com operações de tabulação cruzada por álgebra de mapas. **Revista Brasileira de Geomorfologia**, Uberlândia, v. 13, n. 2, p. 125–137, 2012.

SUBBURAYALU, S. K.; JENHANI, I.; SLATER, B. K. Disaggregation of component soil series on an Ohio County soil survey map using possibilistic decision trees. **Geoderma**, Amsterdam, v. 213, p. 334–345, 2014.

SUBBURAYALU, S. K.; SLATER, B. K. Soil Series Mapping By Knowledge

Discovery from an Ohio County Soil Map. **Soil Science Society of America Journal**, Madison, v. 77, n. 4, p. 1254, 2013.

TACHIKAWA, T. et al. **ASTER Global Digital Elevation Model Version 2 - summary of validation results**. 2011. Disponível em: <http://pubs.er.usgs.gov/publication/70005960>. Acesso em: ago. 2018.

TAGHIZADEH-MEHRJARDI, R. et al. Comparing data mining classifiers to predict spatial distribution of USDA-family soil groups in Baneh region, Iran. **Geoderma**, Amsterdam, v. 253–254, p. 67–77, 2015.

TAGHIZADEH-MEHRJARDI, R. et al. Predicting and mapping of soil particle-size fractions with adaptive neuro-fuzzy inference and ant colony optimization in central Iran. **European Journal of Soil Science**, Oxford, v. 67, n. 6, p. 707–725, 2016.

TEN CATEN, A. et al. Regressões logósticas múltiplas: Fatores que influenciam sua aplicação na predição de classes de solos. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 35, n. 1, p. 53–62, 2011a.

TEN CATEN, A. et al. Componentes principais como preditores no mapeamento digital de classes de solos. **Ciência Rural**, Santa Maria, v. 41, n. 7, p. 1170–1176, 2011b.

TEN CATEN, A. et al. Estatística multivariada aplicada à diminuição do número de preditores no mapeamento digital do solo. **Pesquisa Agropecuária Brasileira**, Brasília, v. 46, n. 5, p. 553–561, 2011c.

TEN CATEN, A. et al. Mapeamento digital de classes de solos: características da abordagem brasileira. **Ciência Rural**, Santa Maria, v. 42, n. 11, p. 1989–1997, 2012.

TEN CATEN, A. et al. An appropriate data set size for digital soil. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 37, n. 1, p. 359–366, 2013.

TENG, H. T. et al. Updating a national soil classification with spectroscopic predictions and digital soil mapping. **Catena**, Cremlingen, v. 164, p. 125–134, 2018.

TERRA, F. S.; DEMATTÊ, J. A. M.; ROSSEL, R. A. V. Proximal spectral sensing in pedological assessments: vis–NIR spectra for soil classification based on weathering and pedogenesis. **Geoderma**, Amsterdam, v. 318, p. 123–136, 2018.

TESKE, R.; GIASSON, E.; BAGATINI, T. Comparação de esquemas de amostragem para treinamento de modelos preditores no mapeamento digital de classes de solos. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 39, n. 1, p. 14–20, 2015a.

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. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 39, n. 4, p. 950–959, 2015b.

VASQUES, G. M. et al. Integrating geospatial and multi-depth laboratory spectral data for mapping soil classes in a geologically complex area in southeastern Brazil. **European Journal of Soil Science**, Oxford, v. 66, n. 4, p. 767–779, 2015.

VASU, N. N.; LEE, S. R. A hybrid feature selection algorithm integrating an extreme learning machine for landslide susceptibility modeling of Mt. Woomyeon, South Korea. **Geomorphology**, Amsterdam, v. 263, p. 50–70, 2016.

VAYSSE, K.; LAGACHERIE, P. Using quantile regression forest to estimate uncertainty of digital soil mapping products. **Geoderma**, Amsterdam, v. 291, p. 55–64, 2017.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. 3. ed. Burlington: Morgan Kaufmann Publishers, 2011.

WOLSKI, M. S. et al. Digital soil mapping and its implications in the extrapolation of soil-landscape relationships in detailed scale. **Pesquisa Agropecuária Brasileira**, Brasília, v. 52, n. 8, p. 633–642, 2017.

YILDIRIM, P. Filter based feature selection methods for prediction of risks in hepatitis disease. **International Journal of Machine Learning and Computing**, Singapore, v. 5, n. 4, p. 258–263, 2015.