UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS RURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DO SOLO

ESTRATÉGIAS PARA PREDIÇÃO DE CLASSES DE SOLO

TESE DE DOUTORADO

Luciano Campos Cancian

Santa Maria, RS, Brasil

2019

ESTRATÉGIAS PARA PREDIÇÃO DE CLASSES DE SOLO

Luciano Campos Cancian

Tese apresentada ao Programa de
Pós-Graduação em Ciência do Solo, da
Universidade Federal de Santa Maria (UFSM, RS),
como requisito parcial para a obtenção do grau de
Doutor em Ciência do Solo.

Orientador: Prof. Dr. Ricardo Simão Diniz Dalmolin

Santa Maria, RS, Brasil

2019

Cancian, Luciano Campos
Estratégias para a predição de classes de solo / Luciano Campos Cancian.- 2019.
100 p.; 30 cm

Orientador: Ricardo Simão Diniz Dalmolin
Coorientador: Alexandre ten Caten
Tese (doutorado) - Universidade Federal de Santa Maria, Centro de Ciências Rurais, Programa de Pós Graduação em Ciência do Solo, RS, 2019

1. mapeamento digital de solos 2. análise bibliométrica 3. dados legados 4. incerteza 5. pedometria I. Dalmolin, Ricardo Simão Diniz II. ten Caten, Alexandre III. Título.

Sistema de geração automática de ficha catalográfica da UFSM. Dados fornecidos pelo autor(a). Sob supervisão da Direção da Divisão de Processos Técnicos da Biblioteca Central. Bibliotecária responsável Paula Schoenfeldt Patta CRB 10/1728.

Luciano Campos Cancian

ESTRATÉGIAS PARA PREDIÇÃO DE CLASSES DE SOLO

Tese apresentada ao Curso de Doutorado do Programa de Pós-Graduação em Ciência do Solo, área de concentração em Processos Físicos e Morfogenéticos do Solo, da Universidade Federal de Santa Maria (UFSM), como requisito básico para a obtenção do grau de **Doutor em Ciência do Solo.**

Aprovado em 25 de fevereiro de 2019:

Ricardo Simão Diniz Dalmolin, Dr. (UFSM)
(Presidente/Orientador)

Alexandre ten Caten, Dr. (UFSC)
(Co-orientador)
participante por videoconferência

Elpídio Inácio Fernandes Filho, Dr. (UFV)
participante por videoconferência

Fabrício de Araújo Pedron, Dr. (UFSM)

Jean Michel Moura-Bueno, Dr. (Embrapa Cocais)

Ricardo Bergamo Schenato, Dr. (UFSM)

Santa Maria, 25 de fevereiro de 2019.

AGRADECIMENTOS

Acredito que essa seja uma das partes mais importantes de uma tese. Esse é o momento de agradecer a todos que estiveram ao meu lado, pois se cheguei até aqui foi porque não estava sozinho. Acredito também que esse pode ser o momento de pedir desculpas à família, aos amigos e colegas que não receberam a atenção merecida e, repetidamente, ouviam como resposta para convites frases como: "- dessa vez não vai ter como ir, o trabalho está apertado". Portanto, a cada um, meu mais sincero agradecimento e também pedido de desculpas.

À Deus, pela vida e por tudo que tem proporcionado. Por me dar forças para que chegasse até aqui.

Aos meus pais, Nelso Piovesan Cancian e Lucia Campos Cancian, por serem meus exemplos, por todo incentivo e compreensão, dedico isso a vocês.

À minha amada noiva, Amanda Grassmann, faltam palavras para agradecer. Por todo o apoio e incentivo em todos os momentos, pela maravilhosa companhia em todos esses anos e pelos conselhos que sempre acalmam. Sem você, eu não teria chegado até aqui.

À toda a família, pelo incentivo e compreensão pela ausência em diversos momentos nesses anos.

Ao mestre Ricardo Dalmolin, pela amizade, pela valiosa orientação, incentivo e confiança, sempre sendo um exemplar profissional e ser humano.

Ao meu co-orientador Alexandre ten Caten, pela disponibilidade, pelos relevantes conselhos e por todo o auxílio; e aos professores de Pedologia da UFSM, Fabrício Pedron e Ricardo Schenato, pela amizade e pelos valiosos ensinamentos repassados.

À Universidade Federal de Santa Maria, ao Programa de Pós-Graduação Ciência do Solo e aos professores do Departamento de Solos da UFSM, pelo apoio e estrutura disponibilizada.

À CAPES, pela concessão da bolsa de estudos.

Aos colegas e amigos do incrível Laboratório de Pedologia da UFSM, Alessandro Samuel-Rosa, André Dotto, Ândrea Franco, Diego Gris, Estelita Penteado (*in memorian*), Gabriel Deobald, Guilherme Reis, Ismael Backes, Jéssica Costa, João Pedro Flores, Jordano Maffinni, Juliana Lorensi, Leonardo Zotelle, Luís Antônio Santos, Miriam Rodrigues, Mario Wolski, Nícolas Rosin, Pedro Nascimento, Pedro Saccol, Rafael Paz Marques e, em especial, a Jean Bueno, Wagner Lopes, Taciara Horst e Daniely Vaz pela convivência diária.

À todos que contribuíram de forma direta ou indireta na concretização dessa importante etapa de minha formação, **MEU MUITO OBRIGADO!**

RESUMO

Tese de Doutorado
Programa de Pós-Graduação em Ciência do Solo
Universidade Federal de Santa Maria

ESTRATÉGIAS PARA PREDIÇÃO DE CLASSES DE SOLO

AUTOR: LUCIANO CAMPOS CANCIAN
ORIENTADOR: RICARDO SIMÃO DINIZ DALMOLIN
Santa Maria, 25 de fevereiro de 2019.

Com o intuito de disponibilizar informações com maior agilidade e resolução espacial adequada para suprir a demanda por informações sobre o solo, o mapeamento digital de solos (MDS) é uma alternativa para mapear classes e propriedades de solo, usufruindo da disponibilidade cada vez maior de técnicas processamento e mineração de dados. Nesse cenário, dados que permitam uma compreensão clara do desempenho científico e as relacionam com os padrões da produção científica global podem auxiliar nos caminhos a serem seguidos pela pesquisa, podendo contribuir inclusive com novas políticas públicas. A possibilidade de se fazer uso de informações previamente geradas sobre o solo, denominadas de dados legados, pode auxiliar com informações de entrada ao MDS a um custo reduzido, visto que não há necessidade de novas coletas. Como os produtos do MDS possibilitam a estimativa da incerteza, e uma análise abrangente pode contribuir para a qualidade dos mapas. Se quantificada e espacializada a incerteza, essas informações podem ser usadas para aprimorar a amostragem e otimizar a geração de informações. Dessa forma, os objetivos deste trabalho foram (1) caracterizar a produção científica em mapeamento digital de solos no Brasil e no mundo, no período de 1996 a 2017, nas bases de dados Scopus e Web of Science; e (2) avaliar técnicas de obtenção de dados adicionais para melhorar as predições de classes de solo com uso de dados legados. Para isso, foram realizados dois estudos. No primeiro, foram pesquisados de termos referentes ao MDS nas bases de dados, incluindo pesquisas de termos nos títulos, resumos e palavras-chave dos artigos. A partir disso, foi gerado um conjunto de índices bibliométricos dos resultados utilizando o pacote Bibliometrix em ambiente R. No segundo estudo, um mapa de classes de solo foi gerado com base em covariáveis ambientais, utilizando dados legados, em uma área de 13000 km² da região Central do Estado do Rio Grande do Sul, que está entre as áreas prioritárias do PronaSolos. Os mapas foram avaliados por validação cruzada e validação externa, além de mapas de incerteza expressarem as áreas com maior confusão do modelo. Adicionalmente, foram testadas estratégias para obtenção de pontos adicionais ao conjunto de calibração com base em mapas legados e reamostragem guiada na incerteza. O estudo 1 demonstrou que, no contexto geral, o crescente número de artigos em MDS foi publicado em sua maior parte na revista Geoderma. Entre os 10 com mais artigos publicados, a Revista Brasileira de Ciência do Solo é o único periódico de acesso aberto. Embora existam países na vanguarda do MDS, como Estados Unidos e Austrália, a posição do Brasil no número de artigos e autores não pode ser menosprezada, mostrando a importância da participação do país na pesquisa em MDS. O estudo 2 resultou em um mapa de classes de solo, gerado apenas com os dados legados, com acurácia de 0,49 na validação externa e incerteza geral de 0,84. Um conjunto híbrido, utilizando os dados legados de diferentes fontes foi capaz de melhorar acurácia para 0,55 e reduzir a incerteza para 0,77. Contudo, embora os dados do mapa legado terem trazido benefícios ao modelo, demonstraram inconsistências devido a sua escala. A reamostragem guiada pela incerteza, pela melhoria trazida ao modelo fazendo uso de uma pequena quantidade de dados, foi a estratégia que demonstrou o maior potencial. Nossos dados demonstram que o MDS é uma técnica promissora, podendo ser utilizado como metodologia no Programa Nacional de Solos (PronaSolos).

Palavras-chave: mapeamento digital de solos; análise bibliométrica; dados legados; incerteza, pedometria.

ABSTRACT

Doctoral Thesis
Graduate Program in Soil Science
Federal University of Santa Maria

STRATEGIES FOR SOIL CLASSES PREDICTION

AUTHOR: LUCIANO CAMPOS CANCIAN
ADVISOR: RICARDO SIMÃO DINIZ DALMOLIN
Santa Maria, February 25, 2019

In order to provide information with greater agility and adequate spatial resolution to supply the demand for soil information, digital soil mapping (DSM) is an alternative to map classes and soil properties, taking advantage of the increasing availability of techniques processing and data mining. In this scenario, data that allow a clear understanding of scientific performance and relate them to the patterns of global scientific production can help in the paths to be followed by the research, and may even contribute with new public policies. The possibility of making use of previously generated information on the ground, called legacy data, can help as input information to the DSM at a reduced cost, since there is no need for new collections. As the DSM products make it possible to estimate uncertainty, a comprehensive analysis can contribute to map quality. If uncertainty is quantified and spatialized, this information can be used to improve sampling and optimize information generation. Thus, the objectives of this work were (1) to characterize the scientific production in digital mapping of soils in Brazil and in the world, from 1996 to 2017, in the Scopus and Web of Science databases; and (2) evaluate additional data collection techniques to improve soil class predictions using legacy data. For this, two studies were carried out. In the first, we searched for terms related to MDS in the databases, including searches for terms in the titles, abstracts, and keywords of articles. From this, a set of bibliometric indexes of the results were generated using the Bibliometrix package in the R environment. In the second study, a soil class map was generated based on environmental covariates using legacy data in an area of 13000 km² of the central region of Rio Grande do Sul State, which is among the priority areas of PronaSolos. The maps were evaluated by cross-validation and external validation, in addition to uncertainty maps expressing the areas with greater confusion of the model. In addition, strategies were tested to obtain additional points to the calibration set based on legacy maps and guided-uncertainty resampling. Study 1 demonstrated that, in the general context, the increasing number of articles in DSM was published for the most part in the Geoderma journal. Among the 10 journals most published articles, the Revista Brasileira de Ciência do Solo is the only open access journal. Although there are countries at the forefront of DSM, such as the United States and Australia, Brazil's position in the number of articles and authors cannot be overlooked, showing the importance of the country's participation in DSM research. Study 2 resulted in a map of soil classes, generated only legacy data, with an accuracy of 0.49 in external validation and a general uncertainty of 0.84. A hybrid set using legacy data from different sources was able to improve accuracy to 0.55 and reduce uncertainty to 0.77. However, while legacy map data has brought benefits to the model, they have shown inconsistencies due to its resolution. The uncertainty-guided resampling, by the improvement brought to the model using a small amount of data, was the strategy that demonstrated the greatest potential. Our data demonstrate that DSM is a promising technique and can be used as a methodology in the Programa Nacional de Solos (PronaSolos).

Keywords: digital soil mapping; bibliometric analysis; legacy data; map uncertainty; pedometry.

LISTA DE TABELAS

CAPÍTULO 2

CAPÍTULO 3

LISTA DE FIGURAS

CAPÍTULO 2

CAPÍTULO 3

SUMÁRIO

1 INTRODUÇÃO GERAL

O solo é um recurso de extrema importância para o desenvolvimento da humanidade, desempenhando diversos serviços ambientais e vital para a produção de alimentos. Por fornecer inúmeros recursos para toda a população, há uma pressão muito grande em seu uso que, consequentemente, o modifica constantemente. Muitas vezes esse uso se dá de maneira inadequada, levando o solo à degradação e comprometendo a sustentabilidade.

Quem trabalha e maneja esse meio tão rico e tão suscetível deve estar ciente da importância que o solo possui e dos riscos que sua degradação proporciona. No intuito de divulgar esse assunto, diversos órgãos têm realizado ações visando a maior percepção da importância que o solo possui por parte de toda sociedade.

Por uma proposição da União Internacional de Ciência do Solo (IUSS), foi anunciada a Década Internacional do Solo (2015 – 2024) com o planejamento de realizar diversas atividades para o período, como campanhas de sensibilização pública, educação, divulgação de informações e a publicação de uma série de livros sobre o solo. No Brasil também vêm sendo desenvolvidas atividades com o mesmo intuito. A Sociedade Brasileira de Ciência do Solo criou, em 2011, a Comissão Especializada em Pedometria dentro da divisão “Solo no Espaço e no Tempo”, e a Embrapa Solos formou a Rede Brasileira de Pesquisa em Mapeamento Digital de Solos (RedeMDS), visando integrar pesquisadores da área. Pode-se citar como uma ação importante no âmbito nacional o projeto de execução do PronaSolos. Considerado o maior programa de pesquisa sobre o solo no Brasil, pelos próximos 30 anos o PronaSolos envolverá dezenas de instituições parceiras, dedicadas à investigação, documentação, inventário e interpretação dos dados de solos brasileiros. O objetivo principal do programa é mapear os solos de 1,3 milhão de km² do País nos primeiros dez anos, e mais 6,9 milhões de km² até 2048, em escalas que vão de 1:25.000 a 1:100.000. Isso demandará não só uma grande quantidade de mão de obra qualificada, mas também o desenvolvimento e aplicação de tecnologia para a execução desse desafio.

Dada a importância, visto o interesse de diversos órgãos pelo assunto e a grande demanda por informações referentes ao conhecimento detalhado da distribuição espacial dos solos e de suas propriedades, muito necessita ser feito para que essas informações sirvam como base para todas as áreas da Ciência do Solo. Por suas limitações, as técnicas convencionais de levantamento de solo apresentam dificuldades para atender tais demandas. O Mapeamento Digital de Solos (MDS), uma subárea da Pedometria, é uma alternativa para mapear classes e propriedades de solo, usufruindo da disponibilidade cada vez maior de

técnicas de processamento e mineração de dados, bem como de informações espaciais da superfície terrestre.

Nos últimos anos, houve avanços efetivos no que diz respeito a métodos e modelos de predição de classes e propriedades do solo pelo MDS. No entanto, ainda permanece nesse cenário a dificuldade em obter informações à campo, visto a demanda de tempo e mão de obra necessárias, implicando em um maior custo aos levantamentos. Uma possibilidade encontrada para contornar essa dificuldade é o uso de informações já disponíveis sobre solos, obtidas por levantamentos convencionais, os quais ganharam a denominação de "dados legados".

Para que se consiga uma cobertura completa das propriedades do solo em resolução ou escalas adequadas, com uma maior eficiência e possibilitando mapeamentos com um baixo custo de amostragem e análise eficiente, esforços devem ser reunidos para o desenvolvimento dessa área tão importante à Ciência do Solo.

Seguindo uma tendência mundial e encarando a necessidade de serem exploradas todas as possibilidades de uso dos dados já existentes para que se obtenha a maior cobertura possível do território mapeada, abre-se uma lacuna no conhecimento que deve ser preenchida não só com as técnicas de MDS, mas com o uso de dados legados e identificar estratégias para superar suas peculiaridades, como a imperfeita cobertura amostral.

O presente trabalho aborda o tema da utilização do MDS como ferramenta para geração de novas informações sobre o solo. No primeiro capítulo, por meio de dados de livros, revisões bibliográficas e artigos, será formada uma base de informações que visam auxiliar na formação de conceitos, enriquecimento de ideias e uso de metodologias.

No segundo capítulo, é apresentado o artigo resultado da segunda pesquisa da presente tese, intitulado "Bibliometric Analysis for Pattern Exploration in Worldwide Digital Soil Mapping Publications", publicado no periódico Anais da Academia Brasileira de Ciências. O artigo busca contextualizar e analisar a pesquisa em MDS no Brasil e no mundo, permitindo a identificação de tendências e possíveis caminhos a serem seguidos para que mais dados sobre o solo sejam gerados em nosso país. A partir dessas tendências, será possível identificar as dificuldades que o MDS encontra e buscar alguns caminhos que possam contribuir para o decorrer dessa pesquisa. O terceiro capítulo tem como intenção fazer uso de dados legados como informações de entrada ao MDS, buscando, com dados já disponíveis, aplicar e avaliar técnicas de reamostragem para o mapeamento de classes de solo na região central do Estado do Rio Grande do Sul.

2 HIPÓTESES

A análise bibliométrica permite avaliar o posicionamento dos trabalhos em MDS desenvolvidos no Brasil comparados com os países de vanguarda nessa temática e identificar caminhos a serem seguidos pela pesquisa brasileira.

Mapas de classes de solo podem ser gerados utilizando dados legados e melhorados fazendo uso de técnicas como a reamostragem a campo guiada pelo mapa de incerteza ou obtenção de dados em mapas legados pode reduzir a incerteza geral e aumentar a acurácia do mapa de solo gerado por técnicas de MDS.

3 OBJETIVO GERAL

O objetivo geral deste trabalho foi avaliar, através de uma análise bibliométrica, as tendências da pesquisa em MDS no Brasil e no mundo e estabelecer estratégias para a predição de classes de solo e possível utilização no PronaSolos.

3.1 OBJETIVOS ESPECÍFICOS

Caracterizar, com base em um conjunto de indicadores bibliométricos, a produção científica sobre MDS para o Brasil e o mundo, visando identificar características e peculiaridades na produção científica nacional e global no MDS, fazendo a previsão das tendências de crescimento nessa área do conhecimento e indicando possíveis caminhos a serem seguidos

Verificar a viabilidade do uso de dados legados, além de avaliar técnicas de obtenção de dados adicionais para melhorar as predições de classes de solo.

4 CAPÍTULO 1 – REVISÃO BIBLIOGRÁFICA

4.1 A IMPORTÂNCIA DE MAPEAR O SOLO

O solo é um componente essencial para os ecossistemas terrestres, pois possui relação desde a produção de alimentos até a regulação do clima. Embora o solo possua fundamental importância, ainda há uma carência muito grande de informações sobre classes e propriedades de solos (BREVIK et al., 2016), impedindo o planejamento de atividades ou de impactos que venham a ser gerados pelo seu uso (HARTEMINK et al., 2013). Além disto, percebe-se que grande parte das informações sobre o solo no mundo não possuem informações quanto a sua exatidão ou não estão disponíveis de forma digital.

A possibilidade de mapear classes de solo foi possível após a determinação dos conceitos iniciais sobre a formação do solo feita pelo cientista russo Vasili Vasilevich Dokuchaev, que em 1883 realizou um estudo sobre os "Chernozems" – solos escuros das estepes da Rússia – onde desenvolveu e modelou conceitos sobre a natureza e a gênese de tipos de solos e relação com a paisagem. Dokuchaev aplicou princípios básicos de morfologia para caracterizar os principais grupos de solos, elaborando o que pode ser chamada de primeira classificação científica de solos e instituindo os alicerces para a ciência do solo moderna.

Em 1941, com base nesses estudos, Hans Jenny concluiu que os fatores de formação não são causas, mas variáveis independentes, e formalizou as bases conceituais sobre a gênese e a distribuição dos solos na paisagem, propondo um modelo clássico que aborda os fatores relacionados com a formação do solo. O solo (S) é função (*f*) do clima (cl), dos organismos (o) e do relevo (r), agindo sobre o material de origem (p) durante um período de tempo (t), formando assim o modelo *clorpt* (Equação 1).

$$S = f\,(cl, o, r, p, t, \ldots) \quad (1)$$

As reticências no final da equação demonstravam que Hans Jenny possuía ciência de que o solo é um sistema aberto, cuja formação poderia ainda incluir outros fatores. A equação de Jenny foi pioneira na representação dos processos pedogenéticos e caracterizou de uma forma complexa a capacidade de predição de um solo em função da interação dos seus fatores de formação. Dessa forma, se o solo é uma função dos processos pedogenéticos e se a

distribuição espacial desses fatores na paisagem for conhecida, conclui-se que o solo pode ser compreendido a partir da obtenção de informações de campo e construção de um modelo de relação solo-paisagem (JENNY, 1941).

O mapeamento do solo exige sistemas de classificação que permitam a comunicação das informações mapeadas, a compreensão do sistema do solo e a criação de modelos de solo para obter essa compreensão (BREVIK et al., 2016). A possibilidade de representar essa informação em mapas pedológicos permitiu o avanço de muitas áreas da ciência do solo, pois esses apresentam informações primordiais e servem de base também para o planejamento de uso das terras (DALMOLIN et al., 2004).

O conhecimento sobre os atributos do solo, bem como a sua distribuição espacial, proporcionam a compreensão do sistema como um todo e direcionam a necessidade de práticas de manejo adequadas (ARRUDA et al., 2013; LAGACHERIE e McBRATNEY, 2006). Além disso, muito vem se debatendo sobre o uso das informações do solo também em aplicações ecológicas e econômicas, especialmente a adaptação às mudanças ambientais (GRIMM e BEHERENS, 2010)

Os mapas de solos são indispensáveis para a compreensão e divulgação do conhecimento por conterem informações fundamentais para sua sustentabilidade respeitando a aptidão de uso das terras (DALMOLIN et al., 2004). Dada a importância dos mapas de solo e sua estreita relação com os ciclos hidrológicos, ciclagem de nutrientes e produção de alimentos e energia (HARTEMINK e McBRATNEY, 2008), houve um aumento considerável na demanda por essas informações nos últimos anos (BAZAGLIA FILHO et al., 2013).

Uma das necessidades mais pertinentes no cenário atual da pesquisa em solos diz respeito ao conhecimento da distribuição espacial dos solos e de suas propriedades e características. Para que essas informações possam ser nem utilizadas, é necessário o mapeamento com alto nível de detalhamento, seja por mapas de classe de solo ou também pela criação de mapas de aptidão agrícola, uso das terras e área de risco de erosão (VAN ZIJL et al., 2014) ou até mesmo definir estratégias sobre a segurança do uso do solo (KIDD et al., 2018). O conceito de segurança do solo vem sendo debatido e aplicado em diversos estudos com o intuito da conservação e uso racional de seus recursos (AMUNDSON et al, 2015; BREVIK et al., 2016; HUANG et al., 2018). A segurança do solo é um conceito abrangente, motivado pelo desenvolvimento sustentável e visando a manutenção e aumento da produção alimentos, fibras e água doce no mundo, além de contribuir para a sustentabilidade energética e climática, mantendo a biodiversidade e a proteção geral do ecossistema (McBRATNEY et al., 2014).

Obter essas informações por técnicas convencionais de levantamento do solo torna-se uma tarefa difícil devido a uma série de limitações, que vão desde maior tempo de execução, maior necessidade de mão de obra, não apresentam a incerteza das informações e não são reprodutíveis, não conseguindo assim atender as crescentes demandas por tais informações. Levantamentos de solos abrangem trabalhos realizados em escritório, campo e laboratório (IBGE, 2018), cujo interesse principal é o registro de observações e interpretações de características do local e dos solos encontrados, a partir da formação de um modelo mental (DALMOLIN et al., 2004; DALMOLIN e TEN CATEN, 2015), utilizando dados do meio físico para definir unidades com características homogêneas na paisagem (BERG e OLIVEIRA, 2000; WOLSKI et al., 2017). Para McBratney et al. (2000) e Bazaglia Filho et al. (2013), um questionamento que também recai sobre os levantamentos convencionais é devido a esses possuírem informações em sua grande maioria qualitativas, sendo subjetivas e podendo variar com a interpretação de cada pedólogo.

Buscando aprimorar a qualidade das informações disponibilizadas, os levantamentos de solos devem se apropriar das novas tecnologias, trazendo informações não só de classes, mas também de atributos do solo com maior nível de detalhe permitindo agilidade na geração das informações (DALMOLIN e TEN CATEN, 2015; PINHEIRO et al., 2018).

Conforme Sanchez et al. (2009), somente 109 países possuem mapas de solos nas escalas 1:1.000.000 ou maiores, o que representa cerca de 30% da superfície terrestre com mapeamentos disponíveis. Embora muito esforço tenha sido aplicado nesses últimos anos para aumentar a cobertura de mapeamento, esse número com certeza ainda não é o suficiente para o estudo dos solos do planeta. No Brasil, todo o território está mapeado nas escalas 1:5.000.000 e 1:1.000.000, que são consideradas inadequadas para determinadas aplicações (DALMOLIN et al., 2004). Contudo, apenas 35% do território possui mapeamento completo com escalas entre 1:100.000 e 1:600.000, e uma área muito restrita tem mapas de solos com maior detalhamento (MENDONÇA-SANTOS e SANTOS, 2006). Isso tem feito com que muitos interessados por informações, inclusive, necessitem busca-las a partir de contratação particular de prestadores de serviço (BAZAGLIA FILHO et al., 2013).

A falta de mapeamentos decorre do maior tempo de execução e do elevado custo do mapeamento através dos métodos convencionais, o que motiva o estudo de métodos alternativos de levantamento pedológico (McBRATNEY et al., 2003). Apesar de terem promovido avanços, os levantamentos convencionais apresentam limitações não só pelo custo e tempo, mas principalmente pelas características dos seus resultados (McBRATNEY et al., 2012; BREVIK et al., 2016).

Métodos que reduzem a quantidade de trabalho em campo possuem potencial para aplicação em pesquisas (FINKE, 2000). Sob essa ótica, o mapeamento por métodos convencionais pode ser considerado mais oneroso e demorado, especialmente para grandes áreas (FORKUOR et al., 2017). A busca por novos métodos passa pela união do conhecimento já existente em solos e pela utilização e adequação de técnicas muitas vezes já utilizadas em outras áreas da ciência.

Os primeiros mapas que se tem conhecimento usavam uma abordagem semelhante aos mapas feitos há 50 anos, onde polígonos delineiam propriedades discretas, tornando a precisão do mapa do solo prejudicada devido à dificuldade de lidar com a complexidade do solo (McBRATNEY et al., 2019). Brevik et al. (2016) ressalta que as limitações dos mapas de solo também estão ligadas ao fato de as unidades de mapeamento serem generalizadas para se ajustarem à quantidade de informação que o pedólogo poderia interpretar a partir das observações de campo disponíveis, bem como serem legivelmente delineadas na escala cartográfica de produção selecionada.

O avanço da tecnologia e a criação de sistemas inteligentes permitiu a integração da ciência do solo com sistemas computacionais (HARTEMINK e McBRATNEY, 2008), principalmente com interesse em usar algoritmos computacionalmente mais intensivos (HENGL et al., 2018) e com capacidade de lidar com a demanda por informações precisas e atualizadas sobre o solo (HARTEMINK e McBRATNEY, 2008). Como resultado dessa maior disponibilidade de tecnologias para coleta e análise de dados, um novo conceito está sendo adotado em levantamento de solos. A Pedometria e o MDS são uma opção com grande potencial para abastecer as novas demandas por informações (MCBRATNEY et al., 2003; MCBRATNEY et al., 2012; HENGL et al., 2018), tornando possível a predição de atributos em toda a paisagem como uma opção para obter informações sobre o solo.

4.2 A PEDOMETRIA E O MAPEAMENTO DIGITAL DE SOLOS

O termo Pedometria deriva-se do grego pedos (solo) e metron (medida). Essa área da Ciência do Solo tem por objetivo principal decifrar os problemas referentes às incertezas intrínsecas ao método convencional de mapeamento (McBRATNEY et al., 2019). A Pedometria faz uso de novas abordagens de modelagem quantitativa, aprofundando o conhecimento da variabilidade espacial dos solos e possibilitando quantificar a qualidade e

precisão das informações geradas com uso de modelos numéricos e/ou estatísticos (MCBRATNEY et al., 2003).

A Pedometria é uma área do conhecimento que faz uso da ciência do solo, geoprocessamento, matemática e estatística. Seu uso permite que sejam criadas relações numéricas entre determinada propriedade ou atributo relevante do solo e variáveis preditoras, permitindo caracterizar a variação espacial do solo de modo quantitativo (McBRATNEY et al., 2000; LAGACHERIE e McBRATNEY, 2007). Com uma maior intenção em compreender a formação dos solos do que propriamente prever a distribuição espacial, a própria abordagem inicial sobre a formação dos solos, o modelo *clorpt* (JENNY, 1941), não considerava o fator espacial – primordial para a geração de mapas - em sua função.

Com a disponibilização e popularização do GPS, tornou-se mais fácil a obtenção precisa de amostras georreferenciadas à campo. Ficou mais fácil também a obtenção de informações que sirvam de apoio à criação de relações matemáticas entre o solo e a paisagem, as chamadas covariáveis ambientais.

Numa busca de suprir a falta de informações quantitativas sobre o solo, como uma subárea derivada da Pedometria, o MDS teve suas bases estabelecidas por McBratney et al. (2003) e surgiu como uma opção para suprir as demandas por informações de solo, visando gerar informações com uma com maior velocidade e resolução espacial adequada sobre a distribuição espacial das classes de solo ou de seus atributos.

O MDS pode ser definido como a criação de sistemas de informação espacial com uso de modelos numéricos para a inferência das variações espaciais e temporais nos diferentes tipos de solos e seus atributos a partir de observações a campo, do conhecimento dos solos e de variáveis ambientais correlacionadas (LAGACHERIE e MCBRATNEY, 2007). Também é fundamental compreender que as bases do MDS, representadas pelo modelo *s.c.o.r.p.a.n* (MCBRATNEY et al., 2003), são uma adaptação moderna da equação de Jenny (1941) para representar os fatores de formação do solo, onde s: solo; c: clima; o: organismos; r: topografia; p: material originário; a: idade; n: localização espacial.

O MDS não se refere apenas a um método de mapear o solo por meio de técnicas computacionais ou geração de relações quantitativas entre covariáveis ambientais e propriedades de solo (MINASNY e McBRATNEY, 2016). Para que a aplicação do MDS seja ainda mais ampla, há necessidade de mais informações e principalmente de um banco com dados precisos e georreferenciados de solos (DALMOLIN e TEN CATEN, 2015).

A partir da introdução do conceito *scorpan* no MDS muitos estudos vêm sendo realizados, alavancando a realização de vários eventos e workshops com o intuito de divulgar

as evoluções obtidas à comunidade científica do solo. Em 2004 foi realizada a primeira oficina mundial de MDS, em Montpellier. A partir desse evento, foi formado um grupo de trabalho da União Internacional de Ciência do Solo, o qual propôs a organização de eventos frequentes relacionados ao assunto. Em escala mundial foram realizadas oficinas globais, sendo o segundo workshop global em MDS no Rio de Janeiro, em 2006; e posteriormente em Logan, Estados Unidos, em 2008; Roma, Itália, em 2010; Sydney, Austrália em 2012; Nanjing, China, em 2014 e em Aarhus, Dinamarca em 2016. A cada dois anos, cientistas do solo se reúnem para obter atualizações sobre os últimos avanços no MDS e discutir ações dos grupos de trabalho (McBRATNEY et al., 2019), demonstrando o interesse global pelo MDS, visando discutir, alavancar e difundir o conhecimento gerado.

O Brasil, seguindo a tendência mundial de crescimento da pesquisa em MDS, também tem buscado desenvolver essa linha de pesquisa dentro do país. Dentre as iniciativas tomadas, cabe destacar a formação da Comissão Especializada em Pedometria, englobada na divisão Solo no Espaço e no Tempo da Sociedade Brasileira de Ciência do Solo (SBCS), além da criação da RedeMDS, que visa o agrupamento dos pesquisadores interessados em desenvolver e divulgar essa área da Ciência do Solo no país. Também foi realizado, em 2016, o I Encontro Brasileiro de Pedometria - Pedometrics Brazil, buscando promover o avanço da pesquisa em Pedometria e MDS no país.

Após a segunda oficina global em 2006, foi iniciado o projeto GlobalSoilMap, considerado o projeto de maior abrangência internacional em ciência do solo (ARROUAYS et al., 2014; HEMPEL et al., 2014). O GlobalSoilMap tem por objetivo principal desenvolver um mapa digital de solos que contemple todas as áreas continentais do globo terrestre, a partir de um banco de dados de propriedades do solo. Esse banco de dados será construído principalmente por meio de informações do solo existentes em dados legados e covariáveis ambientais. O projeto é necessidade da demanda por informações de solos mais precisas e detalhadas por todas as áreas da ciência do solo (ARROUAYS et al., 2014; DALMOLIN e TEN CATEN, 2015). Visando a obtenção de informações dentro de padrões, Arrouays et al. (2014) citam nas especificações do projeto GlobalSoilMap que os dados devem estar de acordo com algumas normas para obtenção de um produto final padronizado.

Cabe salientar também a importância da cada vez maior disponibilização de modelos digitais de elevação (MDE) para o MDS, advindos de diferentes bases e com qualidade e aplicação variada. Esses são fontes importantes de covariáveis preditoras para uso no MDS, visto que é uma das principais representações espaciais do relevo na paisagem, sendo base

para a maioria dos modelos de predição de classes ou de atributos do solo, caracterizando-se como uma importante ferramenta para representar o relevo (WYSOCKI et al., 2012).

Além dos dados que contemplem o relevo, dados oriundos de outras fontes também são aproveitados, como os índices derivados de imagens de sensoriamento remoto ou mapas geológicos (HOFIG et al., 2014). Covariáveis ambientais mais detalhadas, além de transmitirem mais informações e representarem com maior precisão a condição real do solo (SAMUEL-ROSA et al., 2015), também proporcionarão ao modelo matemático capacidade de realizar predições de mais precisas de acordo com a escala (MAYNARD e JOHNSON, 2014).

Fazendo uso de um conceito muito similar ao do levantamento convencional de solos, sistemas de aprendizado de máquina necessitam a inclusão de dados quantitativos que, de uma forma ou de outra, expressem os fatores atuantes na formação dos solos e realize a predição da ocorrência dos solos baseado nessas informações. Diante disso, o MDS é uma opção para facilitar a viabilidade da execução de levantamentos de solos (HÖFIG et al., 2014), pois faz uso principalmente de informações digitais que explicitam a estreita relação entre solo e relevo, como apresentado em Horst et al. (2018).

Essa capacidade de criar relações da distribuição de classes ou atributos do solo na paisagem pode ser tratada como uma das vantagens do MDS (DONG et al., 2019). Com tais informações, posteriormente, pode-se prever também a ocorrência das mesmas classes de solo em áreas ainda não amostradas por meio do uso de dados gerados em áreas geograficamente semelhantes (LAGACHERIE e VOLTZ, 2000) por extrapolação fazendo uso do conceito de área de referência, como demonstrado por Wolski et al. (2017).

Cabe salientar que, independentemente do método utilizado para realizar a predição, a obtenção de um bom resultado está condicionada à qualidade dos dados de entrada, ou seja, a resolução do MDE, das covariáveis ambientais utilizadas, a quantidade dos dados e por último, mas não menos importante, a qualidade das amostras utilizadas para o treinamento do método preditivo.

Dentre os métodos de predição usados no MDS mais utilizados para a predição de classes de solo, pode-se citar regressões logísticas múltiplas multinomiais (GIASSON et al., 2006; TEN CATEN et al., 2012), modelos logísticos com aplicação de componentes principais (TEN CATEN et al., 2011), máquina vetor de suporte (BEHRENS e SCHOLTEN, 2007), árvores de decisão (SILVA et al., 2016) e random forest (TAGHIZADEH-MEHRJARDI et al., 2015).

Dentre os métodos mais robustos, a predição de classes de solo por métodos do tipo árvore de decisão é favorecida pela sua robustez como método preditivo. Contudo, o seu

potencial em explicar e esclarecer as relações existentes entre os fatores de formação, a paisagem e classes de solos é incipiente (TEN CATEN, 2011) e não vem sendo empregado na maioria dos trabalhos realizados.

Tendo alguns conceitos semelhantes aos da árvore de decisão, o random forest foi desenvolvido por Breiman e Cutler (2009), sendo um conjunto de árvores de classificação e regressão. Contudo, não contém apenas uma árvore de regressão padrão, como na árvore de decisão, mas várias árvores de regressão, formando uma "floresta" e permitindo uma maior quantidade de regras de classificação. Cada árvore é treinada a partir de uma amostra de *bootstrap* independente e aleatória, onde um subconjunto de dados é usado para treinar a árvore e os pontos restantes são deixados para validar a árvore. Além disso, a cada divisão, um subconjunto aleatório de covariáveis preditoras é escolhido. A partir desse subconjunto aleatório, a covariável mais importante é selecionada para dividir os dados. Diante disso, o random forest é um dos métodos mais promissores para o mapeamento de classes de solo (PAHLAVAN RAD et al., 2014; TAGHIZADEH-MEHRJARDI et al., 2015), mostrando boa eficiência e robustez (SUBBURAYALU e SLATER, 2013; HEUNG et al., 2016; HEUNG et al., 2017)

Com o avanço das técnicas e a maior geração de dados de solos, torna-se cada vez mais necessário avaliar a qualidade das informações geradas. Qualquer mapa pedológico, seja ele obtido através da forma tradicional ou digital, deverá passar por uma verificação da qualidade e veracidade das informações nele contidas. Essa verificação pode ser realizada através da comparação do mapa com outra referência já existente ou através da verificação em pontos de validação (ROSSITER, 2004).

Para conferência da exatidão e validação da modelagem aplicada, muitas vezes são escolhidos locais representativos para coleta de amostras e visualizadas características, onde, a partir das feições visualizadas a campo, podem ser estabelecidas relações matemáticas para confirmar a ocorrência de determinada classe de solo. Dados qualitativos ou quantitativos, obtidos a partir de relatórios de levantamentos convencionais ou pontos obtidos à campo, tais como as classes dentro de uma área, uma área ocupada por determinada classe ou os locais na paisagem onde predominam determinada classe de solo são possibilidades para avaliar a qualidade dos modelos gerados (ODGERS et al., 2014b). Como o mapeamento convencional baseia-se na descrição qualitativa da relação solo-paisagem (GRUNWALD, 2005), o MDS torna mais fácil a avaliação das informações geradas por sua natureza quantitativa (MCBRATNEY et al., 2003), pois proporciona formas mais adequadas de avaliar sua acurácia.

As medidas de acurácia em mapas têm como principal propósito verificar a magnitude de concordância de um mapa com a realidade de uma área a qual pretende-se representar. A avaliação da acurácia do mapa pode ser obtida a partir da matriz de erros da classificação, da qual pode-se calcular estimativas de exatidão como a acurácia global (CONGALTON, 1991). A acurácia pode ser definida como a estimativa da porcentagem de área mapeada que foi classificada de forma exata quando comparada com a validação dos dados à campo.

Pelo fato de as classes de solo serem variáveis qualitativas, ou seja, não possuem uma variação escalar ou numérica, a avaliação do comportamento da distribuição dos erros ou acertos de predição torna-se muito subjetivo. Há ainda o fato de que classes pertencentes a uma mesma ordem ou que possuem similaridade na paisagem podem ser consideradas morfológica ou taxonomicamente próximas, diminuindo assim o peso do erro gerado na predição dessa classe.

A crescente geração de informações sobre o solo tem despertado um crescente interesse na determinação da incerteza sobre as informações geradas sobre o solo (BREVIK et al., 2016). Como os produtos do MDS possibilitam a estimativa da incerteza (MINASNY e BISHOP 2008), atrelado ao fato dessa análise valorizar os produtos do MDS do ponto da qualidade da estimativa, uma análise abrangente pode contribuir para a qualidade dos mapas (STUMPF et al 2017).

Stumpf et al. (2017) faz uma boa abordagem sobre os diferentes tipos de erro no MDS, descrevendo os componentes de maior influência nas discrepâncias do modelo preditivo. Segundo Nelson et al. (2011), a incerteza no MDS tem origem em quatro fontes de erro, que são erro de covariável, de modelo, posicional e analítico. O erro de covariável tem origem na medição ou interpolação (BISHOP et al., 2006). O erro do modelo preditivo refere-se a uma compreensão insuficiente ou a uma simplificação excessiva dos processos que ocorreram no solo (MINASNY e BISHOP, 2008). O erro de posicionamento vem de dados espaciais localizados de forma errônea ou inadequada (GRIMM E BEHRENS, 2010). O erro analítico refere-se a erros de medição de propriedades do solo que ocorrem durante a análise laboratorial e, conforme Viscarra Rossel e McBratney (1998), são baixos em comparação com outras fontes de erro descritas anteriormente.

Como o MDS tornou possível estimar a incerteza, ele agrega valor aos produtos de levantamento de solo, uma vez que permite indicar a qualidade de previsão das variáveis de estudo. Se a incerteza é quantificada e espacializada, também permite melhorar a amostragem e otimizar a geração de informações do solo. Stumpf et al. (2017) usaram um método de reamostragem guiada pela incerteza para melhorar as previsões de silte e argila em uma

pequena bacia na China. Com a estratégia, a incerteza espacial média foi reduzida para a predição de silte e argila, demonstrando o potencial da técnica para uso no MDS.

4.3 O USO DE DADOS LEGADOS

A principal fonte de informação espacial de solos são os mapas gerados por meio do método convencional. Contudo, esses não disponibilizam as informações em escalas adequadas para estudos que requerem um maior nível de detalhamento, não conseguindo suprir as atuais demandas por informações do solo. Essa crescente demanda pelo conhecimento detalhado sobre o solo e seus atributos torna necessário, além da criação de novas técnicas, obter o máximo de informações possível dos levantamentos já existentes, visando reduzir o tempo e custo da realização de novos estudos (OMUTO et al., 2013), possibilitando também que sejam gerados mapas que descrevam a variação de atributos do solo (ODGERS et al., 2014b).

A maioria dos trabalhos desenvolvidos, focados em MDS, está fundamentada na construção de um modelo numérico entre as observações do solo no campo e relacionando-as com os fatores *scorpan* obtidos por diferentes meios. Nesse método, o modelo é aplicado a dados ambientais distribuídos espacialmente e geralmente com pontos coletados a campo. Para isso, é possível fazer uso de dados já gerados em outros trabalhos. O uso desses dados, chamados de dados legados, é fundamental não só para conseguir suprir a demanda por informações sobre o solo, mas também porque grandes recursos foram investidos para coletar e analisar esses dados (ARROUAYS et al., 2017; ZHANG et al., 2017) sendo utilizados em diversos trabalhos (HEUNG et al., 2016; SARMENTO et al., 2017; ZHANG et al., 2017; MACHADO et al., 2018), atingindo bons resultados.

Contudo, tais dados foram coletados e analisados para um propósito que pode ser diferente do objetivo de seu uso posterior. Características como método de distribuição dos pontos na área, número de pontos coletados ou atributos analisados podem tornar difícil ou inviabilizar sua posterior utilização como um dado legado. O tipo e a quantidade de dados de solo disponíveis é o que determina as técnicas apropriadas que podem ser usados para o MDS (McBRATNEY et al., 2003). Minasny e McBratney (2010) afirmam que a forma com que esses dados podem ser utilizados para o desenvolvimento de mapas digitais do solo dependem da disponibilidade e do tipo de dados, ou seja, se disponíveis em formato digital ou analógico, georreferenciamento das informações, dentre outros.

Os dados legados para mapear as propriedades categóricas do solo vêm tipicamente de uma das duas fontes: dados do perfil ou dados do polígono do solo que foram digitalizados a partir dos mapas criados pelo método convencional de mapeamento (HEUNG et al., 2017). O MDS, no entanto, também pode usar mapas convencionais de solo como entrada para fins de extrapolação para áreas vizinhas, atualização ou desagregação de mapas de solos (KEMPEN et al., 2009; ODGERS et al., 2014a; PAHLAVAN-RAD et al., 2014). A desagregação de mapas de solo consiste no uso de técnicas de MDS para reestruturar os mapas, desagregando espacialmente as informações no mapa do solo em unidades componentes, de uma maneira que represente melhor a natureza contínua dos solos (BAGATINI et al., 2016; SARMENTO et al., 2017; ZERAATPISHEH et al., 2019).

Contudo, muitos mapas não possuem informações suficientemente detalhadas para tal fim, não permitindo que sejam extraídas informações que não foram incluídas no mapa em sua concepção. Como a desagregação de mapas requer uma descrição detalhada das características da paisagem onde ocorrem as classes de solo para permitir a recriação do mapa de solo com um maior nível de detalhe (HEUNG et al., 2017), isso pode ser um empecilho. Pelo fato de muitos levantamentos possuírem um banco de dados de perfis de solo descritos (ROSSITER, 2004) esses dados se tornam úteis em situações onde as informações contidas no levantamento de solos sobre o mapa não permitam um melhor detalhamento pela técnica de desagregação de mapas ou quando a resolução espacial de levantamentos de solo existentes é muito grosseira. (HEUNG et al., 2017)

Um procedimento genérico para o uso de mapas do solo envolve a geração de pontos de treinamento dentro de polígonos onde as unidades de mapeamento são apresentadas. As principais vantagens do método de polígonos incluem a capacidade dos usuários de selecionar um tamanho de amostra suficiente grande, o que é benéfico para capturar mais a variabilidade da paisagem e o espaço de características de uma variável categórica (HEUNG et al., 2014; HEUNG et al., 2016). Heung et al. (2017) usaram pontos de treinamento derivados de polígonos para o mapeamento de classes de solo, obtendo acurácia de 68% ao usar dados de treinamento derivados de polígonos, ressaltando o bom resultado obtido com o uso dessa abordagem.

Lagacherie e McBratney (2007) sugerem que os dados legados do solo disponíveis devem ser bem organizados antes que seja aplicada qualquer metodologia para o MDS. Sendo assim, um banco de dados apropriado deve ser desenvolvido e sistematizado para armazenar os dados legados harmonizados. Diante disso, uma iniciativa que ganhou destaque é o GlobalSoilMap (ARROUAYS et al., 2014), que tem por objetivo o agrupamento de dados

legados do solo, seu processamento e a geração de informações consistentes em escala global. Esse projeto, aliás, teve seu marco inicial em um grupo de trabalho do workshop global em MDS de 2006, realizado no Rio de Janeiro, Brasil. Servindo de base para a compilação de dados para o GlobalSoilMap, Arrouays et al. (2017) estabelecem que a recuperação de dados do solo inclui três etapas principais: (1) manutenção de bibliotecas e acervos, incluindo digitalização de relatórios e mapas de papel analógico para formatos digitais; (2) compilação dos dados do solo sob um padrão comum das fontes de dados resgatadas, a partir do agrupamento de perfis e dados de solo com padronização de dados , harmonização e controle de qualidade; e (3) quando compilados, os dados legados são usados para gerar mapas de propriedades do solo em grade dentro da iniciativa GlobalSoilMap.

Experiências satisfatórias como o Soil Survey Geographic Database (SSURGO), nos Estados Unidos, que auxilia o MDS com o objetivo de mapear todas as terras aráveis do país (CHANEY et al 2016), também podem ser citadas. Outro exemplo é o ISRIC World Soil Information, que em 2014 divulgou o SoilGrids - um sistema global de informações do solo que utiliza cerca de 110.000 perfis de solo (McBRATNEY et al., 2019), compilando dados de solo de todo o mundo. Essas estratégias podem servir de inspiração para serem adotadas por outros países para esse fim, facilitando a obtenção de informações sobre o solo.

Recentemente, o Brasil apresentou uma ferramenta com potencial nesse sentido. O Repositório Brasileiro Livre para Dados do Solo Aberto (RBLDAS) (www.ufsm.br/febr) é uma iniciativa que permite aos cientistas do solo publicarem seus conjuntos de dados. O RBLDAS tem como objetivo organizar o armazenamento e permitir o compartilhamento de todos os tipos de informações do solo no Brasil. Dessa forma, é possível que os dados legados do solo sejam utilizados em outros estudos, aumentando também a colaboração entre os cientistas do solo (SAMUEL-ROSA et al. 2018), sendo essa uma importante fonte de dados legados.

Em uma estrutura organizada com conjunto de dados que dá a flexibilidade para acomodar muitos tipos de dados de qualquer variável do solo, facilita a participação de cientistas do solo na recuperação e avaliação de qualidade de dados legados. O RBLDAS já permite o acesso às 14.477 observações de solo - 42% delas do sul e sudeste do Brasil - de 232 conjuntos de dados, disponibilizando também ferramentas de visualização e pesquisa específicas de dados.

Na maioria dos países existem dados legados obtidos por diferentes pesquisas, em diferentes anos e com diferentes propósitos (SULAEMAN et al., 2013). Nesse contexto, os dados úteis incluem mapas que têm uma boa referência ou observações a campo e do solo que

estão prontos para uso, ou seja, os dados foram completamente tabulados e possuem localização geográfica.

No caso de dados obtidos de levantamentos mais antigos, onde o acesso à equipamentos para georreferenciamento era limitado e as localização eram realizadas de forma empírica, apenas pela descrição do local, há uma maior dificuldade. Nesse caso, é necessário obter a posição geográfica por meio das informações contidas no relatório sobre o local de coleta, geralmente em locais de fácil acesso, como beiras de estrada. Samuel-Rosa et al. (dados não publicados) propuseram uma metodologia para a aquisição de coordenadas pela busca de informações descritas nos relatórios em um serviço de mapeamento da web (www.google.com.br/maps). Esses mesmos autores citam que a partir do mecanismo de busca que pode ser usado para encontrar cidades, estradas e corpos de água, é possível estimar as coordenadas de muitas das observações sem geolocalização.

Isso demonstra os esforços para utilizar dados legados no MDS para diversas possibilidades. Diante disso, o mapa de classes de solo torna-se um meio, e não um fim na obtenção de informações sobre o solo. Mesmo em regiões onde todo o território está mapeado, o MDS ainda se faz extremamente necessário para monitorar a capacidade dos solos para lidar com questões ambientais globais, mapeando aspectos de segurança do solo e identificando áreas com maior necessidade de proteção (McBRATNEY et al., 2019).

As demandas por diferentes tipos de informação sobre o solo surgem de diferentes campos da ciência, e devido a isso o estudo sobre o solo deve sempre permitir a geração de informações sobre esse aspecto. O aproveitamento das diferentes fontes de dados e os avanços tecnológicos deve ser sempre fonte de estudo para que a ciência do solo esteja em vanguarda não só na geração de informações, mas também na aplicação para o grande objetivo geral: a conservação e manutenção da sustentabilidade do solo.

4.4 A BIBLIOMETRIA COMO UMA FERRAMENTA NA PESQUISA

Uma forma de avaliar o vigor de uma área da ciência é mensurar o número de publicações ao longo do tempo, sendo uma indicação da produtividade (HARTEMINK e McBRATNEY, 2008), tornando-se mais importantes ainda pelo fato de que medidas de impacto científico estão sendo cada vez mais utilizadas para promoções acadêmicas, avaliação de bolsas e avaliação de candidatos a vagas de emprego (MINASNY et al., 2013). De acordo com Merton (1968), a publicação científica é a força vital da ciência e um dos

objetivos principais de qualquer atividade científica, isto porque é um procedimento pelo qual os resultados da investigação são comunicados, trocados e verificados com toda a comunidade científica.

O estudo de dados métricos, dentre eles a bibliometria, surgiram justamente da necessidade de avaliar a produção científica (NORONHA e MARICATO, 2008). De acordo com Arvanitis e Chatelin (1988), a intenção da análise bibliométrica é obter uma visão geral de uma ciência, incluindo o material bibliográfico, a análise estatística, instituições e pesquisadores que trabalham em um determinado campo. Essa abordagem sobre a pesquisa também pode apresentar uma visão geral do trabalho e desenvolvimento de determinada área, além do impacto de sua literatura científica.

O estudo da bibliometria mostra-se como um importante instrumento para identificar as particularidades de um determinado ramo da ciência, permitindo obter uma análise abrangente das diferentes áreas, demonstrando os padrões de crescimento, organização do conhecimento e fornecendo perspectivas de cenários futuros (ARVANITIS e CHATELIN, 1994) a partir da evolução ou decréscimo da pesquisa em determinado campo do conhecimento.

Destacam-se como estudos com uso dessas ferramentas a compilação de artigos sobre as tendências da ciência do solo feita por Warkentin (1994) e a análise dos 100 volumes da revista Geoderma publicados entre 1967 e 2001 por Hartemink (2001). Essa última, que identificou e sinalizou áreas do conhecimento que emergiram e outras que praticamente deixaram de ser pesquisadas – ou ao menos publicadas no periódico – nesse período. Já Minasny et al. (2013), concluíram, entre outros resultados, que a taxa de publicação média da ciência do solo é de 2,5 artigos por ano por pesquisador, sendo menor em comparação com a estudos sobre a água e a bioquímica, que possuem média de 3,1 e 3,8 artigos por ano, respectivamente.

Outro trabalho de destaque é o de Minasny et al. (2010), que avaliou as taxas de auto-citação na ciência do solo, permitindo avaliar os países e periódicos que mais citam seus próprios trabalhos. Dentre outras conclusões, os autores do trabalho puderam concluir que, com poucas exceções, as taxas de autocitação na ciência do solo são razoáveis quando comparadas à outras áreas da ciência. Isto se torna importante por permitir avaliar o comportamento de áreas da ciência, uma vez que o aumento do número de citações de um pesquisador ou um grupo de pesquisa eleva o índice h de um cientista individual ou aumenta o fator de impacto de um determinado periódico, sendo que a auto-citação não é exatamente um reflexo da disseminação do trabalho.

Estudar e avaliar como vem sendo desenvolvida a pesquisa de uma determinada área do conhecimento permite que se visualizem perspectivas futuras e tendências sobre as linhas de pesquisa dessa área. Muito vem sendo feito para que a pesquisa em MDS contribua para o desenvolvimento da Ciência do Solo no mundo, e o reconhecimento dos padrões de pesquisa, bem como as características do que foi pesquisado até aqui é fundamental para futuros trabalhos.

4.5 REFERÊNCIAS

AMUNDSON, R.; BERHE, A. A.; HOPMANS, J. W.; OLSON, C.; SZTEIN, A. E.; SPARKS, D. L. Soil and human security in the 21st century. **Science**, v. 348(6235), p. 1261071, 2015.

ARROUAYS, D. et al. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. **GeoResJ**, v. 14, p. 1-19, 2017.

ARROUAYS, D. et al. The Global Soil Map project specifications. In: ARROUAYS, D.; MCKENZIE, N.; HEMPEL, J.W.; RICHER DE FORGES, A, MCBRATNEY, AB. Eds. **Global Soil Map: Basis of the global spatial soil information system.** Taylor & Francis Group, London. p. 9-12, 2014.

ARRUDA, G. P. D.; DEMATTÊ, J. A. M.; CHAGAS, C. D. S. Digital soil mapping by artificial neural networks based on soil-landscape relationships. **Revista Brasileira de Ciência do Solo**, v. 37, n. 2, p. 327-338, 2013.

ARVANITIS, R.; CHATELIN, Y. Bibliometrics of tropical soil sciences: Some reflections and orientations. In: McDONALD, P., Ed. **The literature of soil science**. Ithaca, Cornell University Press, 1994. p. 73-94.

ARVANITIS, R.; CHATELIN, Y. Bibliometrics of Tropical Soil Sciences: Some Reflections and Orientations. In: **Stratégies Scientifiques et Développement: Sols et Agriculture des Régions Chaudes.** Orstom Editions, Paris, 1988.

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

BEHRENS, T.; SCHOLTEN, T. A comparison of data-mining techniques in predictive soil mapping. In: LAGACHERIE, P. et al. (Eds.). **Digital soil mapping, an introductory perspective**. Developments in soil science. Amsterdam: Elsevier, v.31, p. 353-364, 2007.

BERG, M. V. D.; OLIVEIRA, J. B. Variability of apparently homogeneous soilscapes in São Paulo state, Brazil: II. quality of soil maps. **Revista Brasileira de Ciência do Solo**, v. 24, n. 2, p. 393-407, 2000.

BISHOP, T. F.; MINASNY, B.; MCBRATNEY, A. B. Uncertainty analysis for soil-terrain models. **International Journal of Geographical Information Science**, v. 20, n.2, p. 117-134, 2006.

BAZAGLIA FILHO, O.; RIZZO, R.; LEPSCH, I. F.; PRADO, H. D.; GOMES, F. H.; MAZZA, J. A.; DEMATTÊ, J. A. M. Comparison between detailed digital and conventional soil maps of an area with complex geology. **Revista Brasileira de Ciência do Solo**, v. 37, n. 5, p. 1136-1148, 2013.

BREIMAN, L.; CUTLER, A. Random Forests homepage, 2009. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm (acesso em 23 de novembro de 2018).

BREVIK, E. C.; CALZOLARI, C.; MILLER, B. A.; PEREIRA, P.; KABALA, C.; BAUMGARTEN, A.; JORDÁN, A. Soil mapping, classification, and pedologic modeling: History and future directions. **Geoderma**, v. 264, p. 256-274, 2016.

CHANEY, N. W. ; WOOD, E. F. ; MCBRATNEY, A. B. ; HEMPEL, J. W. ; NAUMAN, T. W. ; BRUNGARD, C. W. ; ODGERS, N. P. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. **Geoderma**, v. 274, p. 54-67. 2016.

CONGALTON, R. G. A. Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. **Remote Sensing Environment**, v. 37, n. 1, p. 35-46, 1991.

DALMOLIN, R. S. D.; KLAMT, E.; PEDRON, F. A.; AZEVEDO, A. C. D. Relação entre as características e o uso das informações de levantamentos de solos de diferentes escalas. **Ciência Rural**, 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**, v. 17, n. 2, p. 77-86, 2015.

DONG, W.; WU, T.; SUN, Y.; LUO, J. Digital Mapping of Soil Available Phosphorus Supported by AI Technology for Precision Agriculture. In 2018 7th International Conference on Agro-geoinformatics (Agro-geoinformatics) (pp. 1-5). IEEE.

FINKE, P. A. Updating the (1: 50,000) Dutch groundwater table class map by statistical methods: an analysis of quality versus cost. **Geoderma**, v. 97, n. 3, p. 329-350, 2000.

FORKUOR, G.; HOUNKPATIN, O. K.; WELP, G.; THIEL, M. High resolution mapping of soil properties using remote sensing variables in South-Western Burkina Faso: a comparison of machine learning and multiple linear regression models. **PloS one**, v. 12, n. 1, p. e0170478, 2017.

GIASSON, E.; CLARKE, R. T.; INDA JUNIOR, A. V.; MERTEN, G. H.; TORNQUIST, C. G. Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. **Scientia Agricola**, v.63, p.262-268, 2006.

GRIMM, R.; BEHRENS, T. Uncertainty analysis of sample locations within digital soil mapping approaches. **Geoderma**, v. 155, n. 3-4, p. 154-163, 2010.

GRUNWALD, S. **Environmental soil-landscape modeling**: Geographic information technologies and pedometrics. CRC Press, 2005.

HARTEMINK, A. E. Developments and trends in soil science: 100 volumes of Geoderma (1967–2001). **Geoderma**, v. 100, p. 217-268, 2001.

HARTEMINK, A. E.; KRASILNIKOV, P.; BOCKHEIM, J. G. Soil maps of the world. **Geoderma**, v. 207, p. 256–267, 2013.

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

HEMPEL, JW et al. GlobalSoilMap project history. In: ARROUAYS, D; MCKENZIE, NJ; HEMPEL, JW; RICHER DE FORGES, A; MCBRATNEY, AB. Eds. **Global Soil Map. Basis of the Global Spatial Soil Information System**. Proceedings of the First Global Soil Map Conference. Boca Raton, CRC Press, p. 3-8, 2014.

HENGL, T.; NUSSBAUM, M.; WRIGHT, M. N.; HEUVELINK, G. B.; GRÄLER, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. **PeerJ**, v. 6, p. e5518, 2018.

HEUNG, B.; BULMER, C. E.; SCHMIDT, M. G. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. **Geoderma**, v. 214, p. 141-154, 2014.

HEUNG, B.; HO, H. C.; ZHANG, J.; KNUDBY, A.; BULMER, C. E.; SCHMIDT, M. G. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. **Geoderma**, v. 265, p. 62-77, 2016.

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**, 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**, v. 49, n. 12, p. 958-966, 2014.

HORST, T. Z.; DALMOLIN, R. S. D.; CATEN, A. T.; MOURA-BUENO, J. M.; CANCIAN, L. C.; PEDRON, F. D. A.; SCHENATO, R. B. Edaphic and Topographic Factors and their Relationship with Dendrometric Variation of Pinus Taeda L. in a High Altitude Subtropical Climate. **Revista Brasileira de Ciência do Solo**, v. 42, 2018.

HUANG, J.; MCBRATNEY, A. B.; MALONE, B. P.; FIELD, D. J. Mapping the transition from pre-European settlement to contemporary soil conditions in the Lower Hunter Valley, Australia. **Geoderma**, v. 329, p. 27-42, 2018.

IBGE. **Manual técnico de pedologia**. Coordenação de Recursos Naturais e Estudos Ambientais. 3. ed. - Rio de Janeiro, 2018 .425 p.

JENNY, H. **Factors of soil formation: a system of quantitative pedology**. Courier Corporation, 1941.

KEMPEN, B.; BRUS, D. J.; HEUVELINK, G. B.; STOORVOGEL, J. J. Updating the 1: 50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. **Geoderma**, v. 151, n. 3-4, p. 311-326, 2009.

KIDD, D.; FIELD, D.; MCBRATNEY, A.; WEBB, M. A preliminary spatial quantification of the soil security dimensions for Tasmania. **Geoderma**, v. 322, p. 184-200, 2018.

LAGACHERIE, P.; MCBRATNEY, A. B. Chapter 1: Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. **Developments in soil science**, v. 31, p. 3-22, 2007.

LAGACHERIE, P.; MCBRATNEY, A. B. Spatial information systems and spatial soil inference systems: perspectives for digital soil mapping. **Developments in soil science**, v. 31, p. 3-22, 2006.

MACHADO, I. R.; GIASSON, E.; CAMPOS, A. R.; COSTA, J. J. F.; SILVA, E. B. D.; BONFATTI, B. R. Spatial Disaggregation of Multi-Component Soil Map Units Using Legacy Data and a Tree-Based Algorithm in Southern Brazil. **Revista Brasileira de Ciência do Solo**, v. 42, 2018.

MAYNARD, J. J.; JOHNSON, M. G. Scale-dependency of LiDAR derived terrain attributes in quantitative soil-landscape modeling: Effects of grid resolution vs. neighborhood extent. **Geoderma**, v. 230, p. 29-40, 2014.

MCBRATNEY, A. B. et al. An overview of pedometric techniques for use in soil survey. **Geoderma**, v. 97, n. 3, p. 293-327, 2000.

MCBRATNEY, A.; DE GRUIJTER, J.; BRYCE, A. Pedometrics Timeline. **Geoderma**, v. 338, p. 568-575, 2019.

MCBRATNEY, A. B. et al. Digital Soil Mapping. In: HUANG, P.M.; LI, Y.; SUMNER, M.E. **Handbook of Soil Sciences: Properties and Processes**. Boca Raton: Taylor & Francis, 2012, p. 37-43.

MCBRATNEY, A.; FIELD, D. J.; KOCH, A. The dimensions of soil security. **Geoderma**, v. 213, p. 203-213, 2014.

MCBRATNEY, A. B.; MENDONÇA-SANTOS, M.L.; MINASNY, B. On Digital Soil Mapping. **Geoderma**, v. 117, n. 1, p. 3-52, 2003.

MENDONÇA-SANTOS, M. L.; SANTOS, H. G. The state of the art of Brazilian soil mapping and prospects for digital soil mapping. In: LAGACHERIE, P.; McBRATNEY, A.B.; VOLTZ, M. (Ed.). **Digital Soil Mapping: an introductory perspective**. Amsterdam: Elsevier, 2007. p. 39–54.

MERTON, R. K. The Matthew effect in science. **Science**, v. 159, p. 56-63, 1968.

MINASNY, B.; BISHOP, T. Analysing uncertainty. In: MCKENZIE, N. J.; GRUNDY, M. J.; WEBSTER, R.; RINGROSE-VOASE, A. J. (Eds.), **Guidelines for Surveying soil and Land Resources**, p. 383-393. Australia: CSIRO Publishing, 2008.

MINASNY, B.; HARTEMINK, A. E.; MCBRATNEY, A. Individual, country, and journal self-citation in soil science. **Geoderma**, v. 155, n. 3-4, p. 434-438, 2010.

MINASNY, B.; HARTEMINK, A. E.; MCBRATNEY, A.; JANG, H. J. Citations and the h index of soil researchers and journals in the Web of Science, Scopus, and Google Scholar. **PeerJ**, v. 1, p. e183, 2013.

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

MINASNY, B.; MCBRATNEY, A. B. Methodologies for Global Soil Mapping. In: BOETTINGER, et al. (Ed.). **Digital Soil Mapping: Bridging Research, Environmental Application and Operation**. London: Springer, 2010, p. 429–436.

NELSON, M. A.; BISHOP, T. F. A.; TRIANTAFILIS, J.; ODEH, I. O. A. An error budget for different sources of error in digital soil mapping. **European Journal of Soil Science**, v. 62, n. 3, p. 417-430, 2011.

NORONHA, D. P.; MARICATO, J. M. Estudos métricos da informação: primeiras aproximações. **Revista eletrônica de biblioteconomia e ciência da informação**, v. 13, n. 1, p. 116-128, 2008.

ODGERS, N. P. et al. Disaggregating and harmonising soil map units through resampled classification trees. **Geoderma**, v. 214, p. 91-100, 2014a.

ODGERS, N. P.; McBRATNEY, A. B.; MINASNY, B. Digital soil property mapping and uncertainty estimation using soilclass probability rasters. In: ARROUAYS, D. et al. (Ed.). **GlobalSoilMap: Basis of the global spatial soil information system**. London: Taylor & Francis, p. 341-346, 2014b.

OMUTO, C.; NACHTERGAELE, F.; ROJAS, R. V. **State of the art report on global and regional soil information: Where are we? Where to go?** Global Soil Parthership Technical Report. Roma: FAO, 2013. 69p.

PAHLAVAN RAD, M. R.; TOOMANIAN, N.; KHORMALI, F.; BRUNGARD, C. W.; KOMAKI, C. B.; BOGAERT, P. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. **Geoderma**, v. 232, p. 97-106, 2014.

PINHEIRO, H. S. K.; CARVALHO JUNIOR, W. D.; CHAGAS, C. D. S.; ANJOS, L. H. C. D.; OWENS, P. R. Prediction of Topsoil Texture Through Regression Trees and Multiple Linear Regressions. **Revista Brasileira de Ciência do Solo**, v. 42, 2018.

ROSSITER, D. G. Digital soil resource inventories: status and prospects. **Soil Use Management**, v. 20, p. 296-301, 2004.

SAMUEL-ROSA A et al. 2018. Bringing together Brazilian soil scientists to share soil data. In: 21st World Congress of Soil Science, Rio de Janeiro. **Soil science: beyond food and fuel**, 2 p.

SAMUEL-ROSA, A.; HEUVELINK, G. B. M.; VASQUES, G. M.; ANJOS, L. H. C. Do more detailed environmental covariates deliver more accurate soil maps? **Geoderma**, v. 243, p. 214-227, 2015.

SANCHEZ, P. A. et al. Digital soil map of the world. **Science**, v.325, n. 5941, p. 680-681, 2009.

SARMENTO, E. C.; GIASSON, E.; WEBER, E. J.; FLORES, C. A.; HASENACK, H. Disaggregating conventional soil maps with limited descriptive data: a knowledge-based approach in Serra Gaúcha, Brazil. **Geoderma Regional**, v. 8, p. 12-23, 2017.

SILVA, S. H. G. et al. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. **Geoderma**, v. 267, p. 65-77, 2016.

STUMPF, F.; SCHMIDT, K.; GOEBES, P.; BEHRENS, T.; SCHÖNBRODT-STITT, S.; WADOUX, A.; SCHOLTEN, T. Uncertainty-guided sampling to improve digital soil maps. **Catena**, v. 153, p. 30-38, 2017.

SUBBURAYALU, S. K.,;SLATER, B. K. Soil series mapping by knowledge discovery from an Ohio county soil map. **Soil Science Society of America Journal**, v. 77, n. 4, p. 1254-1268, 2013.

SULAEMAN, Y. et al. Harmonizing legacy soil data for digital soil mapping in Indonesia. **Geoderma**, v. 192, p. 77-85, 2013.

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

TEN CATEN, A. **Mapeamento digital de solos: Metodologias para atender a demanda por informação espacial em solos**. 2011. 108 p. Tese (Doutorado em Ciência do Solo) Universidade Federal de Santa Maria, Santa Maria, RS.

TEN CATEN, A.; DALMOLIN, R. S. D.; PEDRON, F. A.; MENDONÇA SANTOS, M. D. L. Componentes principais como preditores no mapeamento digital de classes de solos. **Ciência Rural**, v. 41, n. 7, p. 1170-1176, 2011b.

TEN CATEN, A.; DALMOLIN, R. S. D.; MENDONÇA-SANTOS, M. D. L.; GIASSON, E.. Mapeamento digital de classes de solos: características da abordagem brasileira. Revista **Ciência Rural**, v. 42, n. 11, p. 1989-1997, 2012.

VAN ZIJL, G. M. et al. Functional digital soil mapping: A case study from Namarroi, Mozambique. Geoderma, v. 219, p.155-161, 2014.

VILLELA, A.L.O. **Mapeamento Digital de Solos da Formação Solimões Sob Floresta Tropical Amazônica**. 2013. 114 f. Tese (Doutorado em Ciências). Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ. 2013.

VISCARRA ROSSEL, R.; MCBRATNEY, A. B. Soil chemical analytical accuracy and costs: implications from precision agriculture. **Australian Journal of Experimental Agriculture**, v. 38, n. 7, P. 765-775, 1998.

WARKENTIN, B. P. Trends and developments in soil science. In: MCDONALD, P. (Ed.). **The literature of soil science**. 1994. Cornell Univ, 1 ed., Hardcover: 448 p.

WYSOCKI, D. A. et al. Soil Landscape Models. In: HUANG, P.M.; LI, Y.; SUMNER, M.E. **Handbook of Soil Sciences: properties and processes**. Boca Raton: Taylor & Francis, 2012. p. 294–298.

WOLSKI, M. S.; DALMOLIN, R. S. D.; FLORES, C. A.; MOURA-BUENO, J. M.; TEN CATEN, A.; KAISER, D. R. Digital soil mapping and its implications in the extrapolation of soil-landscape relationships in detailed scale. **Pesquisa Agropecuária Brasileira**, v. 52, n. 8, p. 633-642, 2017.

ZHANG, G. L.; FENG, L.; SONG, X. D. Recent progress and future prospect of digital soil mapping: A review. **Journal of integrative agriculture**, v. 16, n. 12, p. 2871-2885, 2017.

ZERAATPISHEH, M.; AYOUBIA, S.; BRUNGARD, CW; FINKE, P. Disaggregating and updating a legacy soil map using DSMART, fuzzy c-means and k-means clustering algorithms in Central Iran. **Geoderma**, v. 340, p. 249-258, 2019.

5 CAPÍTULO 2 – BIBLIOMETRIC ANALYSIS FOR PATTERN EXPLORATION IN WOLDWIDE DIGITAL SOIL MAPPING PUBLICATIONS

Anais da Academia Brasileira de Ciências (2018)
(Annals of the Brazilian Academy of Sciences)
Printed version ISSN 0001-3765 / Online version ISSN 1678-2690
http://dx.doi.org/10.1590/0001-3765201820180423
www.scielo.br/aabc | www.fb.com/aabcjournal

Bibliometric Analysis for Pattern Exploration in Worldwide Digital Soil Mapping Publications

LUCIANO C. CANCIAN[1], RICARDO S.D. DALMOLIN[1] and ALEXANDRE T. CATEN[2]

[1]Departamento de Solos, Centro de Ciências Rurais, Universidade Federal de Santa Maria, Avenida Roraima, 1000, 97105-900 Santa Maria, RS, Brazil
[2]Departamento de Agricultura, Biodiversidade e Florestas, Universidade Federal de Santa Catarina, Rodovia Ulysses Gaboardi, Km 3, 89520-000 Curitibanos, SC, Brazil

Manuscript received on September 19, 2017; accepted for publication on July 19, 2018

ABSTRACT

Bibliometric analyses provide a clear understanding of the scientific performance and relate them with standards of the global scientific production. Soil science is an outstanding and developing field among environmental sciences. Knowledge about soil characteristics and their distribution in the environment has been enriched by the use of new geotechnologies, resulting in what is known as digital soil mapping. Thus, the objective of this work was to characterize the scientific production in digital soil mapping in Brazil and in the world, in the period from 1996 to 2017, in databases such as Scopus and Web of Science. In the general context of increasing numbers of papers, the journal Geoderma published the highest number of related papers. Among the 10 with most published papers, the Revista Brasileira de Ciência do Solo is the only open access journal. Although there are countries at the cutting edge of digital soil mapping such as the United States and Australia, the position of Brazil in the number of papers and authors cannot be overlooked, showing the importance of the nation's participation in digital soil mapping, as a field of science that can provide guidelines for public policies for the development of agriculture in the country.

Key words: pedometrics, soil science, soil meta-analysis, soil survey.

INTRODUCTION

One way to assess the strength and productivity of a scientific field is to measure the number of publications over time (Hartemink and McBratney 2008, Mao et al. 2015). Bibliometry arose directly from this need of evaluating scientific production, in view of the large amounts of information available in bibliometric databases (Wallin 2005, Moed 2009, Loudcher et al. 2015). Some bibliometric indices have been used for a strategic planning of research by institutions, universities and research funding agencies (Zhou et al. 2016). A clear understanding of the institutional performance does not only support particular areas of research but also situates them in relation to global scientific production standards. In view thereof, the most productive and influential researchers and countries should be investigated (Cancino et al. 2017), also helping researchers to identify the leading journals in the development of a particular field of science (Shokraneh et al. 2012).

Correspondence to: Ricardo Simão Diniz Dalmolin
E-mail: dalmolin@ufsm.br

In this aspect, soil scientists have increasingly contributed for the myriad of publications.

The importance of soils for ecosystems, food production, and climate regulation is more and more viewed as fundamental (Sanchez et al. 2009, Amundson et al. 2015). A growing interest in agriculture has also put soil back on the global research agenda. The increasing need for up-to-date information on soil has been highlighted in several recent studies of the United Nations and other international organizations (Robinson et al. 2017). Soil science is a knowledge area that can help find answers to these challenges (McBratney et al. 2014). Nevertheless, more could be achieved through meta-analysis of what has already been published (Roudier et al. 2015, Arrouays et al. 2017).

Bibliometric studies are not new in soil science. Based on bibliometric tools and a study review, Warketin (1994) sought for trends in soil science studies, underlying the description and determination of the evolution and main topics addressed in this field. Another example is the review of the 100 volumes published by Geoderma, between 1967 and 2001 (Hartemink 2001), evaluating the temporal behavior and characteristics of these publications in one of the main journals of soil science, showing the development of subareas of soil science over the years. Studies as that of Hartemink (2015), with inferences on how the new generation of soil scientists has been using soil classification, demonstrate the importance of bibliometric studies for science. They equip researchers, especially the new generation and young researchers, with an outlook on the headway already made on a given topic, identifying the difficulties, peculiarities and thus, finding ways for its evolution.

In the search for an enhanced acquisition of soil information, digital soil mapping (DSM) emerged by integrating subareas as a technique to generate new soil studies and meet the demand for information in terms of detailed knowledge on spatial distribution and properties (Arrouays et al. 2017). DSM is benefitted by the increasing availability of spatial data of the Earth's surface (McBratney et al. 2012), and is a constantly increasing field of science, in which additional possibilities of applications are continuously being explored. Based on a search in the *Scopus database* for articles containing the keywords "digital soil mapping", Minasny and McBratney (2016) stated that publications in the area increased at a rate of 12 papers per year and the number of citations increased by 384 citations per year. Much effort has been invested so that research on DSM will contribute to the further development of soil science in the world.

In Brazil, with a huge territorial extension available for food and fiber production, there is a lack of information about sustainable land use that supports production activities. As the soil databases available in the country do not cover the entire territory, DSM would contribute in a practical way to complete this information. The application of DSM is a relatively new area of science in Brazil (ten Caten et al. 2012, Dalmolin and ten Caten 2015), and the first paper on DSM in Brazilian territory was published only in 2006 (Giasson et al. 2006). However, there is no information on the amount of the scientific production or how many researchers work with DSM in Brazil. This information would be useful for the specific orientation of public policies for compiling soil information, as of the program "Pronasolos" (Polidoro et al. 2016) for example, dedicated to resume pedological surveys in Brazil, for which DSM could be useful.

In this regard, there is still a lack of studies that characterize the main publications, not only in Brazil, but in other countries as well. Based on a comparison of the production of the main authors and their respective countries and between the DSM studies developed in Brazil and the global trends, the national and global research characteristics

in DSM research could be identified, aside from predicting future scenarios and indicating new research lines. In this context, the objective of this study was to characterize, based on a set of bibliometric indicators, the scientific production on DSM for Brazil and worldwide, between 1996 and 2017, to identify characteristics and peculiarities in the national and global scientific production on DSM, making the prediction of growth trends in this area of knowledge possible and indicating paths to be followed.

MATERIALS AND METHODS

DATA ORIGIN AND SEARCH PROCEDURE

For a pre-analysis of the databases, data were obtained from the Clarivate Analitics Web of Science (WoS) and Scopus databases. All records characterized by articles and bibliographic reviews detected by a query of the subject areas connected to agrarian sciences, published between September 1996 and December 2017, were stored and included in the study.

From combinations of terms referring to DSM, queries were carried out including searches for terms in the titles, abstracts, and keywords of papers. The words and terms used for this study were previously tested in the main scientific journals of the area. For our analysis, they were limited to those that were most relevant and published only in articles specifically about DSM. With this limitation by the search for terms only in titles, abstracts and keywords, articles from other areas with complementary methodologies or only cite of DSM where not taken into consideration. If the amount of results were too large, it would be necessary to check if the publications found fit in any topic of the research area, making the search subjective and not automatic. It is worth mentioning that unless some of the query terms were found in the keywords, title, or abstract, the papers were not included even if they used the term DSM, and some papers may not appear, due to this methodology. The database was searched for the following terms: “digital soil map” OR “digital soil mapping” OR “digital map of soil” OR “digital mapping of soil” OR “GlobalSoilMap”. The plural terms were also included in the search. In both databases, the search was performed in the “Advanced search” field, making use of the boolean operator “OR” between the terms. After, the results were filtered, limiting the years from 1996 to 2017.

From the publications found by this method, all basic information was extracted, including the author’s names, affiliation, country, language of publication, type of document (article or bibliographic review), number of times it was cited, journal name, year of publication, keywords, and subject category. Thereafter the data were saved in BibTex format, as recommended by Aria and Cuccurullo (2018).

BIBLIOMETRIC INDICES

The bibliometric analysis of the complete search results was performed using the package Bibliometrix (Aria and Cuccurullo 2018) version 1.9, in R environment (R Core Team 2018). The two files in BibTex format were uploaded, the “readFiles” function applied and converted to a data frame. Some 670 files were obtained from Scopus and 557 from WoS, consisting of journal articles and bibliographic reviews. After merging the two files, duplicated records were eliminated by the function “remove.duplicated”. Then, to avoid any language conflict that would make an inclusion of duplicate documents possible, a manual screening was performed, removing the duplicate files. The total of 1,227 (Scopus + WoS) files was reduced to a final number of 727 files to be analyzed.

First, the “biblioAnalysis” function was applied, returning an object of class “bibliometrix”, to which the “biblioNetwork” function is applied, which generates a set of bibliometric indices

together with the results from the two databases. With this tool, the annual results of publications and citations were classified as DSM-related. In addition, a keyword network of all papers included in the study was created, allowing to analyze the subject trends in DSM research.

The scientific journals were evaluated as follows: identification of the 10 journals with the highest number of publications on DSM, also by function "biblioAnalysis". The output of papers on DSM of the 10 most productive journals was recorded for the study period. Complementarily, as a way of evaluating the journal quality, the following impact factors were analyzed: SCImago Journal Rank (SJR), a factor generated by Scopus, and the Journal Citation Reports (JCR), published by the Institute for Scientific Information. The JCR is a recognized base for evaluating journals indexed in the WoS. As a complementary metric of the journals with most publications on DSM, the Eigenfactor Score was calculated, considered adequate to mediate the quality of these journals (Cantín et al. 2015). The Eigenfactor is calculated from the number of times the articles published in the journal were cited by the JCR in the last five years, similarly to the Journal Impact Factor (JIF) or the 5-year JIF. In addition, the Eigenfactor assigns a weight or value to each citation, according to the journal of citation.

The advantage of the bibliometrix package is that more than one database can be analyzed, however, only the country of the first author is taken into consideration for the calculation. Therefore, since this study investigated in more detail in which countries the DSM research was carried out, we used only the data obtained by the Scopus database for this purpose, allowing the countries of all authors to be counted. This database was preferred to the others included in the previous analyses due to the higher number of articles found in the Scopus database (670 papers), the highest number of represented journals (Chadegani et al. 2013), the highest number of journals published only in this database (Barnett and Lascar 2012), and the highest number of papers on the subject "soil" (Minasny et al. 2013).

To evaluate the scientific output of a country, papers were counted for the entire study period from 1996 to 2017. For the number of citations, on the other hand, due to a limitation of the maximum time period of the database, only those of the period from 2002 to 2017 were counted, coinciding with the largest expansion of DSM research. The percentages of self-citation of each country were also recorded. A country self-citation means the percentage of citations received by papers from the same country in which the papers were published.

To evaluate the geographical distribution of the authors, the institutional addresses cited in the studies were captured in the Scopus database and their geographic coordinates recorded. To assess the increase of research in the countries over the years, this information was captured for the period from 1996 to 2007 and then from 1996 to 2017, in order to identify how the growth occurred in the first years of research and after the boost in publications on DSM. Maps were created with software QGIS 2.18 (Qgis Development Team et al. 2018) with the point-in-polygon tool, considering the number of authors of each country in the evaluated periods. The data of the institutional address indicated by the authors were also plotted on the map, showing a geographical distribution of authors working with DSM. In addition, the most representative Brazilian institutions were mapped.

RESULTS AND DISCUSSION

CHARACTERISTICS OF THE GLOBAL DSM RESEARCH

From 727 papers retrieved in the Scopus and WoS database, we observed an increased number of publications on DSM at an annual rate of 19.6%, with an average of 15.4 citations per paper (Figure

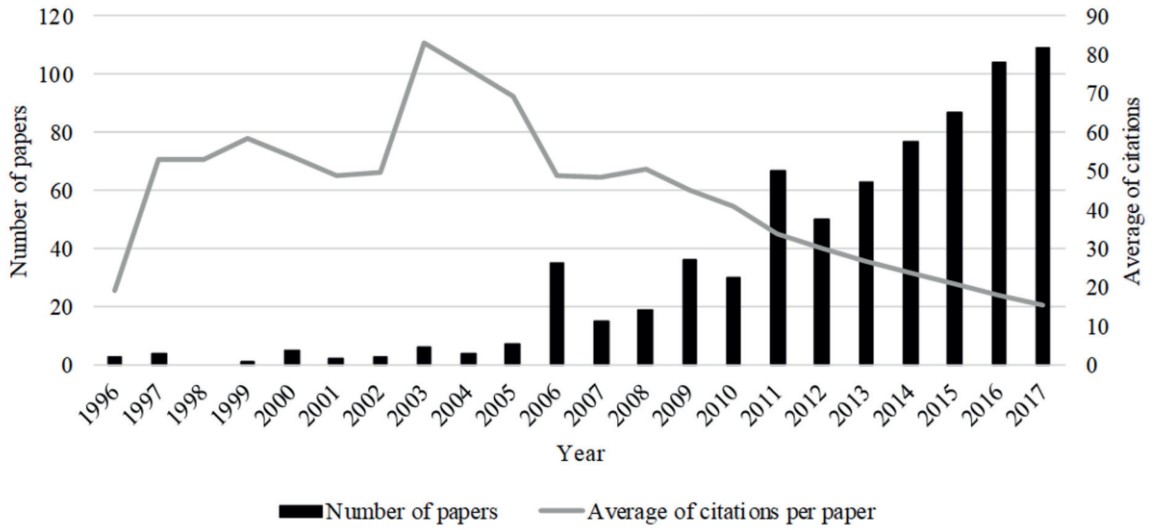


Figure 1 - Evolution of the number of papers and average citations per paper on DSM from 1996 to 2017.

1). Except in 1998, at least one publication was released per year, and a remarkable increase in the number of published papers was observed after 2006. A higher number of publications in 2006 was also observed, with discrepant values in relation to the growth trend observed until then. This higher number of papers can be explained by the occurrence of the event Second Global Workshop on Digital Soil Mapping, in 2006. The annual publication increased steadily from 13 in 2007 to 100 in 2017, i.e., a 10-fold increase in those 10 years.

The curve of citations per paper shows growth in the first years, becoming more pronounced in 2003, the year of publication of the paper "On digital soil mapping" (McBratney et al. 2003), which established the bases and defined concepts of DSM, with 950 citations, serving as a reference for several subsequent studies. Thereafter, the curve declined, and the citations per paper rate was diluted by the higher number of papers published until 2017. On this decrease, it should be noted that more recent articles are in the so-called citation window, which is the time needed for the paper to be read and subsequently quoted.

The great majority of papers investigated here deal with the application of DSM techniques in different regions of the world. The increase in data availability can break boundaries and leverage the coverage and availability of soil maps and properties constructed by DSM (Omuto et al. 2013, Sulaeman et al. 2013). A considerable portion of the most cited papers (data not shown) addresses the evaluation of different models and techniques, with a view to evaluating methodologies that are more adequate for the prediction of soil properties. Confirming the statements of Arrouays et al. (2017), there is also an increasing tendency to carry out studies on a local or regional scale. This contributes to a higher number of citations because the newly discovered methodologies can be applied in regions with similar characteristics. However, studies on a regional scale not only contribute to a higher number of citations, but also allow a flow of the already discovered and tested knowledge, to be disseminated and applied in different regions of the world, generating new information about soils.

Taking into consideration all papers published in the last seven years, 80 studies were found where legacy data was exploited or suggesting its use as a way to feed the demand for DSM input data.

Legacy data are also mentioned by Zhang et al. (2017) as one of the trends in the development of DSM and to obtain soil data. Legacy data are data of former soil surveys or previous studies, usually by the so-called "conventional method" (Omuto et al. 2013), and can be used as input data for the application of DSM methodologies (Odgers et al. 2014). This shows that one of the possible limiting factors, with regard to the scope of the papers, is the availability of samples for the training and validation of the generated prediction models. This is due to the need to consider a large amount of samples, limiting DSM research in view of the time and financial resources required for sample collection and analysis.

As the DSM addresses diverse soil information, an overview of the main topics covered in the papers helps identify the most frequent themes dealt with. Figure 2 shows the 15 most frequent keywords in the DSM papers in the survey period, demonstrating that the central theme is the use of "models" for "spatial prediction", mainly "soil organic carbon", soil class mapping (from soil "landscape" relation) or "land use" of the soil.

The keywords also draws attention to the issues of obtaining "information" at appropriate "scales" in the studies, as reported by Arrouays et al. (2017), performing the "digital mapping" of the soil, allowing the formation of "databases" or using information from them.

PERIODIC EVALUATION OF DSM-RELATED PUBLICATIONS

The variability of journals publishing DSM-related articles is wide. A total of 727 papers was published by 171 journals, and approximately 44% were released in 10 journals (Figure 3). Geoderma published most papers on the subject (159, or 22% of the total number of publications), followed by the Soil Science Society of America Journal (28, or 3.8% of the total) and by the Revista Brasileira de Ciência do Solo (RBCS), with 26, or 3.6% of the total.

Representing not only the high number of publications on DSM, but also the representativeness and relevance of journals for soil science, the indices JCR and SJR reflect the importance and tradition of these journals adequately, with emphasis on

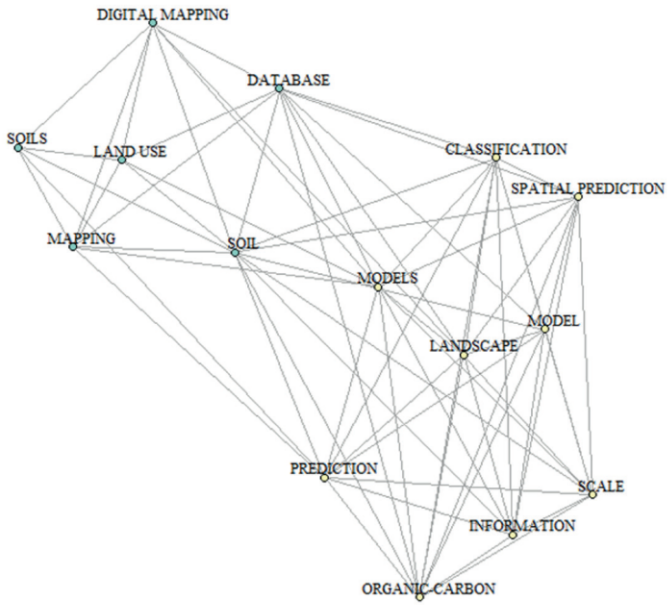


Figure 2 - The 15 keywords with highest frequency in DSM papers.

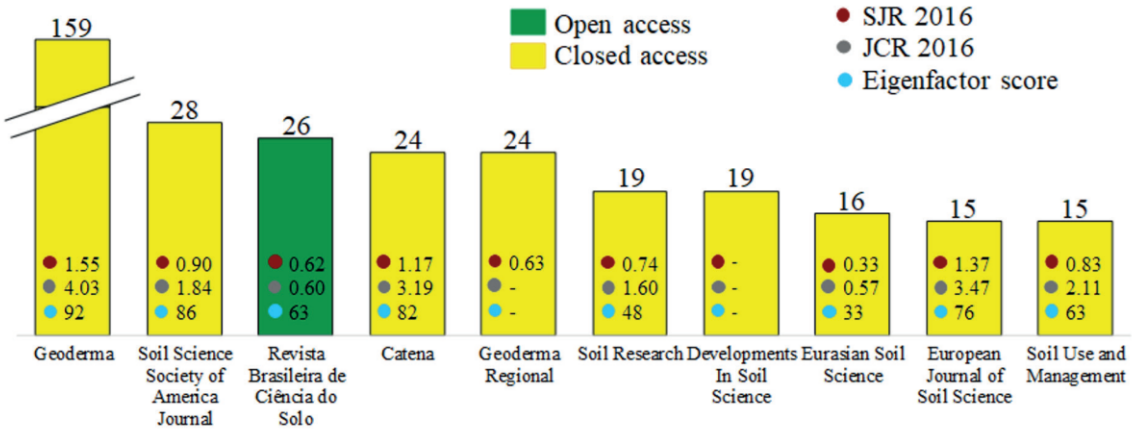


Figure 3 - Comparison of the 10 journals with highest output of DSM-related publications between 1996 and 2017.

Geoderma as one of the main journals in soil science (SJR 1.55, JCR 4.03) (Hartemink 2001). The only journals not represented in these indices are Developments in Soil Science, which was discontinued in 2010, and Geoderma Regional, which is a new journal with an insufficient publication period to calculate the indices.

In spite of the late start of research on DSM in Brazil, the RBCS, the country's leading soil science journal, ranks fifth among the journals with most DSM publications, indicating an increase in DSM research in Brazil. With a considerably higher number of papers than other journals, the RBCS is also the only one in the top 10 with open access to all publications. This is an important aspect of research, since this publication format ensures a more efficient distribution of scientific knowledge than the standard publication model (Martínez-Quintana and Penagos-Corzo 2012).

The restricted availability of the vast majority of articles in closed access journals possibly affects the availability of knowledge to the scientific community, especially in developing countries. In Brazil, the scientific and educational institutions have free access to the largest databases. However, papers provided by scientists and researchers through social networking sites such as Research Gate have a noteworthy influence, facilitating the flow of scientific information (Thelwall and Kousha 2014), increasing the chances of citations and, consequently, raising the bibliometric indices.

Among soil science journals, the RBCS has an outstanding position (Minasny et al. 2013). Nevertheless, RBCS also has one of the lowest impact factors among the top 10 in DSM. In spite of its longstanding tradition in soil science in Brazil, indices are still low, possibly due to the fact that the great majority of articles were published in Portuguese until 2013, when English became the only language of publication (Vargas et al. 2014), facilitating the reading and citation of articles by the international scientific community. However, an analysis of the Eigenfactor Score (63) shows an approximation of the RBCS to other journals with better classification in the other two metrics studied, SJR and JCR, with even higher scores than some other journals. This means that despite the low impact factor, RBCS is cited in influential articles and journals.

EVALUATION OF THE COUNTRIES AND THE POSSIBILITIES OF BRAZIL

The evolution of DSM research in the last years is evident. However, not only the number of

publications increased, but also the number of authors involved with the topic from more countries. Not only this increase was observed, but also the continuity of the authors involved in the first studies on DSM (Figure 1). From 1996 to 2007, the highest number of publishing authors were concentrated in the United States, followed by China. The distribution from 1996 to 2017 shows that these countries continue in evidence, beside the emergence of countries in Europe. Aside from Australia and Brazil followed by Netherlands, France and Germany, which are already very well-represented by the number of authors, it is worth mentioning the increasing participation of Iran, noted as outstanding in the DSM world scenario. Also noteworthy is the large dissemination of authors in the United States, where, in addition to having a high concentration of authors per area from 1996 to 2017, there is also a good distribution of authors across the country. This may be a result of the great efforts of the United States to harmonize and optimize the use of data and soil maps (Thompson et al. 2012) and get a comprehensive coverage of the country's agricultural land (Lobry de Bruyn et al. 2017).

These countries, whether at the forefront of DSM research or through collaboration in studies in other countries, are directly involved in the development of DMS tools and technical applications. Even though to a lesser extent, several countries in the different continents have a significant concentration of authors, while other countries participated with the publication of at least one paper during the survey period. On the other hand, the absence of DSM researchers in several African countries was noted in both survey periods, since although the soil of a good part of the territory with different properties is already mapped (Hengl et al. 2015), few countries have researchers working in institutions of the continent.

Observing the 10 countries with the highest output of DSM papers (Table I), the information shown in Figure 4 is confirmed. The United States is represented by 133 papers, closely followed by Australia, with 124. In addition to this ample dissemination in the United States, Australia stands out not only for dissemination of research, but also of tools for end users (Minasny and McBratney 2016) because it has a qualified research team that has been making great efforts in the development of new techniques, highlighting this country in this research area.

Also noteworthy are France and The Netherlands, which, even with a lower number of papers, had mean citations per paper of 34.7 and 31.6, respectively. The lowest citation means were found for Germany and China, which have a larger production of papers, together with Belgium and Canada (17.2, 16.7, 15.9, and 14.3, respectively). When the self-citation of the top 10 countries in DSM was evaluated, Belgium and China obtained the highest percentages (26.9% and 25.7%, respectively). In comparison, Minasny et al. 2013, studying soil science journals, found an average of 12% self-citations. On the other hand, Minasny et al. (2010) also reported that in soil science publications, the countries with the highest self-citation percentages are China (63%) and the United States (48%), exceeding the values found for DSM publications. The study of self-citations also sheds light on the scientific output of a country, since the more papers a country produces, the more likely it is to cite articles from its own nation, whereas countries with fewer researchers and less published articles are more likely to cite papers from other nations (Minasny et al. 2010).

Evaluating the scientific production in Brazil, an important number of papers and citations was found. Seventy-nine papers were published in the analyzed period, with an average of 24.3 citations per paper, exceeding the number of the United States and China. Despite the delay in starting the application of DSM at the national level which may be related to the later access to software

TABLE I
Comparison of the 10 countries with most published DMS papers between 1996 and 2017 and citations, mean citations per paper and self-citations between 2002 and 2017.

Country	Papers	Citations 2002-2017	Mean citations 2002-2017	Self-citations (%) 2002-2017
United States	133	2758	20.7	22.5
Australia	124	4002	32.3	18.8
China	93	1555	16.7	25.7
Germany	82	1414	17.2	22.3
Brazil	79	1923	24.3	18.1
Netherlands	46	1453	31.6	17.0
France	41	1424	34.7	19.3
Belgium	34	542	15.9	26.9
United Kingdom	33	769	23.3	15.1
Canada	28	401	14.3	22.7

Citations: citations of papers of the proper country, also called self-citations, from 2002 to 2017.

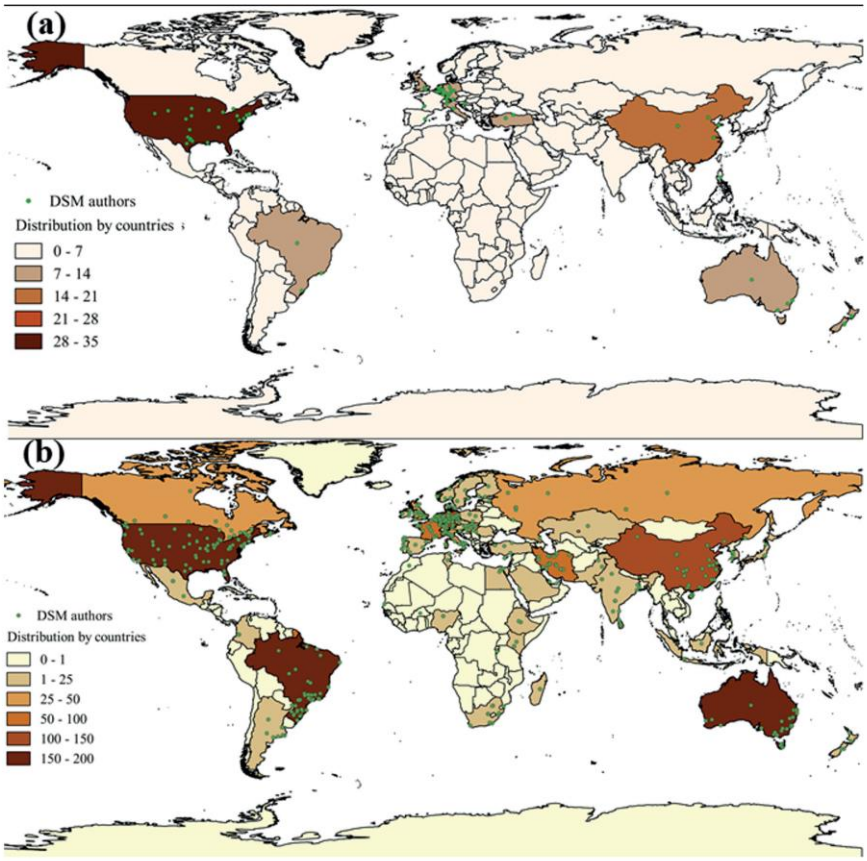


Figure 4 - Geographical distribution of authors of papers on DSM, **(a)** first decade of analysis from 1996 to 2007 and **(b)** second decade of analysis from 1996 to 2017.

and hardware technology in the country, the conservatism of many pedologists in migrating to more advanced techniques or to the lack of qualified researchers in applying the new techniques (ten Caten et al. 2012), Brazil plays a prominent role in the world scene of publications on DSM.

The good performance of DSM in Brazil is also an expression of the level of Brazilian scientific production, ranking among the world's top 25 countries in scientific quality and first in South America (Nature Index 2017), and of the evident growth of soil science in the country (Trajano et al. 2013). This is a sign of the potential and ability not only to leverage scientific production even more, but also that research can be applied to generate knowledge and information in the country.

Figure 5 shows the distribution of Brazilian institutions investing in DSM research. Most of these are located in the south and southeast of the country, e.g.: agency of the Empresa Brasileira de Pesquisa Agropecuária specialized in soil research, Embrapa Solos, which accounts for 20 papers, followed by the Universidade Federal de Santa Maria, with 14 publications and ESALQ - Universidade de São Paulo, with 12 papers. Other institutions also made significant contributions, such as the Universidade Federal do Rio Grande do Sul (7), Universidade Federal de Lavras (7) and Universidade Federal de Santa Catarina (8).

Considering that there are already at least 200 researchers (data not shown) working directly or indirectly with DSM in Brazil, the creation

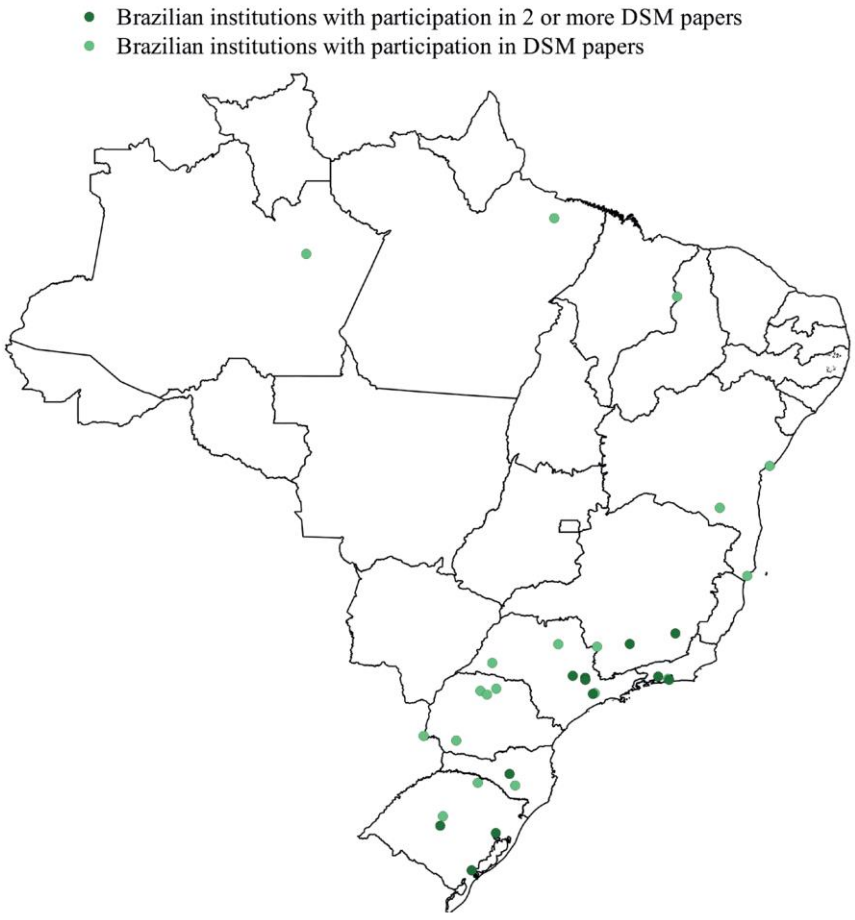


Figure 5 - Brazilian institutions with production of DSM papers.

and application of public policies, programs and research projects is fundamental, since this that may not only leverage Brazilian journals, but will increase the recognition and qualification of Brazilian researchers in scientific and social aspects. In Brazil, the implementation of projects such as SSURGO in the United States, where DSM is understood as a tool to map the country's entire arable land (Chaney et al. 2016), could supply the demand for information on Brazilian soils.

In this regard, Brazil already has fruitful initiatives like the Free Brazilian Repository for Open Soil Data (RBLDAS) (www.ufsm.br/febr), an unprecedented initiative that allows soil scientists to publish their datasets. The RBLDAS aims to centralize storage and allows the sharing of all types of soil information in Brazil. In this way, it is possible for soil legacy data to be used in other studies, also increasing collaboration among soil scientists (Samuel-Rosa et al. 2018). This knowledge, which has already been created and development is underway, could help significantly in programs such as "Pronasolos" (Polidoro et al. 2016), a long-term project to obtain information on soil in Brazil. In order to map the entire national territory, the possibility of using sophisticated techniques for high-precision, fine-resolution modelling of soil properties (Zhang et al. 2017) can be considered a renaissance of pedology in Brazil.

CONCLUSIONS

Publications on DSM are increasing at an accelerated pace, with the most significant contributions coming from Australia, the United States, China, Germany, and Brazil. The vast majority of articles was published in Geoderma, but other journals have also been achieving notable success.

The DSM research in Brazil has been gaining a prominent position in the world scenario, not only in the number of papers, but also with good quotation. From the knowledge already generated and the apparent evolution of DSM in Brazil, public policies and financial support could contribute not only to Brazilian research, but also to the social and technological development of the country by participation in programs to obtain soil information in the country.

ACKNOWLEDGMENTS

The first author would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the scholarship, and the second author is indebted to the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the scholarship and financial support.

REFERENCES

AMUNDSON R, BERHE AA, HOPMANS JW, OLSON C, SZTEIN AE AND SPARKS DL. 2015. Soil and human security in the 21st century. Science 348(6235): 1261071.

ARIA M AND CUCCURULLO C. 2018. Bibliometric and Co-Citation Analysis Tool. Bibliometrix. Available at: https://cran.r-project.org/web/packages/bibliometrix/bibliometrix.pdf. Accessed on February 15, 2018.

ARROUAYS D, LAGACHERIE P AND HARTEMINK AE. 2017. Digital soil mapping across the globe. Geod Region 9: 1-4.

BARNETT P AND LASCAR C. 2012. Comparing unique title coverage of Web of Science and Scopus in Earth and atmospheric sciences. Issues Sci Technol Librariansh 70: 1-20.

CANCINO CA, MERIGÓ JM AND CORONADO FC. 2017. A bibliometric analysis of leading universities in innovation research. Journ Innov Knowl 2(3): 106-124.

CANTÍN M, MUÑOZ M AND ROA I. 2015. Comparison between Impact Factor, Eigenfactor Score, and SCImago Journal Rank Indicator in Anatomy and Morphology Journals. Int J Morph 33(3): 1183-1188.

CHADEGANI AA, SALEHI H, YUNUS M, FARHADI H, FOOLADI M, FARHADI M AND ALE EBRAHIM N. 2013. A comparison between two main academic literature collections: Web of Science and Scopus databases. As Soc Sci 9(5): 18-26.

CHANEY NW, WOOD EF, MCBRATNEY AB, HEMPEL JW, NAUMAN TW, BRUNGARD CW AND ODGERS NP. 2016. POLARIS: A 30-meter probabilistic soil series map of the contiguous United States. Geoderma 274: 54-67.

DALMOLIN RSD AND TEN CATEN A. 2015. Mapeamento Digital: nova abordagem em levantamento de solos. Invest Agrar 17(2): 77-86.

GIASSON E, CLARKE RT, INDA JUNIOR AV, MERTEN GH AND TORNQUIST CG. 2006. Digital soil mapping using multiple logistic regression on terrain parameters in Southern Brazil. Sci Agric 63(3): 262-268.

HARTEMINK AE. 2001. Developments and trends in soil science: 100 volumes of Geoderma (1967–2001). Geoderma 100: 217-268.

HARTEMINK AE. 2015. The use of soil classification in journal papers between 1975 and 2014. Geod Region 5: 127-139.

HARTEMINK AE AND MCBRATNEY AB. 2008. A soil science renaissance. Geoderma 148(2): 123-129.

HENGL T ET AL. 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. PLoS ONE 10(6): e0125814.

LOBRY DE BRUYN L, JENKINS A AND SAMSON-LIEBIG S. 2017. Lessons Learnt: sharing soil knowledge to improve land management and sustainable soil use. Soil Sci Soc Am J 81(3): 427-438.

LOUDCHER S, JAKAWAT W, MORALES EPS AND FAVRE C. 2015. Combining OLAP and information networks for bibliographic data analysis: a survey. Scientometrics 103(2): 471-487.

MAO G, LIU X, DU H, ZUO J AND WANG L. 2015. Way forward for alternative energy research: A bibliometric analysis during 1994–2013. Ren Sust Ene 48: 276-286.

MARTÍNEZ-QUINTANA MU AND PENAGOS-CORZO JC. 2012. Open access in the dissemination of scientific knowledge in psychology. Problems of Psychology in the 21st Century 1: 36-46.

MCBRATNEY AB ET AL. 2012. Digital Soil Mapping. In: Huang PM, Li Y and Sumner ME. Handbook of Soil Sciences: Properties and Processes. Boca Raton: Taylor & Francis, 1442 p.

MCBRATNEY AB, FIELD DJ AND KOCH A. 2014. The dimensions of soil security. Geoderma 213: 203-213.

MCBRATNEY AB, MENDONÇA-SANTOS ML AND MINASNY B. 2003. On Digital Soil Mapping. Geoderma 117(1): 3-52.

MINASNY B, HARTEMINK AE AND MCBRATNEY A. 2010. Individual, country, and journal self-citation in soil science. Geoderma 155(3-4): 434-438.

MINASNY B, HARTEMINK AE, MCBRATNEY A AND JANG HJ. 2013. Citations and the h index of soil researchers and journals in the Web of Science, Scopus, and Google Scholar. PeerJ 1: e183.

MINASNY B AND MCBRATNEY AB. 2016. Digital soil mapping: A brief history and some lessons. Geoderma 264: 301-311.

MOED HF. 2009. New developments in the use of citation analysis in research evaluation. Arch Immunol Ther Exp 57(1): 13-18.

NATURE INDEX. 2017. A guide to the Nature Index. Available at: http://www.natureindex.com/. Accessed on February 20, 2018.

ODGERS NP, MCBRATNEY AB AND MINASNY B. 2014. Digital soil property mapping and uncertainty estimation using soil class probability rasters. In: Arrouays D, Mckenzie N, Hempel J, Richer de Forges A and Mcbratney AB (Eds), GlobalSoilMap: Basis of the global spatial soil information system. London: Taylor & Francis, 494 p.

OMUTO C, NACHTERGAELE F AND ROJAS RV. 2013. State of the art report on global and regional soil information: Where are we? Where to go? Global Soil Partnership Technical Report. Roma: FAO, 81 p.

POLIDORO JC ET AL. 2016. Programa Nacional de Solos do Brasil (PronaSolos). Rio de Janeiro: Embrapa Solos, 54 p.

QGIS DEVELOPMENT TEAM ET AL. 2018. QGIS geographic information system. Open Source Geospatial Foundation Project. Available at: http://qgis.osgeo.org. Accessed on February 12, 2018.

R CORE TEAM. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria: 2018. Available at: http://www.R-project.org/. Accessed on January 28, 2018.

ROBINSON DA, PANAGOS P, BORRELLI P, JONES A, MONTANARELLA L, TYE A AND OBST CG. 2017. Soil natural capital in Europe; a framework for state and change assessment. Sci Rep 7(1): 6706.

ROUDIER P, RITCHIE A, HEDLEY C AND MEDYCKYJ-SCOTT D. 2015. The rise of information science: a changing landscape for soil science. IOP Conf Ser Earth Environ Sci 25: 1755-1315.

SAMUEL-ROSA A ET AL. 2018. Bringing together Brazilian soil scientists to share soil data. In: 21st World Congress of Soil Science, Rio de Janeiro. Soil science: beyond food and fuel, 2 p.

SANCHEZ PA ET AL. 2009. Digital soil map of the world. Science 325(5941): 680-681.

SHOKRANEH F, ILGHAMI R, MASOOMI R AND AMANOLLAHI A. 2012. How to select a journal to submit and publish your biomedical paper? BioImpacts 2(1): 61-68.

SULAEMAN Y, MINASNY B, MCBRATNEY AB, SARWANI M AND SUTANDI A. 2013. Harmonizing legacy soil data for digital soil mapping in Indonesia. Geoderma 192: 77-85.

TEN CATEN A, DALMOLIN RSD, MENDONÇA-SANTOS MDL AND GIASSON E. 2012. Mapeamento digital de classes de solos: características da abordagem brasileira. Cienc Rural 42(11): 1989-1997.

THELWALL M AND KOUSHA K. 2014. ResearchGate: Disseminating, communicating, and measuring scholarship? J Assoc Inf Sci Technol 66(5): 876-889.

THOMPSON JA, NAUMAN TW, ODGERS NP, LIBOHOVA Z AND HEMPEL JW. 2012. Harmonization of legacy soil maps in North America: status, trends, and implications for digital soil mapping efforts. In: Digital Soil Assessments and Beyond: Proceedings of the Fifth Global Workshop on Digital Soil Mapping. Sydney, University of Sydney, 482 p.

TRAJANO MA, RAZUCK FB, CERETTA CA AND SCHETINGER MC. 2013. Evolução da produção científica em Ciência do Solo no Brasil: um olhar sobre o Qualis. Geografia 22(3): 93-105.

VARGAS RA, VANZ SAS AND STUMPF IR. 2014. The role of National journals on the rise in Brazilian Agricultural Science Publications in Web of Science. J Scient Res 3(1): 28-36.

WALLIN JA. 2005. Bibliometric methods: pitfalls and possibilities. Basic Clin Pharmacol Toxicol 97(5): 261-275.

WARKENTIN BP. 1994. Trends and developments in soil science. In: Mcdonald P (Ed), The literature of soil science. Hardcover: Cornell University, 448 p.

ZHANG GL, FENG LIU AND SONG XD. 2017. Recent progress and future prospect of digital soil mapping: A review. J Integ Agric 16(12): 2871-2885.

ZHOU P, TIJSSEN R AND LEYDESDORFF L. 2016. University-Industry Collaboration in China and the USA: A Bibliometric Comparison. PLoS ONE 11(11): e0165277.

CAPÍTULO 3 – ESTRATÉGIAS PARA A PREDIÇÃO DE CLASSES DE SOLO COM DADOS LEGADOS NA REGIÃO CENTRAL DO ESTADO DO RS

RESUMO

Com a grande demanda por informações sobre o solo, é necessária a disponibilidade de mapas de solo precisos, ponto onde o Mapeamento Digital de Solos pode contribuir muito. Uma grande quantidade de dados legados, juntamente com alguns mapas de classe do solo legados na forma de polígonos estão disponíveis. Aliado a isto, a aplicação de técnicas como a espacialização da incerteza dos modelos preditivos pode auxiliar a obter mapas mais precisos. Sob a hipótese de que, embora os dados legados tragam uma imperfeita cobertura de pontos amostrais, técnicas como a reamostragem à campo guiada pelo mapa de incerteza ou obtenção de dados em mapas legados pode reduzir a incerteza geral e aumentar a acurácia, o objetivo desse trabalho foi avaliar estratégias de obtenção de dados adicionais para melhorar as previsões de classes de solo gerados a partir de dados legados. Com auxílio de covariáveis ambientais, um mapa de classes de solo foi criado com base em dados legados do Repositório Brasileiro Livre para Dados Abertos de Solo, com 1922 pontos amostrais usando random forest, e avaliados por validação cruzada e validação externa. Adicionalmente, foram testadas estratégias para obtenção de pontos adicionais ao conjunto de calibração com base em mapas legados e reamostragem guiada na incerteza. O mapa gerado apenas com os dados legados obteve acurácia de 0,49 na validação externa, com incerteza geral de 0,84. Um conjunto híbrido, utilizando os dados legados de diferentes fontes foi capaz de melhorar acurácia para 0,55 e reduzir a incerteza para 0,77. Os dados do mapa legado, embora com benefícios ao modelo, demonstraram inconsistências devido a sua resolução espacial grosseira. A reamostragem guiada pela incerteza, que elevou a acurácia para 0,51 e reduziu a incerteza para 0,81, foi a estratégia que demonstrou o maior potencial pela melhoria trazida ao modelo em relação à menor quantidade de dados necessários.

5.1 INTRODUÇÃO

Informações sobre o solo e suas propriedades são essenciais para gerenciar questões agronômicas e ambientais (BRUNGARD et al., 2015) sendo que a demanda por tais informações é cada vez maior em estudos recentes (ROBINSON et al., 2017) devido a importância do solo. O mapeamento convencional do solo, por suas características e

peculiaridades, não possui formas de suprir essa demanda por informações (BAGATINI et al., 2016) devido a maior necessidade de mão de obra, tempo e custo para a geração dos dados. Os mapas de classes de solo convencionais, em particular, possuem a inadequação da escala e questões relativas à precisão do atributo predito (HARTEMINK et al., 2010). Apesar disso, os levantamentos de solos realizados no passado possuem informações valiosas dos ambientes mapeados, permitindo que novos estudos sejam feitos utilizando como base esses dados.

Com o intuito de disponibilizar informações com maior agilidade e resolução espacial adequada, o Mapeamento Digital de Solos (MDS) apresenta uma abordagem semelhante ao mapeamento convencional, baseado na relação solo-paisagem. Contudo, devido a sua natureza quantitativa, os produtos do MDS são reprodutíveis e permitem a representação contínua de dados (STUMPF et al., 2017), além de permitir quantificar a incerteza das informações geradas (ODGERS et al., 2014). Dessa forma, o MDS é uma alternativa para mapear classes e propriedades de solo, usufruindo da disponibilidade cada vez maior de técnicas de processamento e mineração de dados (DALMOLIN e TEN CATEN, 2015; WOLSKI et al., 2017).

A aplicabilidade do MDS depende da disponibilidade e distribuição das observações do solo e da disponibilidade e qualidade das covariáveis ambientais (KROL, 2008). Mesmo com os avanços significativos no que diz respeito a métodos e modelos de predição, a obtenção de dados ainda pode ser um entrave. A obtenção desses dados a campo passa pelos mesmos empecilhos que passa o mapeamento convencional, como maior tempo, custo, e disponibilidade de mão de obra especializada para sua obtenção.

No Brasil, diversos mapeamentos de solo foram realizados. Esses levantamentos foram gerados para diferentes fins, resultando não só nos mapas de solos, mas também em relatórios que incluem as descrições de perfis onde constam todos os dados morfológicos descritos. Esses dados, chamados de dados legados, podem ser utilizados no MDS com um custo reduzido, pois, muitas vezes, não há necessidade de novas coletas no campo (STUMPF et al., 2016). Dada a demanda por informações sobre o solo, os dados legados são fundamentais por servirem como base no MDS para a geração de modelos preditivos de classes ou propriedades do solo (OMUTO et al., 2013; HOUNKPATIN et al., 2018).

O uso de dados legados, juntamente com a grande disponibilidade de covariáveis ambientais, oferecem informações valiosas para sobre os ambientes onde foram coletados (SULAEMAN et al., 2013). Isto pode auxiliar a execução de programas como o PronaSolos (POLIDORO et al., 2018), considerado o maior programa de investigação do solo no Brasil, que tem como objetivo principal mapear os solos de 1,3 milhão de km² do País nos primeiros

dez anos, e mais 6,9 milhões de km² até 2048, em escalas que vão de 1:25.000 a 1:100.000. Projetos nesse âmbito já mostraram sucesso, como é o caso do Soil Survey Geographic Database - SSURGO (SOIL SURVEY STAFF, 2019), onde o MDS é utilizado como ferramenta, juntamente com dados legados, para mapear os solos do país.

As informações geradas pelo MDS possibilitam a estimativa da incerteza (MINASNY e BISHOP 2008), que além de valorizar os produtos do MDS por informar a qualidade da estimativa, permite utilizar as informações sobre a incerteza pode contribuir também para a melhoria da acurácia dos mapas (STUMPF et al 2017).

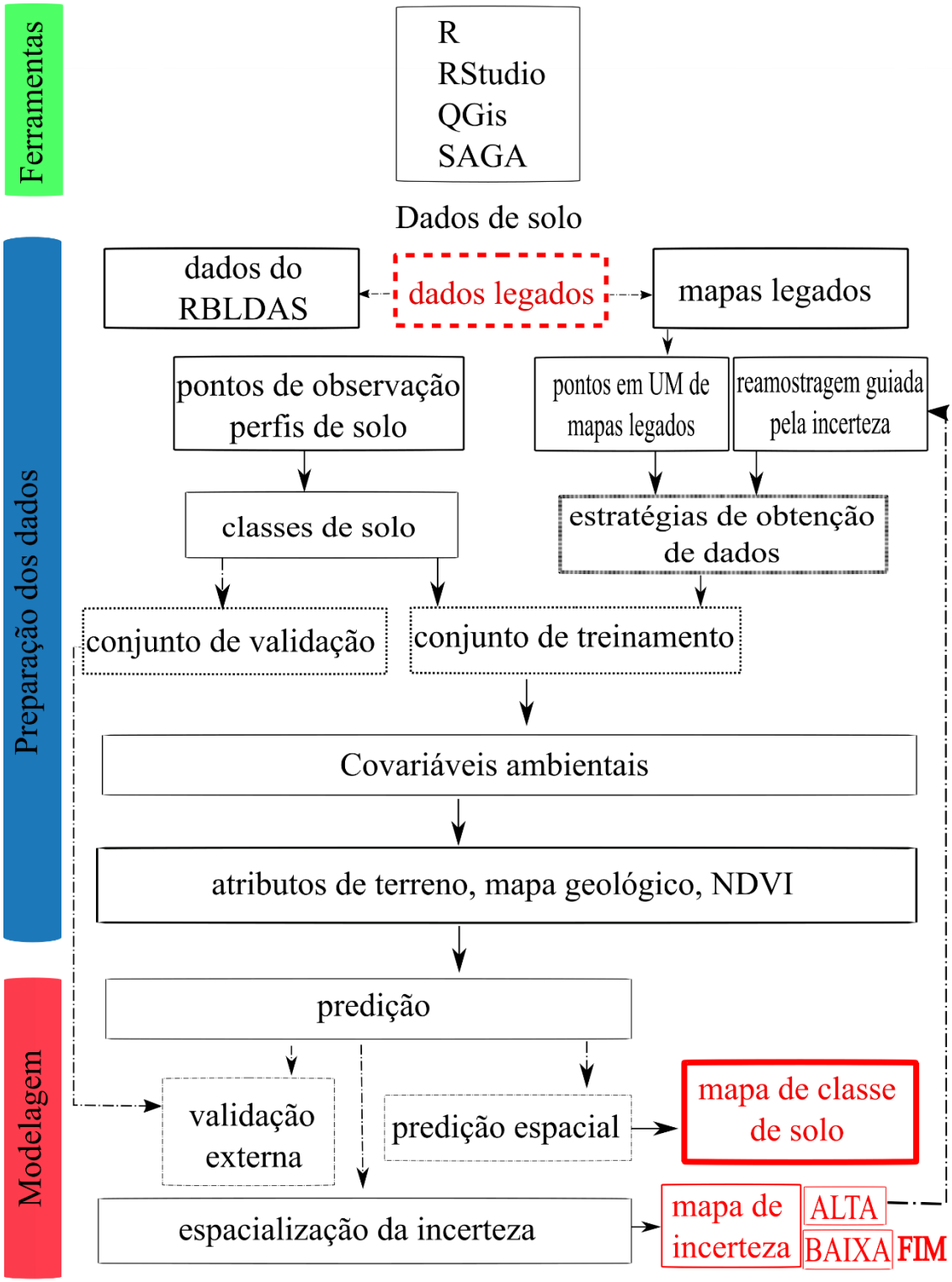
De acordo com Stumpf et al. (2017), a incerteza, se quantificada e espacializada, pode ser usada para aprimorar a amostragem e otimizar a geração de informações. Esses autores usaram um método de reamostragem guiada pela incerteza para melhorar as previsões de silte e argila em uma pequena bacia hidrográfica na China. Com a estratégia proposta, a incerteza espacial média foi reduzida para a predição de silte e argila, demonstrando o potencial da técnica para uso no MDS. Estratégias como essa se tornam importantes sob a ótica de que são necessários métodos eficientes de identificação de locais de amostragem a campo (PAHLAVAN-RAD et al., 2014).

Sob a hipótese de que, embora os dados legados possuam uma imperfeita distribuição de pontos amostrais, técnicas como a reamostragem a campo guiada pelo mapa de incerteza ou obtenção de dados em mapas legados irá aumentar a acurácia do mapa de solo predito, o objetivo do presente trabalho foi avaliar técnicas de obtenção de dados adicionais para melhorar as previsões de classes de solo com uso de dados legados.

5.2 MATERIAL E MÉTODOS

A Figura 1 apresenta um breve roteiro da metodologia do trabalho, que por sua vez será descrita a seguir.

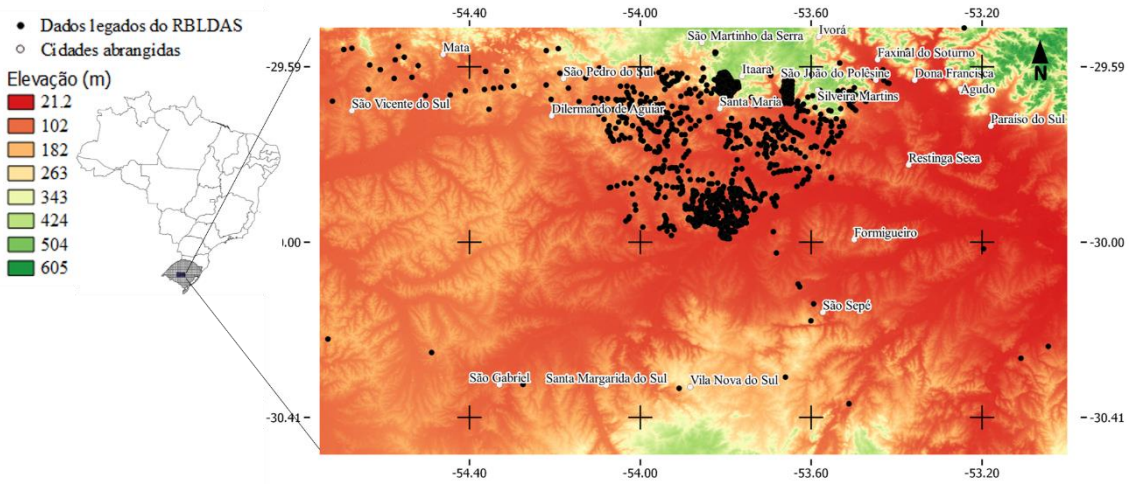
Figura 1 - Fluxograma com o roteiro da metodologia.



5.2.1 Descrição da área

A área de estudo compreende aproximadamente 13000 km² da região Central do Estado do Rio Grande do Sul (Figura 2), que está entre as áreas prioritárias do PronaSolos. Nela, se insere a transição entre as regiões fisiográficas do Planalto Rio-Grandense e a Depressão Central, incluindo a região de transição entre essas, denominada Rebordo do Planalto. O clima da região, conforme o sistema de classificação de Köppen, é subtropical do tipo Cfa (ALVARES et al., 2013). O Rebordo do Planalto se destaca pela geomorfologia complexa (SARTORI, 2009), composta pela Sequência Superior da Formação Serra Geral (rochas ígneas–riolito-riodacito), Sequência Inferior da Formação Serra Geral (rochas ígneas–basalto-andesito), Formação Botucatu (rochas sedimentares - arenito eólico), Formação Caturrita (rochas sedimentares – arenito fluvial) e depósitos fluviais recentes. Já a Depressão Central é uma região composta por rochas sedimentares diversificadas, originadas por deposição de sedimentos de composição e granulometria variada, apresentado uma sucessão de diferentes camadas de rochas sedimentares (STRECK et al., 2018). A área de estudo apresenta relevo que varia de plano a montanhoso, com altitudes entre 21 e 605 m.

Figura 2 - Localização da área de estudo, representando a variação de relevo, os pontos amostrais disponíveis no RBLDAS e as cidades inseridas na área de estudo.



5.2.2 Obtenção dos dados legados

Os dados legados foram obtidos a partir do Repositório Brasileiro Livre para Dados Abertos do Solo - RBLDAS (www.ufsm.br/febr) desenvolvido pelo Grupo de Pedologia da UFSM. O RBLDAS foi criado com o propósito de servir como meio para a compilação, organização e publicação de todos os tipos de dados do solo no Brasil. Para a área de estudo, o RBLDAS disponibiliza 1922 pontos amostrais legados, obtidos em estudos previamente realizados, onde constam informações sobre a classe de solo, sendo esses os dados legados na forma de pontos amostrais utilizados no presente trabalho. Desses dados, 226 são representados por perfis completos descritos no campo, classificados pelo menos até o 2º nível categórico do Sistema Brasileiro de Classificação do Solo (SiBCS) (SANTOS et al., 2013), conforme apresentado na Tabela 1.

Tabela 1 - Distribuição das 13 classes encontradas nos dados legados e sua identificação conforme o SiBCS.

n	%	Classe de solo	Sigla da classe
334	17.4	Argissolo Vermelho	PV
329	17.1	Neossolo Litólico	RL
265	13.8	Argissolo Bruno-Acinzentado	PBAC
219	11.4	Argissolo Vermelho-Amarelo	PVA
184	9.6	Planossolo Háplico	SX
151	7.9	Argissolo Amarelo	PA
113	5.9	Neossolo Flúvico	RY
110	5.7	Cambissolo Háplico	CX
65	3.4	Neossolo Regolítico	RR
22	1.1	Argissolo Acinzentado	PAC
8	0.4	Gleissolo Háplico	GX
6	0.3	Luvissolo Háplico	TX
3	0.2	Neossolo Quartzarênico	RQ

Pelo fato dos dados terem sido adquiridos de diferentes fontes, as unidades de análise, sistema de referência geográfico e profundidade do solo observadas também serão diferentes. Isso trouxe a necessidade de realizar a adequação e harmonização dos dados, que foi realizada em duas etapas, conforme sugerido por Sulaeman et al (2013). A primeira etapa diz respeito à padronização da localização das amostras. Grande parte dos pontos coletados foram georreferenciados por sistemas atualmente desatualizados e por esse motivo tiveram suas

coordenadas geográficas convertidas para o sistema WGS84. Sem essa padronização, a integração dos dados de solo não seria possível devido à incompatibilidade geodésica. A segunda etapa trata da harmonização das informações contidas. A principal etapa dessa padronização foi a reclassificação dos pontos amostrais conforme o SiBCS (SANTOS et al., 2013) até o 2º nível categórico (Subordens), visto que a maior parte dos mapeamentos é classificada com edições anteriores ou aproximações do SiBCS.

5.2.3 Obtenção das covariáveis ambientais

Os dados foram obtidos dos seguintes bancos de dados: Google Earth Engine, disponibilizando informações dos satélite ALOS e LANDSAT com resolução de 30 m; Serviço Geológico do Brasil (CRPM), disponibilizando o mapa geológico do Estado do Rio Grande do Sul na escala de 1:750.000; e TOPODATA (VALERIANO, 2008), disponibilizando covariáveis ambientais previamente derivadas, além de disponibilizar o MDE com resolução de 30 m. Do MDE foram derivadas covariáveis ambientais de acordo com Wilson e Gallant (2000), processadas pelo software SAGA-GIS 2.3.2 (CONRAD et al., 2015). Como preditores, um conjunto de 22 covariáveis ambientais relacionadas a fatores de formação do solo foi utilizado para a predição das classes de solo (Tabela 2).

Tabela 2 - Covariáveis ambientais utilizadas, apresentando sua origem, descrição e fator de formação do solo a qual representa

Origem	Covariável ambiental	Descrição	Fator "scorpan"[a]
MDE TOPODATA	Elevação	MDE representando a elevação da superfície terrestre	r
	Analytical hillshading	Apresenta o ângulo em um feixe de luz proveniente da posição fonte atingiria a superfície	r
	Aspecto	Expressa a orientação da encosta	r
	Área de contribuição	Representa a área de captação de fluxo	r
	Índice de convergência	Apresenta um índice de convergência / divergência em relação ao fluxo terrestre	r
	Cross-sectional curvature	Representa a curvatura tangencial que se intersecta com o plano definido pela superfície normal e uma tangente ao contorno	r
	Curvatura vertical	Apresenta a curvatura média do terreno no sentido vertical da rede de drenagem	r
	Curvatura geral	Apresenta uma medida geral da convexidade do terreno	r
	Curvatura longitudinal	Representa a curvatura do perfil que se cruza com o plano definido pela direção do gradiente máximo da superfície	r
	Declividade	Expressa a mudança de elevação sobre certa distância	r
	Fator LS	Representação da declividade que prevê o potencial de erosão do solo	r

	Curvatura planar	Se refere ao caráter divergente ou convergente dos fluxos de matéria e energia nas vertentes, em projeção horizontal	r
	Curvatura de perfil	Se refere à forma da vertente, podendo ser convexa, côncava ou retilínea, ao ser analisada em perfil	r
	Curvatura tangencial	Computa a tangente de um plano em relação à topografia da superfície	r
	Índice de umidade do terreno (TWI)	Caracteriza a distribuição espacial de zonas de saturação superficial e conteúdo de água nas paisagens	r
Serviço Geológico do Brasil - CRPM	Geologia	Mapa de geologia (escala 1:750.000)	p
ALOS	Índice de posição topográfica	Indica a posição para cada local em relação a elevação média dentro de uma vizinhança	r
	Geoformas	Classes de forma de terreno, evidenciando formas terrestres e padrões fisiográficos	r
Landsat 8	Índice de Vegetação da Diferença Normalizada	Indica a condição da vegetação natural, buscando evidenciar a ação de organismos pela vegetação presente	o
	Talvegues e divisores de água	Evidenciação de talvegues e divisores de água a partir de curvaturas do terreno	r
TOPODATA	Orientação de vertentes	Classes de orientação de vertentes (em octantes)	r
	Formas do terreno	Classes de curvatura horizontal e vertical	r

[a] O, organismos; R, relevo; MO, material de origem.

5.2.4 Predição e avaliação dos modelos

Os modelos de predição foram gerados pelo algoritmo Random Forest (RF), implementado no pacote Caret (KUHN, 2013) pelo software R (R DEVELOPMENT CORE TEAM, 2018) usando os parâmetros disponíveis. O RF é um termo utilizado para um conjunto de classificadores baseados em árvores, desenvolvido com base no classificador de árvore de regressão, que busca melhorar o desempenho do modelo e pode ser aplicado a predições de variáveis contínuas e categóricas (HEUNG et al., 2016; PINHEIRO et al., 2018). O RF se diferencia do método de árvores por construir muitas árvores, sem processo de poda, obtendo assim “florestas”, e para cada árvore, apenas um subconjunto das variáveis de previsão é utilizado.

Como os diferentes conjuntos de dados foram compostos por dados advindos de diferentes fontes, que por sua vez, podem trazer consigo inconsistências no que diz respeito a qualidade da informação, informando ao modelo o valor de confiança para cada ponto amostral. O valor de confiança varia de 0 a 100 %, sendo 0 % a menor confiança e 100 % a maior confiança na informação. Esse é um passo importante, visto que mapas de classes de solo legados, por sua natureza, permitem até 20 % de outras classes, na forma de inclusões, dentro de uma unidade de mapeamento simples. Dessa forma, dados legados de mapas foram introduzidos no modelo com 80 % de confiança e dados legados que foram coletados a campo foram introduzidos com 100 % de confiança.

O treinamento do modelo usou validação cruzada e os modelos de predição de classe de solo tiveram seu desempenho avaliado pela acurácia geral. Buscou-se também uma avaliação mais rigorosa dos mapas de classe de solo gerados. Adicionalmente, cada mapa gerado teve seu desempenho avaliado por um conjunto de validação externa, gerado utilizando metade dos pontos amostrais na forma de perfis de solo descritos (n=113). Os pontos do conjunto de validação externa foram selecionados aleatoriamente e retirados do conjunto de treinamento antes da geração dos modelos. Dessa forma, o conjunto de validação externa serviu para validar todos os mapas gerados a partir de uma matriz de confusão, de onde foram derivados os índices acurácia geral e acurácia balanceada de classe (CONGALTON, 1991).

Também para a avaliação dos modelos, foi derivada uma medida do índice de confusão (BURROUGH et al., 1997) espacialmente explícito usando as previsões realizadas pelos modelos, que será denominado como incerteza. O índice de confusão pode ser caracterizado por valores de probabilidade que são produzidos como um subproduto da classificação. Essas medidas fornecem uma informação útil sobre a qualidade da classificação resultante em termos das incertezas envolvidas. Esse índice traz uma medida da confusão que o modelo preditivo faz entre as duas classes de solo mais prováveis, variando entre 0 e 1, onde 1 significa máxima confusão. O índice de confusão é calculado usando a seguinte equação (Equação 1):

$$IC = 1 - (P1 - P2) \quad (1)$$

Em que: P1 e P2 são, respectivamente, as probabilidades da classe de solo mais provável e a segunda classe de solo mais provável.

5.2.5 Conjuntos de dados e estratégias de obtenção de novos dados

A estratégia inicial do presente trabalho foi gerar um mapa de classes de solo utilizando apenas dados legados do RBLDAS. O conjunto de dados para esse fim foi chamado de estratégia 0 (E0), sendo composto pelos 1922 pontos obtidos no RBLDAS, dos quais foram subtraídos os 113 pontos de perfis utilizados para o conjunto de validação externa, finalizando o conjunto E0 com 1809 pontos amostrais compostos por perfis descritos

e pontos amostrados a campo. Como todos os pontos amostrais obtidos do RBLDAS tiveram as classes de solo identificadas a campo, o valor de confiança informado para o modelo preditivo foi de 100 %.

Buscando melhorar a acurácia e reduzir a incerteza das informações contidas no mapa de classe de solo utilizando apenas dados legados (E0), foram traçadas duas formas para obtenção de novos dados com o intuito de agregar informações para o treinamento do modelo: (E1) utilização de amostragem em mapas de classes de solo legados e (E2) utilização de reamostragem guiada pela incerteza. O objetivo dessas estratégias é tentar adquirir mais informações para que, a partir de um conjunto de treinamento mais robusto o modelo preditivo consiga capturar com mais precisão as variações do solo na paisagem e, consequentemente, gerar mapas com mais acurácia. Ambas essas estratégias serão agora detalhadas.

5.2.5.1 Estratégia 1 (E1): predição com obtenção de pontos adicionais em mapas de classes de solo legados.

Nessa estratégia foram usadas informações contidas nos levantamentos de solos descritos na Tabela 3.

Tabela 3 - Informações contidas nos levantamentos utilizados.

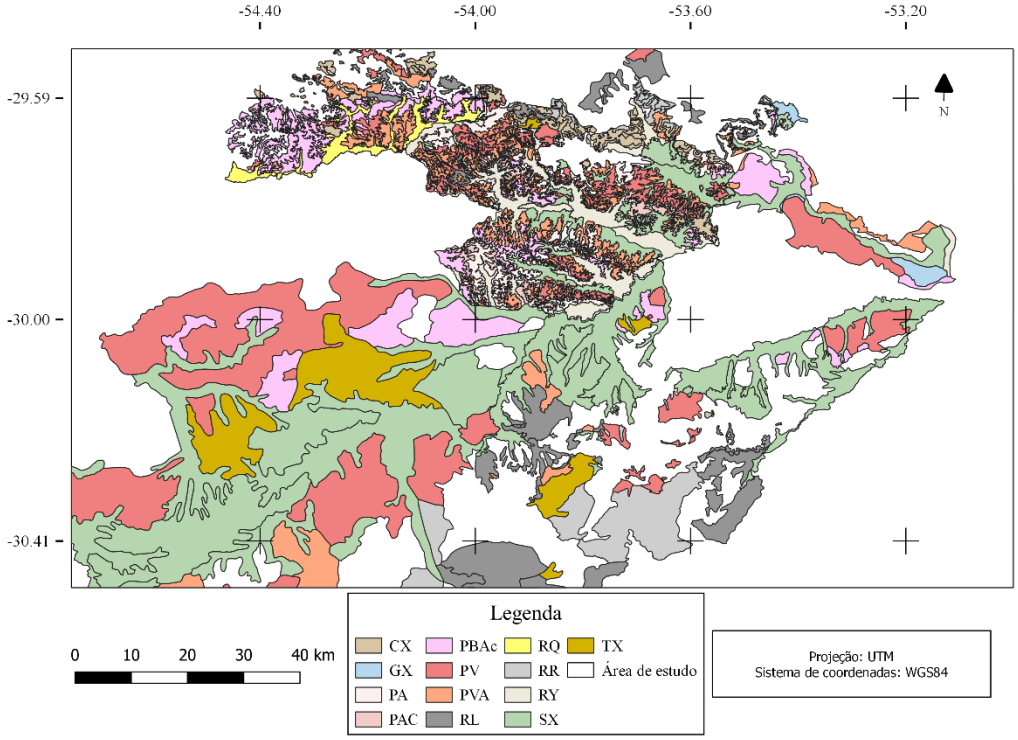
Levantamento de solos	Escala	Tipo de levantamento	Ano de realização
Santa Maria	1:50.000	Semidetalhado	2015
São Gabriel	1:450.000	Reconhecimento / alta intensidade	1968
São Sepé	1:100.000	Semidetalhado	1972
Bacia do Rio Vacacaí	1:100.000	Semidetalhado	1988
São João do Polêsine	1:20.000	Semidetalhado	1997
São Pedro do Sul	1:50.000	Semidetalhado	2001

Esses mapas de solos cobrem uma área de 8397 km², representando 65 % do total da área de estudo. Nos mapas, predominam as classes SX (33 % da área), PV (20 %), PBAC (10

%), PVA (7 %) e RL (7 %), seguidas de TX (6 %), RR (5 %), RY (4 %), PA (3 %), CX (2 %), PAC, (1 %), RQ (1 %) e GX (0,6 %) (Tabela 1).

Os mapas foram digitalizados e tiveram seus polígonos vetorizados no software QGis 3.2.3 (QGIS DEVELOPMENT TEAM, 2018). Nas áreas onde ocorreu a sobreposição de mapas foram utilizados aqueles com escala mais detalhada. As classes de solo foram atualizadas conforme o SiBCS 2013 (SANTOS et al., 2013) e apenas foram utilizados polígonos de classes já encontradas no conjunto de dados E0 (Figura 3).

Figura 3 - Polígonos dos mapas de classes de solo legados dentro da área de estudo.



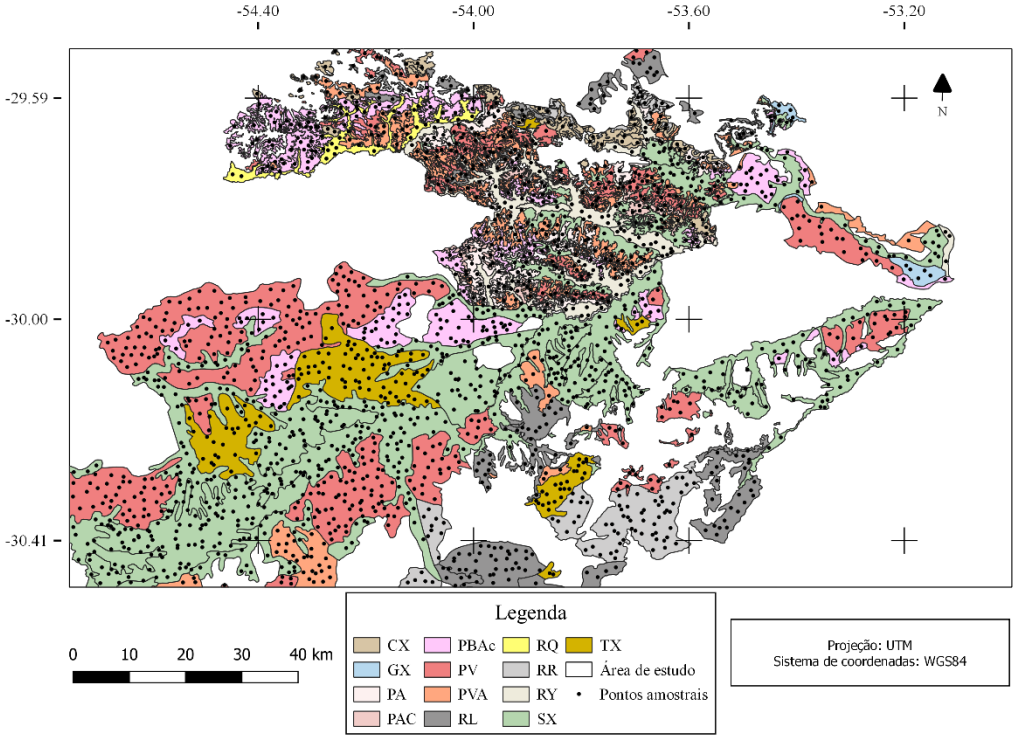
*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

As informações foram utilizadas para gerar pontos amostrais sobre as unidades de mapeamento, constando suas respectivas classes de solo conforme os mapas. Apenas as unidades de mapeamento simples foram consideradas para extração das informações pontuais, ou seja, os polígonos com unidades de mapeamento compostas foram desconsiderados.

Toda metodologia acima descrita foi baseada em trabalhos realizados por Heung et al. (2014) e Heung et al. (2016), onde os mapas legados foram utilizados para a geração de pontos para a predição. Baseado na metodologia utilizada por Heung et al. (2016), foi

realizada a distribuição de 2000 pontos de forma aleatória ponderada em relação a área de cada polígono (Figura 4). Dessa forma, cada classe recebeu o número de pontos ponderado de acordo com a sua área. Esses pontos foram gerados em uma distância mínima de 60 metros entre si e com uma distância mínima de 160 metros da borda dos polígonos, conforme ten Caten et al. (2012), visando eliminar áreas próximas da borda do polígono, consideradas mais incertas devido a possíveis erros de localização ou determinação errônea do limite de transição entre as classes no mapa (PELEGRINO et al., 2016).

Figura 4 - Representação dos 2000 pontos amostrais gerados sobre os mapas de classe de solo legados.



*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

Assim, considerou-se que cada um dos pontos da Figura 4, por estarem dentro da unidade de mapeamento, são representativos da classe taxonômica.

Para testar o aumento do número de pontos na calibração e verificar se haveria melhoria na qualidade dos mapas, foi realizada uma análise de sensibilidade, criando-se diferentes conjuntos acrescentando um número crescente de pontos ao conjunto da estratégia E0, obtendo assim os seguintes conjuntos de calibração:

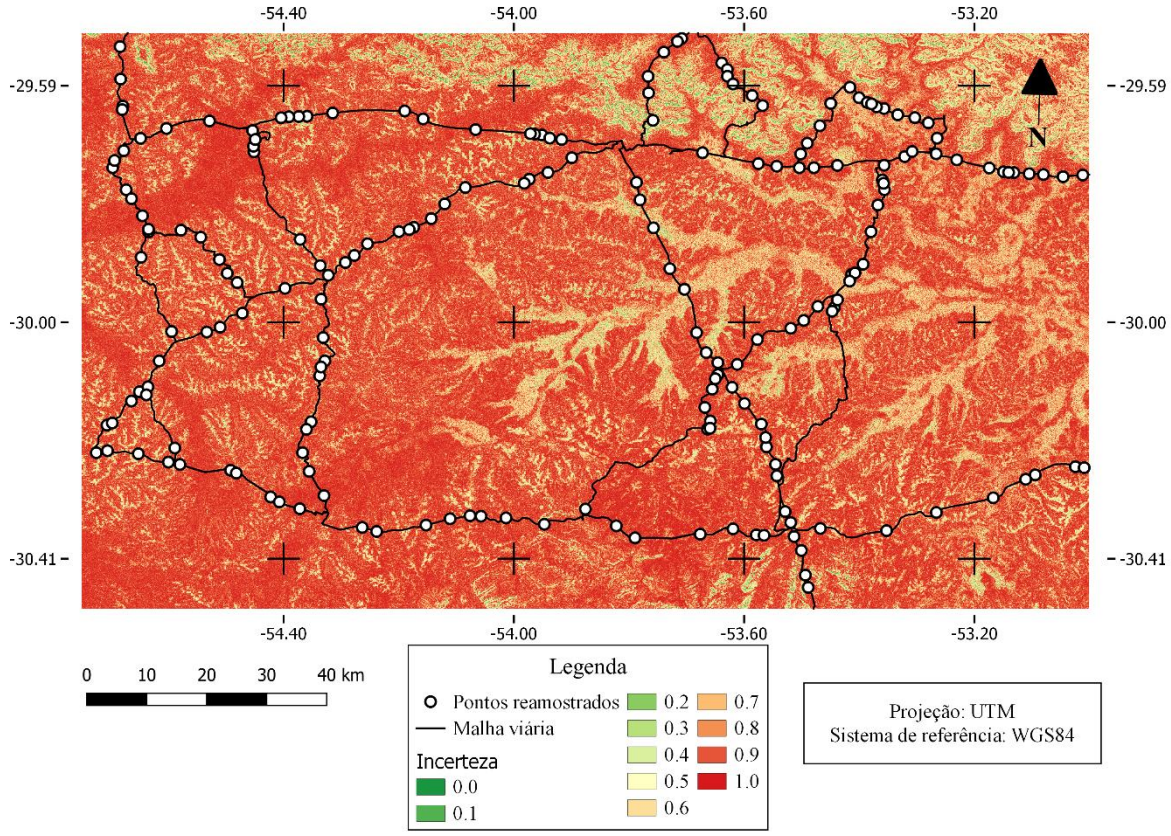
- E1-400 (1809 pontos legados + 400 pontos do mapa legado);
- E1-800 (1809 pontos legados + 800 pontos do mapa legado);
- E1-1200 (1809 pontos legados + 1200 pontos do mapa legado);
- E1-1600 (1809 pontos legados + 1600 pontos do mapa legado);
- E1-2000 (1809 pontos legados + 2000 pontos do mapa legado).

Para esses pontos foi considerado um valor de confiança de 80% para que a classe de solo fosse a representante da UM em questão, pois em até 20% da área pode haver outras classes na forma de variação e/ou inclusão.

5.2.5.2 Estratégia 2 (E2): reamostragem guiada pela incerteza

Como as amostras do conjunto E0 não foram coletadas propositalmente para cobrir a variabilidade espacial de classes de solo, foi realizada a amostragem sobre as áreas de maior incerteza identificadas no mapa gerado pelo conjunto E0. Assim, foi utilizada uma estratégia na metodologia utilizada por Stumpf et al. (2017), em que amostras adicionais foram obtidas a campo em 200 pontos dispostos aleatoriamente sobre as áreas de incerteza, em uma amostragem facilitada por vias de acesso, com um buffer de 200 metros para cada lado das vias. Nesse buffer, os 200 pontos amostrais foram aleatoriamente alocados em pixels de máxima incerteza (Figura 5) do mapa de incerteza gerado pelos dados da estratégia E0. A campo, em cada um dos pontos as classes de solo foram identificadas com auxílio de um trado e classificadas até o segundo nível categórico conforme o SiBCS (SANTOS et al., 2013).

Figura 5 - Pontos para a estratégia de reamostragem com base na incerteza.



Essas novas amostras coletadas foram agrupadas às do conjunto E0 em número crescente, com o propósito de verificar a variação na acurácia proporcionada pela adição de novos pontos reamostrados pela incerteza, formando os seguintes conjuntos de dados:

- E2-50 (1809 pontos legados + 50 pontos de reamostragem guiada);
- E2-100 (1809 pontos legados + 100 pontos de reamostragem guiada);
- E2-150 (1809 pontos legados + 150 pontos de reamostragem guiada);
- E2-200 (1809 pontos legados + 200 pontos de reamostragem guiada).

Como esses pontos tiveram as classes de solo identificadas a campo, o valor de confiança informado para o modelo preditivo é de 100 %.

5.2.5.3 Estratégia 3 (E3) – Predição usando apenas pontos gerados em mapas de classes de solo legados

Buscando verificar a viabilidade dos pontos gerados sobre mapas legados, foi criado um conjunto de dados apenas com os pontos adicionais obtidos pela estratégia E1, dessa vez sem o uso dos dados legados do RBLDAS, totalizando 2000 pontos para o conjunto de calibração E3. Essa estratégia está mais próxima a utilizada por Heung et al. (2016), onde os autores utilizaram somente pontos gerados sobre o mapa legado para usar como conjunto de calibração para um novo mapa digital. Como esses pontos amostrais foram gerados sobre um mapa onde, mesmo sendo uma unidade taxonômica simples, podem conter até 20% de outras classes na forma de inclusões, esses pontos tiveram valor de confiança de 80% informado ao modelo preditivo.

5.2.5.4 Estratégia 4 (E4) – Predição usando um banco de dados híbrido com as duas formas de obtenção de dados

Para testar se um maior número de pontos aumentaria a acurácia dos mapas gerados, foram criados dois conjuntos de dados. O primeiro conjunto de dados foi criado utilizando a maior quantidade de informações possível, utilizando todos os pontos gerados por ambas as metodologias. Para isto, o banco de dados E4-1 conta com os dados do conjunto E0 (1809 pontos), com valor de confiança de 100%; somados aos pontos amostrais gerados sobre os mapas de classes de solo legados – conjunto E3 (2000 pontos), com valor de confiança de 80%; somados aos dados obtidos pela reamostragem guiada pela incerteza (200 pontos), com nível de confiança de 100%. Ao total, o conjunto E4-1 conta com 4009 pontos amostrais, advindos das diferentes fontes.

O segundo conjunto de dados foi criado para verificar se a quantidade de informações não condiciona melhor acurácia aos mapas, mas sim a qualidade das informações que permitem melhores resultados na predição de classes de solo. O conjunto de dados E4-2, que une os conjuntos E0 (1809 pontos), com valor de confiança de 100%; somado ao conjunto de dados que obteve o melhor desempenho na estratégia E1 (E1-800), com 800 pontos amostrados nos mapas legados com valor de confiança de 80%; somados ao conjunto de dados com melhor desempenho da estratégia E2 (E2-200), com 200 pontos reamostrados com

base na incerteza e valor de confiança de 100%. Dessa forma, o conjunto E4-2 conta com um total de 2809 pontos amostrais advindos de diferentes fontes de dados.

O resumo da distribuição de classes de todos os conjuntos de dados testados está apresentado na Tabela 4.

Tabela 4 – Percentual de cada classe de solo em cada conjunto de dados testado.

Conjunto	PV	PBAC	PVA	SX	RL	PA	CX	RY	RR	PAC	RQ	GX	TX
E0	18.46	14.65	12.11	10.17	18.19	8.35	6.08	6.25	3.59	1.22	0.17	0.44	0.33
E1-400	18.83	13.76	11.23	14.35	16.16	7.33	5.34	5.89	3.76	1.18	0.36	0.45	1.36
E1-800	19.13	13.11	10.66	17.25	14.72	6.67	4.83	5.63	3.91	1.11	0.50	0.46	2.03
E1-1200	19.37	12.70	10.23	19.07	13.73	6.17	4.47	5.47	4.00	1.10	0.63	0.50	2.57
E1-1600	19.45	12.32	9.86	20.98	12.91	5.75	4.17	5.31	4.05	1.06	0.70	0.50	2.93
E1-2000	19.58	12.02	9.61	22.26	12.28	5.43	3.94	5.20	4.12	1.05	0.76	0.50	3.25
E2-50	18.44	15.74	11.81	9.97	17.95	8.14	6.15	6.09	3.61	1.19	0.16	0.43	0.32
E2-100	18.19	16.46	11.78	9.94	17.67	7.94	6.31	6.05	3.63	1.16	0.16	0.42	0.32
E2-150	18.01	17.03	11.75	9.75	17.39	7.95	6.36	5.90	3.59	1.13	0.15	0.67	0.31
E2-200	17.89	17.59	11.62	9.67	17.28	7.77	6.41	5.81	3.51	1.15	0.20	0.80	0.30
E3	20.59	9.65	7.35	33.18	6.95	2.80	2.00	4.25	4.60	0.90	1.30	0.55	5.90
E4-1	19.24	13.61	9.48	21.44	12.11	5.28	4.20	5.03	4.05	1.03	0.75	0.68	3.10
E4-2	18.67	15.31	10.41	16.42	14.31	6.37	5.15	5.36	3.83	1.07	0.50	0.72	1.90

PV: Argissolo Vermelho; PBAC: Argissolo Bruno-Acinzentado; PVA: Argissolo Vermelho-Amarelo; SX: Planossolo Háplico; RL: Neossolo Litólico; PA: Argissolo Amarelo; CX: Cambissolo Háplico; RY: Neossolo Flúvico; RR: Neossolo Regolítico; PAC: Argissolo Acinzentado; RQ: Neossolo Quartzarênico; GX: Gleissolo Háplico; TX: Luvissolo Háplico.

5.3 RESULTADOS E DISCUSSÃO

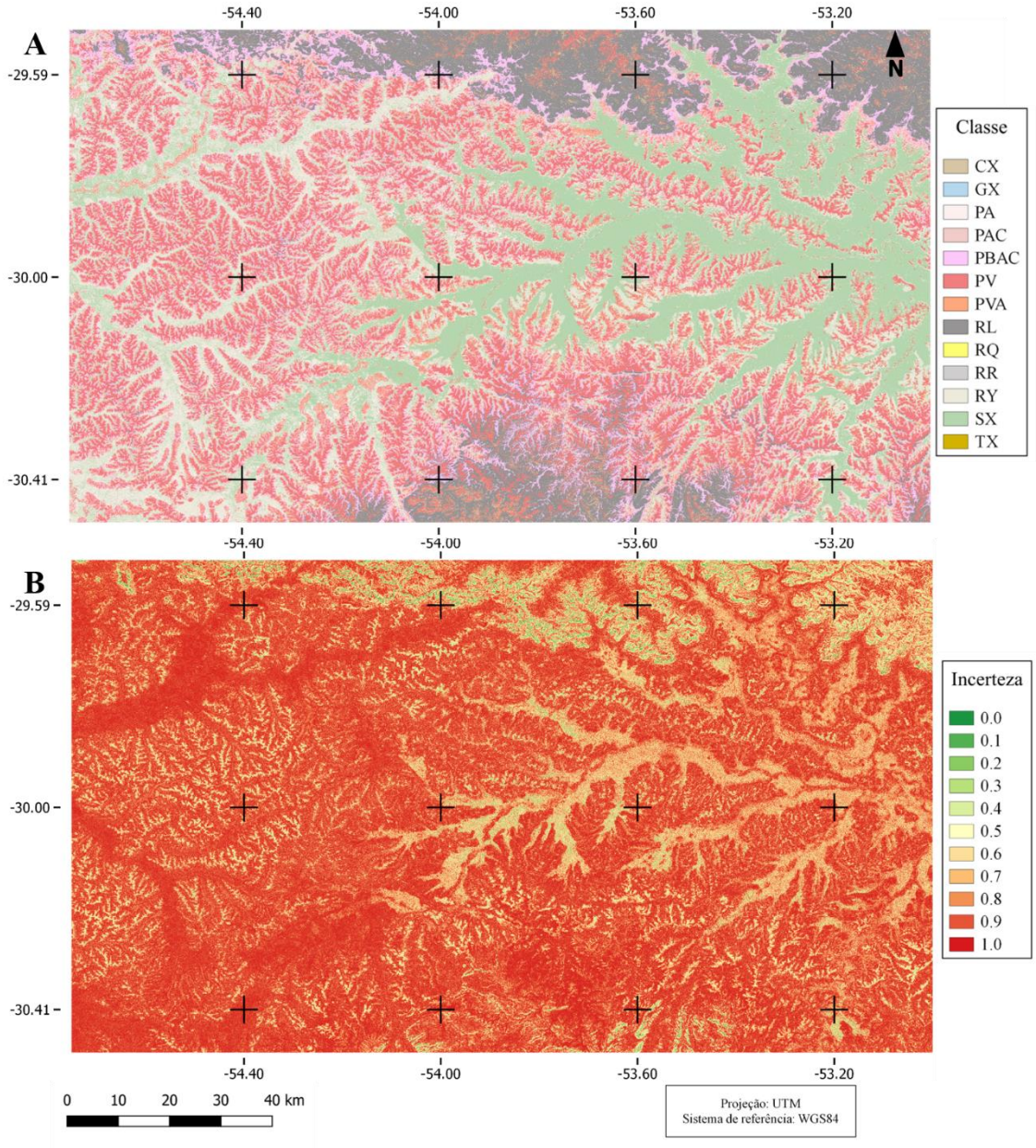
5.3.1 Estratégia E0 – Predição usando os dados legados do RBLDAS

Com base nos dados do conjunto E0, um total de 11 classes de solo foram mapeadas até o segundo nível categórico (Figura 6a). O mapa gerado pelos dados do conjunto E0 apresentou acurácia de 0,59 na validação cruzada e 0,49 na validação externa. Embora não exista um valor de acurácia estabelecido como ótimo pela comunidade cientifica, o valor de acurácia obtido está de acordo com o trabalho de Jeune et al. (2018), que obteve acurácia de 0,52, embora não usando dados legados. Contudo, os resultados de acurácia ficaram bem abaixo dos 0,60 obtidos por Adhikari et al. (2014), que usaram dados legados para predizer 8 classes de solo, em primeiro nível categórico, mesmo em uma área de 43.000 km².

Sob essa visão, os resultados de acurácia podem ser considerados satisfatórios, considerado o fato de ser uma área de grandes proporções e se fazer uso de dados legados

dispostos irregularmente na área de estudo. Ressalta-se que um grande número de classes na área reduz significativamente os resultados de acurácia (BRUNGARD et al., 2015), além de que o aumento do detalhe taxonômico reduz a precisão da predição (JAFARI et al., 2013).

Figura 6 - (A) Mapa de classe de solo gerado pelo conjunto E0 e (B) seu respectivo mapa de incerteza.



*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

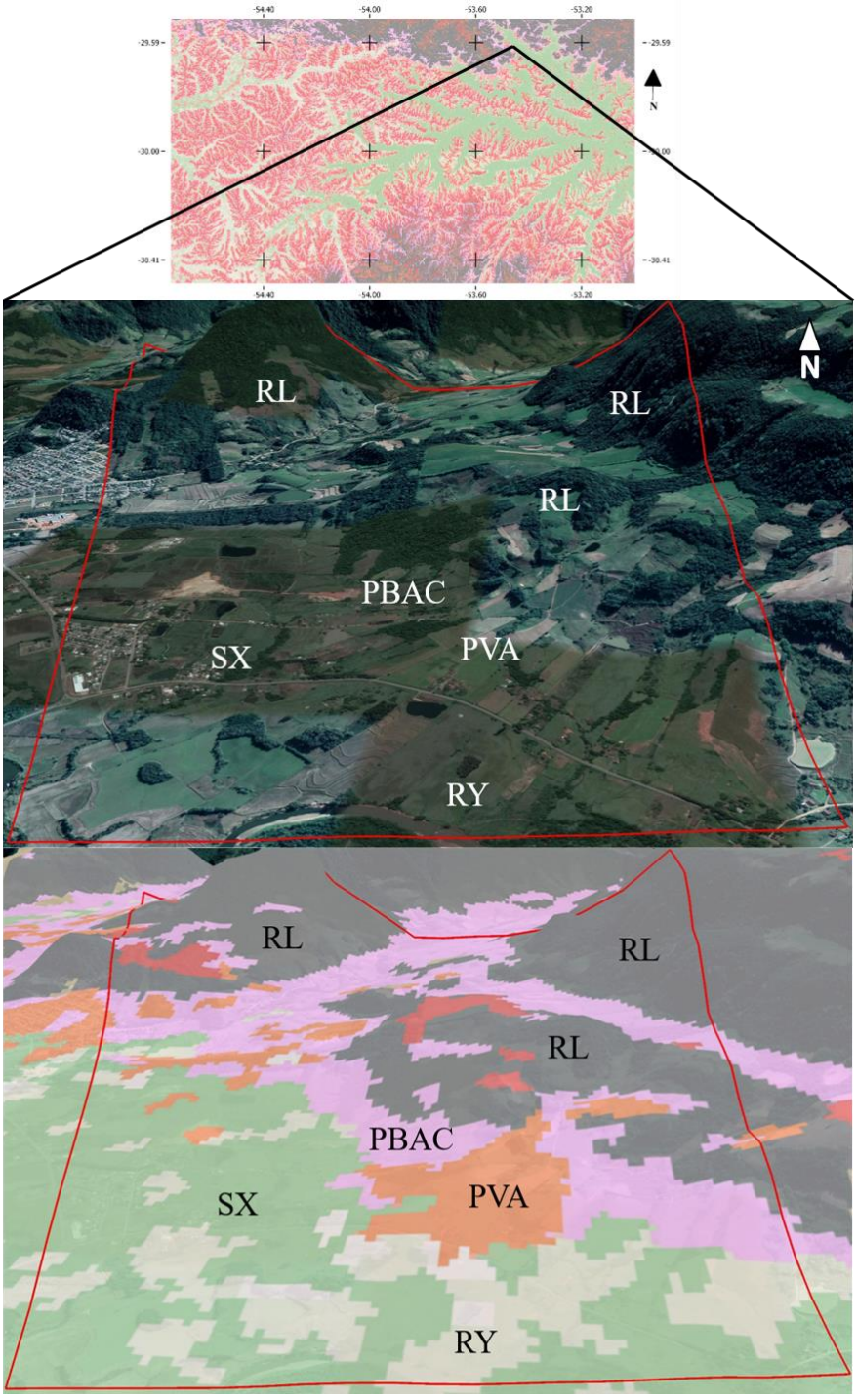
Diante disso, a área apresenta uma grande quantidade de classes distintas em segundo nível categórico, como no caso dos Argissolos, onde as subclasses ocorrem em um relevo semelhante e apenas a cor realiza a distinção das cinco subordens mapeadas. Conforme Meier et al (2018), a pequena variação na matiz, valor e croma do solo, que distingue as classes de solo, torna difícil a distinção das classes na paisagem por parte do modelo. Nas áreas de ocorrência de Argissolos, sua distribuição na paisagem está condicionada aos fluxos de água e ao seu regime hídrico. Comumente, nas porções com maior elevação, onde a drenagem da paisagem ocorre mais rapidamente, ocorrem principalmente PV e PVA. Nas porções mais baixas da paisagem, que por sua vez comumente tem drenagem mais lenta devido a maior proximidade com o lençol freático, ocorrem PBAC, PA e PAC. Tal condição ocorre devido à grande variação do material de origem da área de estudo. Pedron et al. (2006) destacam que sedimentos oriundos de diferentes composições granulométricas podem alterar a drenagem e umidade dos solos, diferenciando os Argissolos, muitas vezes em áreas vizinhas, ocasiona tamanha variação na cor (subordem) das classes de solo.

O mapa gerado pelo conjunto de dados E0, contudo, apresentou uma elevada incerteza geral, obtendo média de 0,84 (Figura 6b). Os resultados de incerteza estão próximos aos encontrados por Zeraatpisheh et al. (2019), que na maior parte da área obtiveram valores de incerteza entre 0,6 e 0,8. Como a incerteza é uma medida da confusão que o modelo apresenta sobre as duas classes de solo mais prováveis, é possível que ele confunda, por exemplo, PV com PVA, ou PVA com PA, que são classes taxonomicamente muito próximas e diferenciadas apenas pela cor. Conforme Vicent et al. (2018), valores de incerteza próximos de 1 representam que a probabilidade de ocorrência das duas classes de solo mais prováveis era muito próxima. Dessa forma, o modelo pode estar certo sobre a ordem de solo a ser predita, mas como a intenção do modelo é predizer também as subordens, o modelo entende essas como “classes” diferentes, elevando, dessa forma, o valor de incerteza.

A Figura 7 apresenta uma visão do relevo típico do Rebordo do Planalto, considerado a transição entre a Região do Planalto e a Depressão Central do Estado do Rio Grande do Sul. A distribuição do mapa de classes de solo gerado pelo conjunto de dados E0, plotado sobre o relevo da região, demonstra a predominância das classes RL, SX, RY, PBAC e PVA. A distribuição das classes está de acordo com o que é comumente encontrado na região, conforme relatado por Klamt et al. (1997). Nas regiões mais baixas da área, é típica a ocorrência das classes RY e SX, sendo essa última indicação de inundação sazonal nas regiões mais planas da paisagem. Embora RY e SX ocorram em áreas próximas e, visualmente, apresentem confusão na sua distribuição, a classe RY predomina nas regiões que

margeiam o Rio Soturno, situado no extremo sul da área delimitada pela Figura 7. Já SX, embora também encontrada nesta região, predomina nas áreas mais baixas e planas do relevo, com uso predominante pelo cultivo agrícola.

Figura 7 - Distribuição de classes sobre o relevo de uma área típica da transição entre a Região do Planalto e a Depressão Central do Estado do RS.



*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

Observa-se também a predominância das classes RL e, em alguns locais, RR, classes essas com ocorrência nas áreas com relevo mais íngreme do Rebordo do Planalto, em que predomina a vegetação nativa da região. Já em áreas de transição entre as planícies e as regiões íngremes da paisagem, há maior ocorrência das PBAC e PVA, além de algumas inclusões da classe PV em regiões mais elevadas da área, ambas ocorrendo em relevo ondulado e suave-ondulado.

A partir da matriz de confusão gerada pela validação externa (Tabela 5), percebe-se uma maior acurácia das classes CX, PBAC, RL e TX. Nota-se também uma dificuldade por parte do modelo para predizer corretamente as classes RL e RR, que por serem da mesma ordem, possuírem características correlatas e ocorrerem em condições de relevo muito semelhantes, torna difícil mapeá-las com precisão. Contudo, por serem classes com aptidão ao uso muito semelhantes, a confusão entre RL e RR não representa um erro tão grave por parte do modelo quanto a confusão de RL com outras classes, como por exemplo PV ou PVA, essas de aptidão ao uso significativamente diferentes. Um fato que chama atenção é a confusão realizada pelo modelo entre as classes PV e SX, visto que são distintas não só por suas características, mas também pelo seu local característico de ocorrência possuir poucas semelhanças. A classe SX também apresentou confusão com a classe RY, possivelmente pelas semelhanças de relevo plano em que as duas tipicamente ocorrem.

Tabela 5 - Matriz de confusão da validação externa do mapa gerado com o conjunto de dados E0.

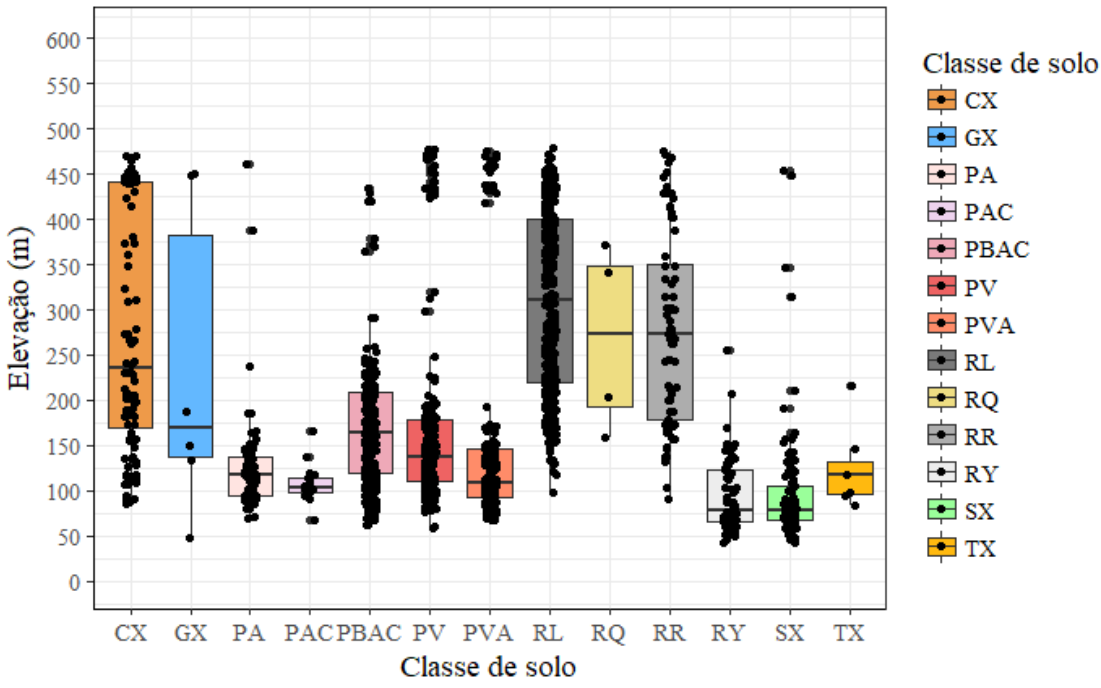
Classe	CX	GX	PA	PAC	PBAC	PV	PVA	RL	RQ	RR	RY	SX	TX
CX	**6**	0	0	0	0	0	0	0	0	0	0	0	0
GX	0	**0**	0	0	0	0	0	0	0	0	0	0	0
PA	0	0	**0**	0	0	1	0	0	0	0	0	0	0
PAC	0	0	0	**0**	0	0	0	0	0	0	0	0	0
PBAC	0	0	1	1	**12**	3	1	0	0	2	0	1	0
PV	0	0	3	1	1	**15**	3	2	1	0	0	6	1
PVA	0	0	0	0	2	1	**1**	0	0	0	0	2	0
RL	1	0	0	0	1	1	0	**11**	0	5	1	1	0
RQ	0	0	0	0	0	0	0	0	**0**	0	0	0	0
RR	0	0	0	0	0	0	0	0	0	**3**	0	0	0
RY	0	0	0	0	0	1	0	0	0	0	**2**	5	0
SX	0	1	0	0	0	1	1	0	0	0	6	**4**	0
TX	0	0	0	0	0	0	0	0	0	0	0	0	**1**
Ac. classe	0.93	0.50	0.50	0.50	0.83	0.73	0.56	0.87	0.50	0.65	0.58	0.56	0.75
Val. cruzada	0,59	Ac. Geral	0.49										

*Ac. Classe: acurácia balanceada de classe; Val. Cruzada: acurácia pela validação cruzada; Ac. Geral: acurácia pelo conjunto de validação externa; CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

Por outro lado, as classes GX, PA, PAC e RQ obtiveram os menores valores de acurácia, ambas com valor de 0,50. O baixo resultado da classe GX deve-se, possivelmente, ao fato dessa ocorrer em relevo com características semelhantes aos locais em que a classe SX comumente ocorre, ou seja, áreas planas e com encharcamento. De modo geral, essas classes diferenciam-se pelo fato de GX ocorrer em áreas com encharcamento permanente, o que dificulta a predição dessa classe por parte do modelo.

Uma possível causa dessa confusão por parte do modelo em mapear mesmo classes que comumente ocorrem em relevos distintos é demonstrada na Figura 8. Percebe-se que, embora PV ocorra em sua maioria em elevações maiores que PVA, essa regra não vale para as demais classes. Todos os Argissolos ocorrem em elevações muito semelhantes, ocorrendo até mesmo nas mesmas elevações que a maioria dos pontos de classes como CX, RL, RY e SX. A elevação em que ocorrem as classes ajuda explicar também a confusão criada pelo modelo entre as classes PV e SX, que, embora taxonomicamente distantes, parte de seus pontos ocorrem em elevações parecidas.

Figura 8 - Representação da elevação em que ocorrem as diferentes classes de solo do conjunto de dados E0.



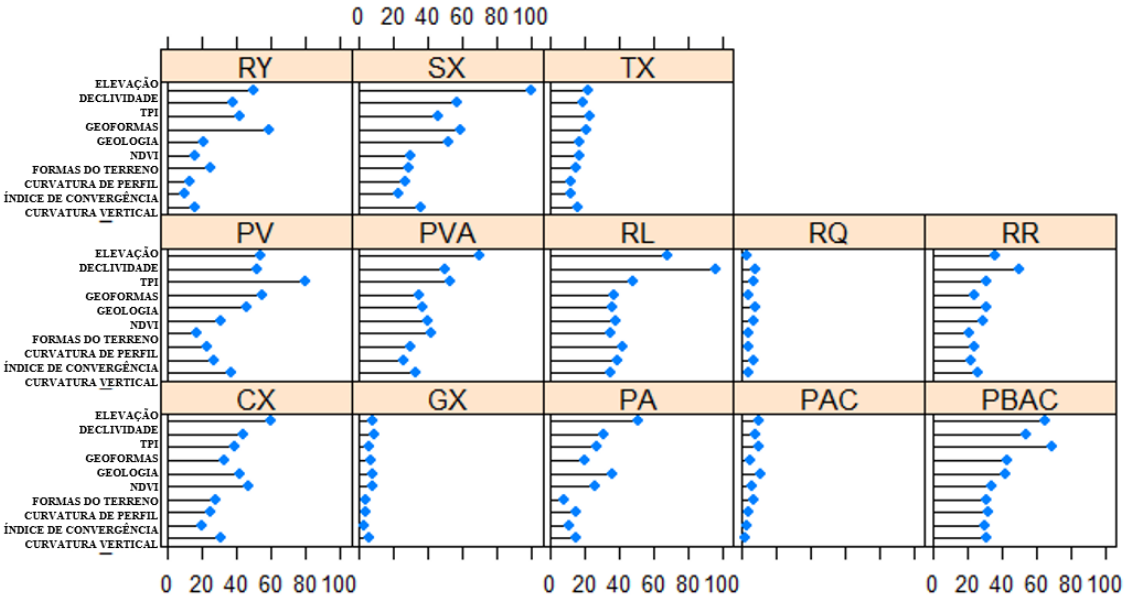
*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

É possível explicar também a semelhança entre as classes RL e RR e RY e SX, que apresentam grande confusão entre si e ocorrem em elevações semelhantes e, dada a importância da elevação (Figura 9) para o modelo preditivo, explica a confusão entre essas classes. Embora a elevação não seja a única covariável ambiental utilizada, a maior parte das covariáveis representando o relevo dentro do modelo "scorpan" é derivada dela.

O reflexo da baixa acurácia de algumas classes de solo também é demonstrado na Figura 9, que apresenta a importância que as covariáveis utilizadas pelo modelo E0 tiveram para cada classe em sua predição. A grande maioria das covariáveis tiveram baixa importância justamente nas classes GX, PA, PAC e RQ, ambas com baixa acurácia. Isso significa que as covariáveis não conseguiram distinguir o local de ocorrência dessas classes na paisagem.

Classes como GX e SX, que são morfologicamente semelhantes, apresentam comportamento distinto no que diz respeito a importância das covariáveis. Pela baixa ocorrência da classe GX quando comparada a SX, nenhuma das covariáveis utilizadas conseguiu, com destaque, determinar o local de ocorrência da classe GX na paisagem.

Figura 9 - Importância das 10 covariáveis mais utilizadas pelo modelo E0 nas classes preditas.



*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico; NDVI: índice de vegetação de diferença normalizada; TPI: índice de posição topográfica.

A baixa acurácia das classes PA e PAC pode ser explicada não só pela baixa quantidade de pontos dessas classes no conjunto de predição, mas também, conforme relatado por Meier et al. (2018), pela dificuldade de distinção da característica cor do solo em nível de paisagem, que, em segundo nível categórico, é o que diferencia essas classes. Dessa forma, mesmo as covariáveis que representem com fidelidade os fatores de formação do solo podem não auxiliar a distinguir tais classes na paisagem.

Há necessidade de ser considerada também a escala pouco detalhada do mapa geológico usado como covariável. Isso se torna relevante pelo fato da covariável geologia, que mesmo com pouco detalhamento, foi a quarta covariável mais importante para a predição do modelo gerado pelo conjunto de dados E0. Como a ocorrência de classes de solo, principalmente em nível de subordem, pode estar intimamente ligada ao material de origem (TERAMOTO et al, 2001), que na classe dos Argissolos pode condicionar diferentes classes de drenagem (PEDRON et al., 2006) e, consequentemente, diferentes subordens de Argissolos. Para a correta predição e obtenção de melhores resultados, principalmente dessa classe, seria fundamental a obtenção de informações geológicas em escala mais detalhada. Isso é demonstrado pelo fato da covariável geologia, mesmo em escala pouco detalhada, estar entre as mais importantes para o modelo preditivo, atrás apenas das covariáveis altitude, declividade e índice de posição topográfica. Os trabalhos de Barthold et al. (2013) e Adhkari et al. (2014) também afirmam que a geologia foi um dos mais importantes fatores que influenciaram a distribuição espacial das classes de solo.

5.3.2 Estratégia E1 - Predição com obtenção de pontos adicionais em mapas de classes de solo legados

Todos os modelos preditivos usando pontos adicionais amostrados nos mapas legados conseguiram mapear as mesmas 11 classes de solo apresentadas no mapa gerado pelo conjunto E0, não se diferenciando, sob esse aspecto, do conjunto de predição E0. Contudo, observa-se, em ambos os mapas gerados pelos conjuntos usando pontos adicionais coletados em mapas legados, uma menor capacidade de mapear algumas classes de solo. As classes de solo predominantes em todos os modelos da estratégia E1, que são apresentadas na Figura 10, foram SX, PV e RL. Houve uma redução da capacidade dos mapas do conjunto E1 representarem na paisagem as classes de solo PA, PBAC, PVA e RY classes essas com grande ocorrência no mapa gerado pelo conjunto E0. Por outro lado, as classes SX e PV,

predominantes nos mapas dos conjuntos de dados E1, apresentaram um aumento em suas áreas de ocorrência. Isso possivelmente seja resultado da mudança do percentual de cada classe dentro dos conjuntos de dados com a adição de pontos legados do mapa, apresentados na Tabela 4. É possível relacionar esses fatos pela redução do percentual de pontos nas classes PBAC, PVA e RY, que por sua vez perderam área de ocorrência, enquanto classes que tiverem seu número de pontos elevado pela adição de dados também tiveram sua área de ocorrência aumentada pelos modelos preditivos da estratégia E1.

Visualmente, percebe-se que os mapas gerados pelos conjuntos de dados da estratégia E1 foram muito mais generalistas no que diz respeito às classes PV e SX, tornando as áreas localizadas a leste e sudeste do mapa – onde haviam poucos pontos de calibração no conjunto E0 e somente dados de mapas legados treinaram o modelo no local (Figuras 3 e 4) – com pouco detalhamento de outras classes, como é comum observar em uma toposequência. Conforme Kempen et al. (2009), por vezes os modelos, apesar de estatisticamente sólidos, não apresentam uma representação pedológica plausível.

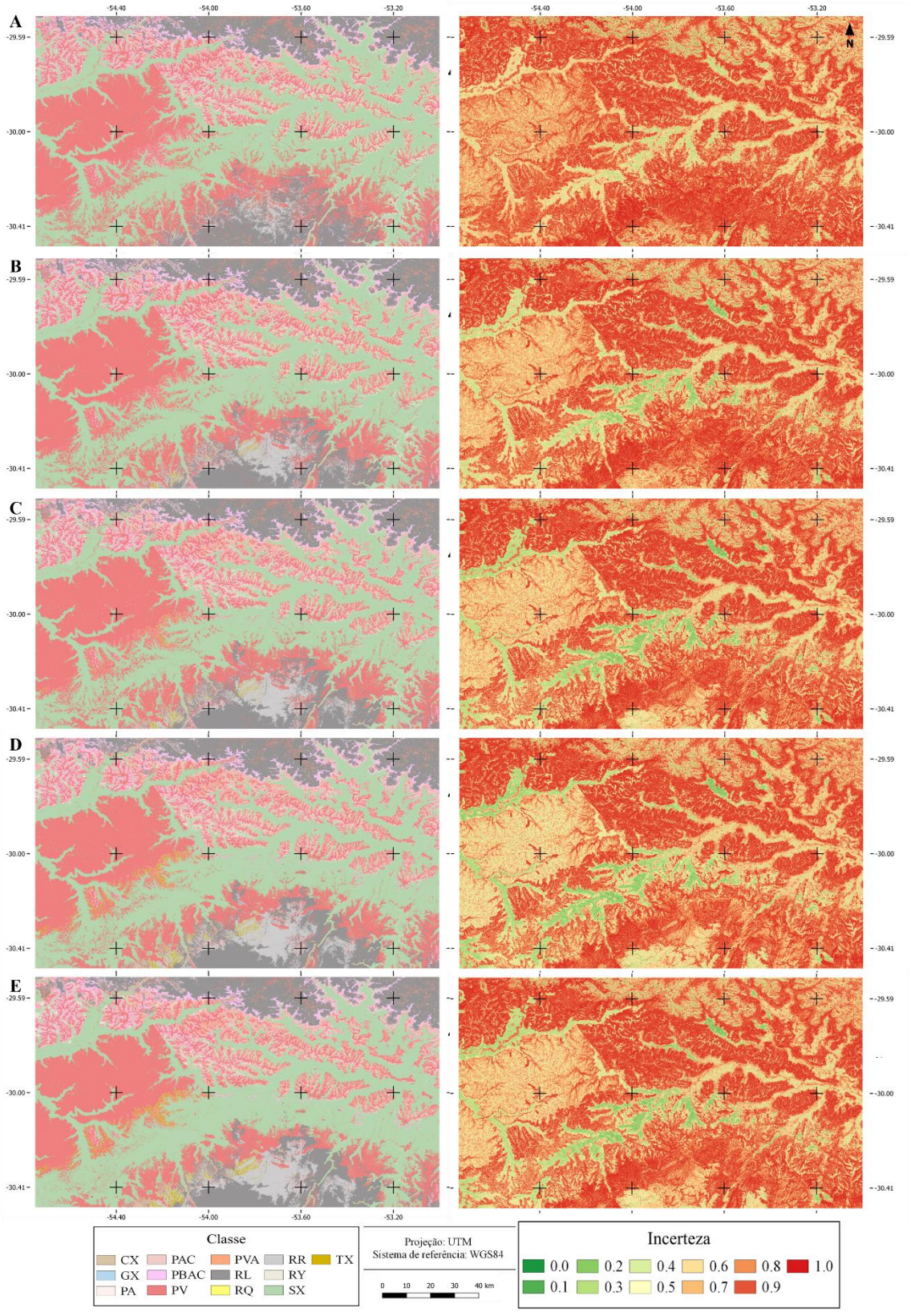
É possível verificar também um aumento nas áreas de RR e TX, ficando mais nítido esse aumento no mapa do conjunto de dados E1-2000 (Figura 10e). Isso também é um reflexo do aumento do número de pontos dessas classes, pouco presentes no conjunto de dados E0, porém mais expressivas nos conjuntos de dados da estratégia E1 devido a maior área – e consequentemente um maior número de pontos – nos mapas de solos legados. O aumento no número de dados proporciona ao modelo maior capacidade de representar sua ocorrência na paisagem pela maior possibilidade de formar relações entre as covariáveis e os dados, permitindo uma maior compreensão das relações solo-paisagem. Por outro lado, a presença de dados pode criar um número muito grande de relações, que por sua vez podem confundir o modelo ou criar uma incapacidade de representar as classes de solo na paisagem.

Os modelos gerados pelos conjuntos de dados da estratégia E1 obtiveram acurácia geral igual ou superior ao modelo gerado pelo conjunto de dados E0. Embora o modelo gerado pelo conjunto de dados E1-400 tenha obtido menor acurácia na validação cruzada comparada ao conjunto E0 (0,56 ante 0,59), na validação externa o resultado apresentou melhoria, com acurácia de 0,52 (ante 0,49 em E0). Houve redução também no valor de incerteza geral do mapa para 0,78, um valor considerável frente a média de 0,84 encontrada no mapa do conjunto E0.

O modelo gerado pelo conjunto de dados E1-800 foi o que apresentou o melhor resultado dentre os conjuntos da estratégia E1. Esse modelo obteve acurácia de 0,56 na validação cruzada e 0,54 na validação externa, tendo também sua incerteza geral reduzida

para 0,76. Os demais modelos da estratégia E1, com acréscimo no número de pontos, obtiveram desempenho abaixo do E1-800. O modelo gerado pelo conjunto de dados E1-1200 obteve acurácia de 0,56 na validação cruzada e 0,50 na validação externa. Embora a incerteza geral do modelo E1-1200 tenha reduzido para 0,74, a validação externa demonstra o resultado inferior desse modelo frente aos demais da estratégia E1. Os modelos gerados pelos conjuntos de dados E1-1600 e E1-2000 obtiveram desempenhos ainda menores, apresentando, respectivamente, acurácia de 0,57 e 0,57 na validação cruzada, 0,48 e 0,49 na validação externa, com incerteza geral de 0,73 e 0,72. Embora os valores de incerteza reduzam, para efeito de comparação devem ser observados os valores da validação externa, que reduziram em ambos os casos. Isso vai de encontro ao que afirmam Pahlavan-Rad et al., (2014), onde ressaltam a importância do mapa convencional do solo, indicando que são importantes covariáveis para o mapeamento digital do solo, e que esses mapas podem aumentar a precisão do modelo.

Figura 10 - (A) Mapas de classe de solo e mapas de incerteza gerados pelos conjuntos (A) E1-400, (B) E1-800, (C) E1-1200, (D) E1-1600 e (E) E1-2000.



*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

A Tabela 6 apresenta a matriz de confusão do conjunto de dados E1-800, que por sua vez obteve o melhor resultado de acurácia dentre os conjuntos da estratégia E1. Nesse conjunto, a classe PA apresentou acurácia de 0,63, melhoria frente ao valor de 0,50 apresentado no conjunto de dados E0. Apesar de, visualmente, a classe SX ter sido generalista na predição do mapa, a acurácia balanceada de classe apresentou valor de 0,68 (frente a 0,56 obtido em E0), demonstrando que o incremento de dados realizado pelo conjunto E1-800 foi relevante para o modelo melhor representar a variação das classes de solo na paisagem.

Tabela 6 - Matriz de confusão da validação externa do mapa gerado com o conjunto de dados E1-800.

	CX	GX	PA	PAC	PBAC	PV	PVA	RL	RQ	RR	RY	SX	TX
CX	**6**	0	0	0	0	0	0	0	0	0	0	0	0
GX	0	**0**	0	0	0	0	0	0	0	0	0	0	0
PA	0	0	**1**	0	0	0	0	0	0	0	0	0	0
PAC	0	0	0	**0**	0	0	0	0	0	0	0	0	0
PBAC	0	0	1	1	**12**	2	1	0	0	2	0	1	0
PV	0	0	2	1	2	**16**	4	1	1	0	0	7	1
PVA	0	0	0	0	1	1	**0**	0	0	0	0	0	0
RL	1	0	0	0	0	1	0	**12**	0	5	1	1	0
RQ	0	0	0	0	0	0	0	0	**0**	0	0	0	0
RR	0	0	0	0	1	1	0	0	0	**3**	0	0	0
RY	0	0	0	0	0	0	0	0	0	0	**1**	1	0
SX	0	1	0	0	0	2	1	0	0	0	7	**9**	0
TX	0	0	0	0	0	0	0	0	0	0	0	0	**1**
Ac. classe	0.93	0.50	0.63	0.50	0.83	0.74	0.49	0.92	0.50	0.64	0.55	0.68	0.75
Val. Cruzada	0,56	Ac. geral		0,54									

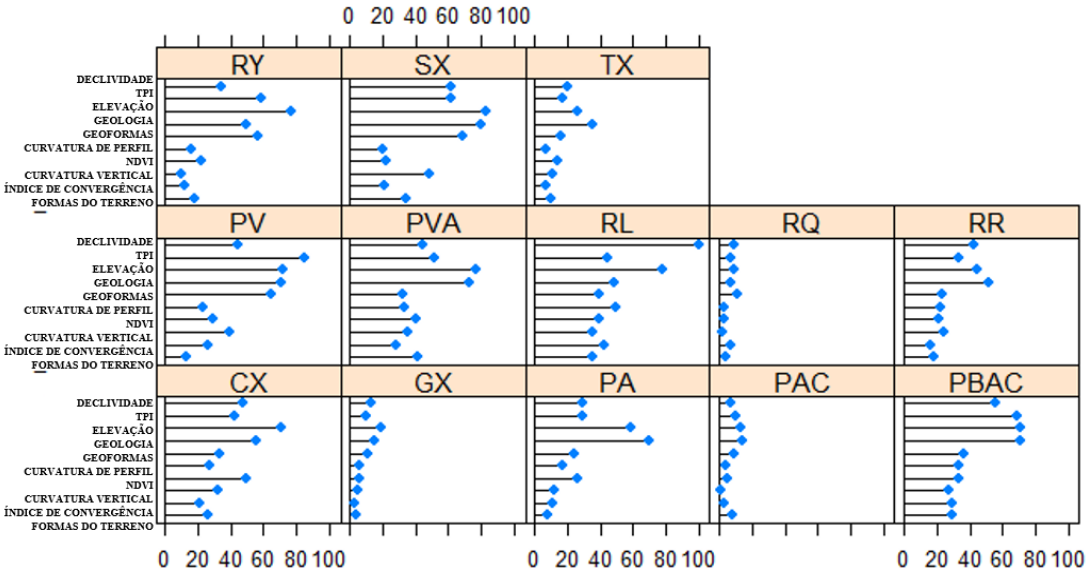
*Ac. Classe: acurácia balanceada de classe; Val. Cruzada: acurácia pela validação cruzada; Ac. Geral: acurácia pelo conjunto de validação externa; CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

O conjunto E1-800 também trouxe sensíveis aumentos na acurácia das classes PV, RL, RR e RY, trazendo ainda uma pequena redução na acurácia da classe PVA. Essa redução na acurácia da classe PVA possivelmente está relacionado a diminuição da área dessa no mapa E1-800, explicitando a relação da inserção de dados com o correto mapeamento das classes pelo modelo.

A inserção de novos dados em relação ao conjunto E0 também alterou as covariáveis ambientais com maior importância ao modelo de predição (Figura 11). Dessa vez, as

covariáveis declividade e TPI foram as mais importantes para o modelo, de modo que no modelo E0 foram elevação e declividade. Cabe salientar a presença da covariável geoformas nos modelos preditos, essa que também foi descrita por Jafari et al. (2012) como uma importante covariável preditora de classes de solo.

Figura 11 - Importância das 10 covariáveis mais utilizadas pelo modelo E1-800 nas classes preditas.



*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

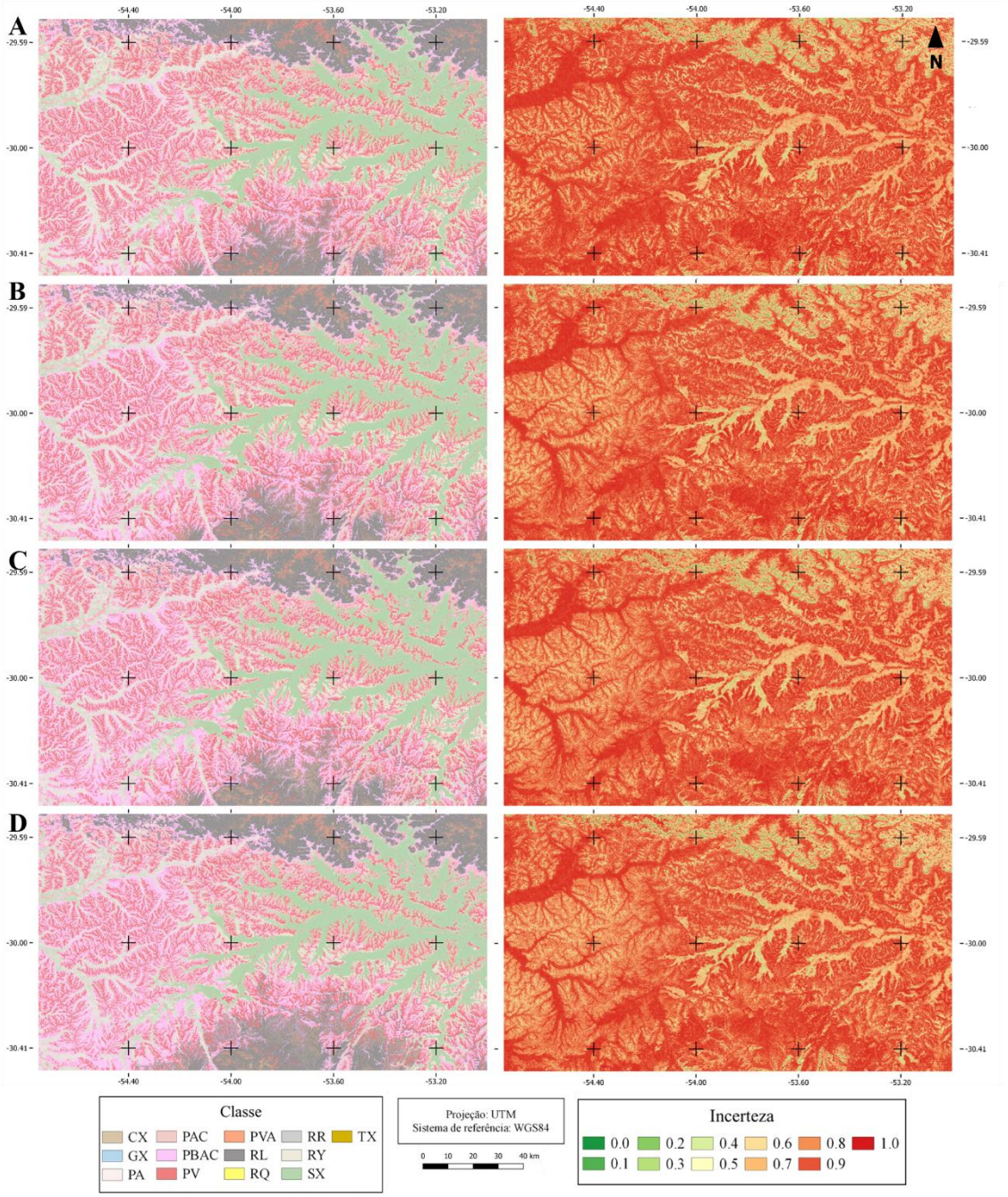
Novamente observa-se que as classes com baixa acurácia - GX, PAC e RQ – apresentam baixa importância na maioria das covariáveis que foram consideradas importantes pelo modelo, justificando sua baixa acurácia. Contudo, cabe salientar a situação da classe PA, que teve sua acurácia elevada no conjunto E1-800 em relação ao E0. Dessa vez, o aumento da acurácia na classe PA proporcionado pela inserção de dados está refletida na importância das covariáveis nessa classe. Geologia e elevação, nessa ordem, mostraram-se com importância elevada, demonstrando que um mapa geológico mais detalhado poderia trazer reflexos positivos não somente sobre uma ou outra classe, mas sobre todo o mapa de classes de solo.

5.3.3 Estratégia E2 - Predição com pontos adicionais reamostrados a campo pela incerteza

Os conjuntos de dados contendo pontos adicionais reamostrados conforme a incerteza do mapa E0 obtiveram desempenho semelhante à E0, conseguindo representar também 11 classes no mapa. Em ambos os modelos, as classes predominantes foram PBAC, PV, SX e RL. O aumento da área ocupada pela classe PBAC pode ser explicada pela maior amostragem dessa classe em relação ao conjunto E0 (Tabela 4). No entanto, as outras três classes que predominaram na área mantiveram os percentuais semelhantes aos observados no conjunto E0, sendo sua maior ocorrência devido a melhor compreensão por parte do modelo da relação solo paisagem dessas. Observando visualmente os mapas (Figura 12), é possível observar a melhor diferenciação de classes em relação aos mapas das estratégias E1, tendo uma distribuição mais coerente das classes dentro de uma toposequência, não apresentando generalizações de classes em partes do mapa.

No que diz respeito a acurácia dos mapas, os modelos gerados a partir dos dados da estratégia E2, obtiveram desempenho igual ou superior ao da estratégia E0. Contudo, eles diferenciam-se por apresentarem comportamento crescente linear com o número de pontos adicionados, sempre apresentando aumento da acurácia do modelo mesmo com a inserção de uma pequena quantidade de pontos. O conjunto E2-50 apresentou acurácia de 0,58 na validação cruzada e 0,49 na validação externa, com incerteza geral de 0,83 ao passo que os conjuntos E2-100 e E2-150 já apresentaram desempenho superior ao E0, ambos com acurácia geral de 0,57 na validação cruzada e 0,50 na validação externa, com incerteza geral de 0,82. Já o conjunto E2-200 foi o que obteve o melhor resultado, com acurácia de 0,56 na validação cruzada e 0,51 na validação externa, reduzindo a incerteza geral para 0,81. Embora com uma diferença considerada baixa entre a acurácia dos conjuntos de dados, percebe-se uma tendência de aumento com a inserção de mais dados, onde a maior acurácia foi obtida no conjunto de dados onde o maior número pontos foi adicionado ao conjunto de calibração.

Figura 12 - Mapas de classe de solo e mapas de incerteza gerados pelos conjuntos (A) E2-50, (B) E2-100, (C) E2-150 e (D) E2-200.



*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

Se analisados os números brutos, o incremento nos valores de acurácia da estratégia E2 podem ser considerados baixos frente aos da estratégia E1. Contudo, se considerados o número de pontos de cada estratégia, os resultados podem ser vistos sob uma outra

perspectiva. Embora o incremento gerado pela estratégia E2 usando reamostragem guiada pela incerteza esteja distante do obtido por Stumpf et al. (2017), com melhoria de 31% nos resultados, mesmo que não modelando o mesmo atributo do solo, os resultados podem ser considerados promissores. Considerando-se o pequeno número de pontos adicionais frente a grande área estudada, a quantidade de classes e subclasses, os resultados são considerados satisfatórios.

Os dados da matriz de confusão do conjunto de dados E2-200 (Tabela 7) demonstram que a adição de dados com reamostragem guiada pela incerteza resultou em melhoria na acurácia quando comparado ao conjunto E0. Os dados demonstram um sensível aumento na acurácia das classes PV (0,73 para 0,75), RY (0,58 para 0,59), SX (0,56 para 0,59) e PVA (0,56 para 0,65).

Tabela 7 - Matriz de confusão da validação externa do mapa gerado com o conjunto de dados E2-200

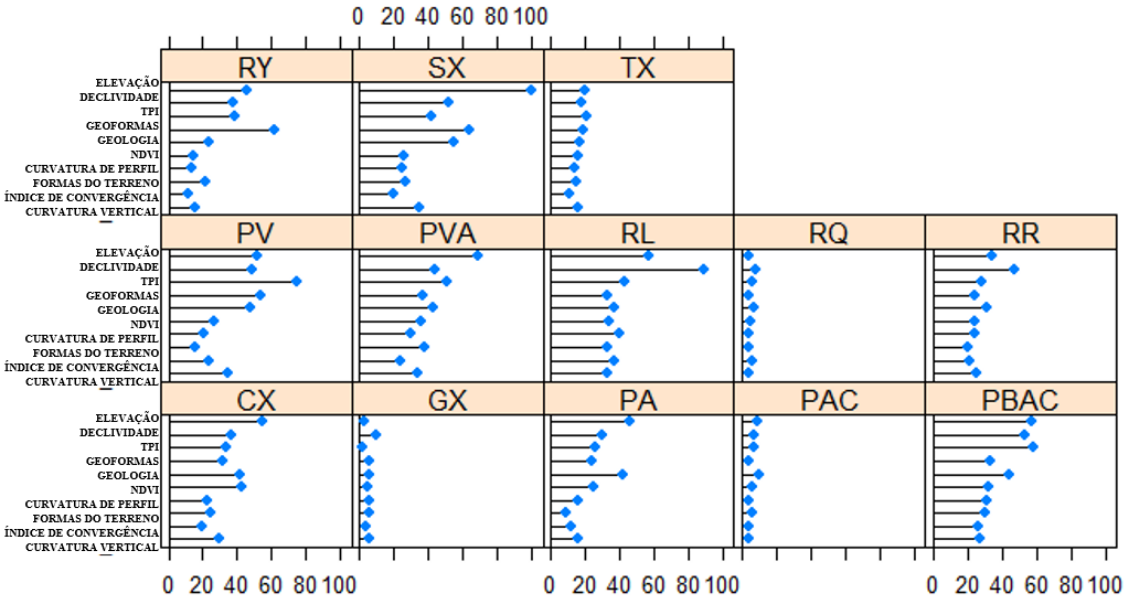
	CX	GX	PA	PAC	PBAC	PV	PVA	RL	RQ	RR	RY	SX	TX
CX	**6**	0	0	0	0	0	0	0	0	0	0	0	0
GX	0	**0**	0	0	0	0	0	0	0	0	0	0	0
PA	0	0	**0**	0	0	0	0	0	0	0	0	0	0
PAC	0	0	0	**0**	0	0	0	0	0	0	0	0	0
PBAC	0	0	1	1	**12**	3	1	0	0	2	2	2	0
PV	0	0	3	1	1	1**6**	2	2	1	0	0	6	1
PVA	0	0	0	0	1	1	**2**	0	0	0	0	1	0
RL	1	0	0	0	2	2	0	**11**	0	5	1	1	0
RQ	0	0	0	0	0	0	0	0	**0**	0	0	0	0
RR	0	0	0	0	0	0	0	0	0	**3**	0	0	0
RY	0	0	0	0	0	0	0	0	0	0	**2**	4	0
SX	0	1	0	0	0	1	1	0	0	0	4	**5**	0
TX	0	0	0	0	0	0	0	0	0	0	0	0	**1**
Ac. classe	0.93	0.50	0.50	0.50	0.81	0.75	0.65	0.86	0.50	0.65	0.59	0.59	0.75
Val. cruzada	0,56	Ac. geral	0.51										

*Ac. Classe: acurácia balanceada de classe; Val. Cruzada: acurácia pela validação cruzada; Ac. Geral: acurácia pelo conjunto de validação externa; CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

Observando o conjunto de covariáveis mais importantes para o modelo E2 – 200 (Figura 13), novamente são observadas similaridades entre os conjuntos de dados. Nas classes com baixa acurácia balanceada, tais como RQ, GX e PAC, nenhuma das covariáveis mais importantes para o modelo foi importante para as classes, demonstrando que, para essas classes, o conjunto de covariáveis ou o conjunto de dados não é capaz de formar relações que

represente o local de ocorrência das classes na paisagem. Tal situação pode se dar possivelmente pela baixa ocorrência dessas classes de solo na paisagem, o que dificulta sua amostragem e torna improvável sua predição pelo modelo.

Figura 13 - Importância das 10 covariáveis mais utilizadas pelo modelo E2-200 nas classes preditas.



*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

Embora com menor acurácia que o conjunto de dados E1-800, o modelo E2-200 pode ser considerado mais realista quando comparado ao modelo gerado com o conjunto de dados E0. Isso pode ser verificado pelo resultado da importância das covariáveis, onde as covariáveis utilizadas pelos modelos E0 e E2-200 foram as mesmas, apenas com variação em sua ordem de importância. Isso pode ser uma das explicações para a semelhança dos mapas dos conjuntos E0 e E2-200, que tem por afinidade terem sido gerados apenas com dados coletados a campo, sem pontos gerados em unidades de mapeamento de mapas legados. Embora os pontos gerados sobre os mapas legados fossem informados para o modelo como 80% de precisão, relações errôneas podem ser formadas a partir da inserção de dados que não tem, por natureza, total certeza da classe dentro da unidade taxonômica onde foi gerado o ponto amostral.

5.3.4 Estratégia E3 - Predição usando apenas pontos gerados em mapas de classes de solo legados

Em uma tentativa de verificar a viabilidade dos dados legados de mapas de classe de solo, o modelo gerado pelo conjunto de dados E3 foi o que mais apresentou diferenças ao conjunto de dados E0, fato resultante da natureza diferente dos dados. Apesar de também ter conseguido mapear as mesmas 11 classes que o mapa E0, a análise visual apresenta uma forte generalização causada pelos dados na região sul e sudeste do mapa (Figura 14). As classes predominantes foram SX e PV, com uma grande generalização dessas em seus locais de ocorrência, de modo que outras classes tiveram dificuldade de serem mapeadas nas proximidades das classes predominantes. Isso pode ser resultado do grande aumento da proporção de pontos da classe SX (Tabela 4) em detrimento da diminuição do número de pontos de classes que predominaram nos outros modelos preditivos. Os dados obtidos por Jafari et al. (2012) e Barthold et al. (2013) corroboram com essa afirmação, de modo que os autores afirmaram que classes de solo com frequências de amostragem mais altas tiveram maior área predita em seus estudos.

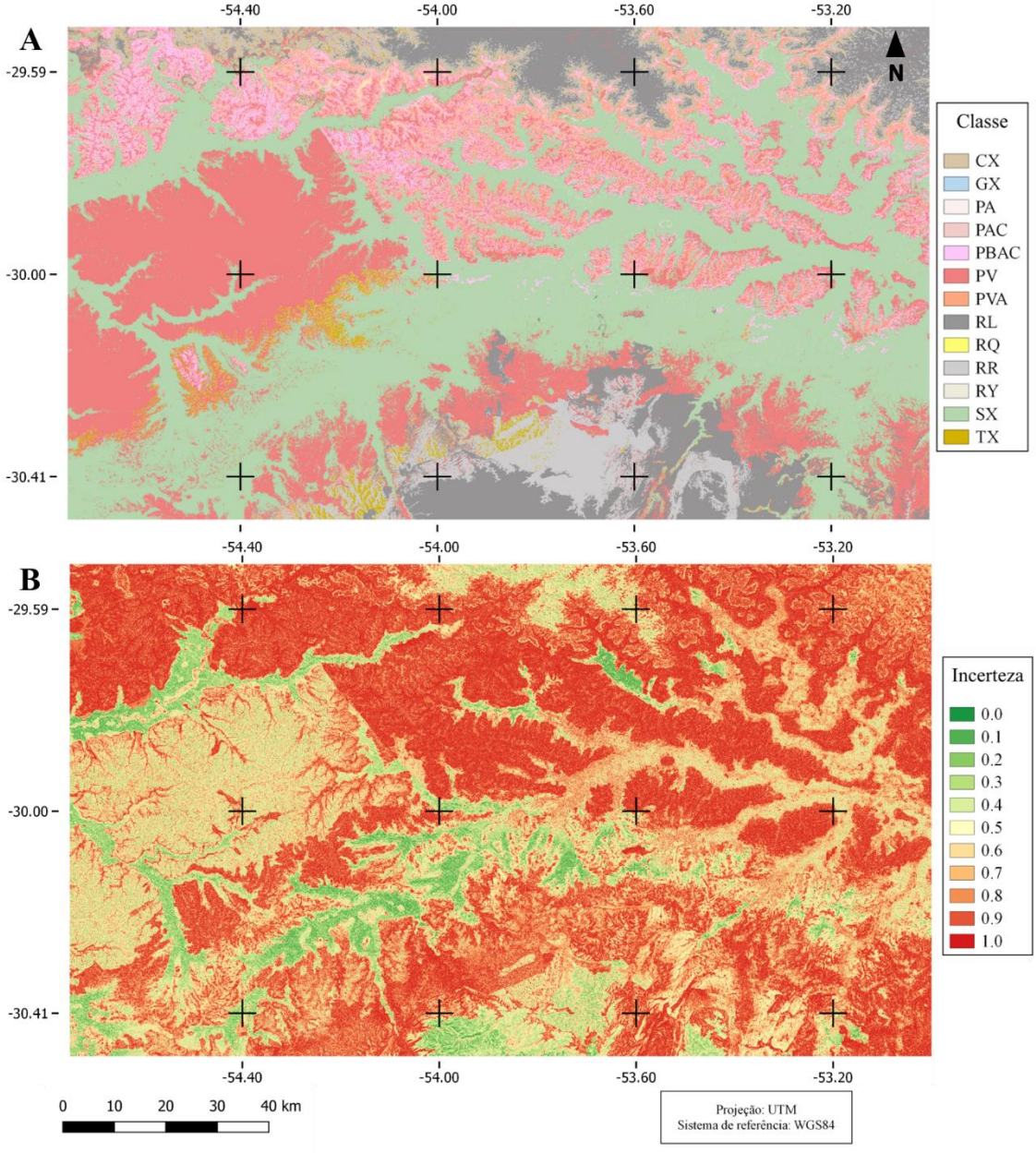
Tal generalização é demonstrada pelos resultados de acurácia do modelo E3. Apesar de ter obtido acurácia de 0,57 na validação cruzada e incerteza geral de 0,72, valores semelhantes aos demais modelos, a acurácia da validação externa foi de apenas 0,27. Isso demonstra que, apesar de o modelo ser robusto e conseguir classificar de forma satisfatória as informações que foram passadas a ele, essas informações não conseguem refletir com exatidão a real ocorrência de classes de solo a campo. Os valores de acurácia podem ser considerados baixos se comparado com a literatura. Heung et al. (2017), comparando modelos preditivos gerados a partir de dados de treinamento derivados de perfis obteve acurácia de 0,61 e usando dados de treinamento derivados de polígonos, como nessa estratégia, obteve acurácia de 0,68. A acurácia do presente estudo também diverge da encontrada por Heung et al (2016), que utilizando metodologia semelhante obteve acurácia de 0,58. Contudo, é necessário frisar que no presente estudo os mapas legados são obtidos de diferentes fontes, com diferentes escalas, além de apresentarem cobertura incompleta da área.

A grande parte dos trabalhos utiliza a técnica de desagregação de polígonos para predição e atualização de mapas de classes de solo, obtendo bons resultados de acurácia (ODGERS et al., 2014; SUBBURAYALU et al., 2014). Contudo, tais estudos fazem uso de informações muito mais refinadas, que na maioria das vezes não estão incorporadas ao mapa

apenas por questão de escala, mas disponíveis e detalhadas em relatórios. No caso do presente estudo, a escala pouco detalhada dos mapas de solo utilizados (HEUNG et al., 2017), além de não haver a cobertura total da área pelos mapas, limitaram os resultados.

Cabe salientar que o propósito original dos mapas de classe de solo legados não é servir como dado de calibração de um modelo preditivo, por isso o uso de suas informações deve ser ponderado e analisado quanto a confiabilidade. Heung et al. (2016) relataram também que métodos de amostragem diferentes podem gerar resultados mostrando padrões de solo inconsistentes com a compreensão dos solos da área de estudo.

Figura 14 - (A) Mapa de classe de solo e (B) mapa de incerteza gerados pelo conjunto E3.



*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL:

Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

Essa baixa capacidade de predição do conjunto E3 também pode ser observada na Tabela 8. Apesar de classes como PVA, RL e SX apresentarem resultados de acurácia balanceada adequados, observa-se maior erro na predição de classes como RL, RR e CX. O erro entre as classes citadas até seria aceitável, visto que o relevo em que ocorrem pode ser semelhante. Contudo, o modelo fez uma classificação errônea, por exemplo, entre CX e PBAC, que possuem características taxonômicas e morfológicas distintas.

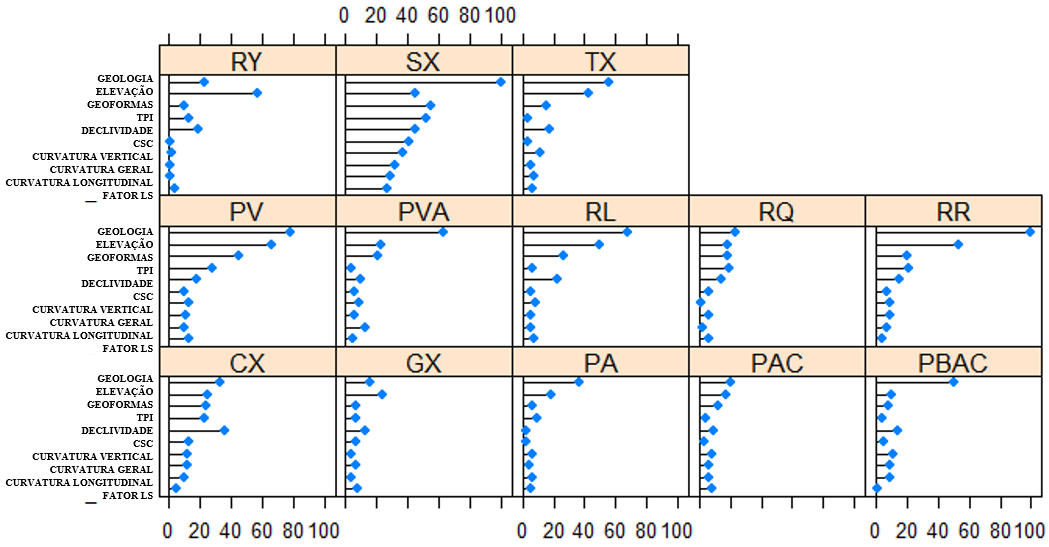
Tabela 8 - Matriz de confusão da validação externa do mapa gerado com o conjunto de dados E3

	CX	GX	PA	PAC	PBAC	PV	PVA	RL	RQ	RR	RY	SX	TX
CX	**2**	0	0	0	9	0	0	2	0	2	0	0	0
GX	0	**0**	0	0	0	0	0	0	0	0	0	0	0
PA	0	0	**0**	0	0	0	0	0	0	0	0	0	0
PAC	0	0	0	**0**	0	0	0	0	0	0	0	0	0
PBAC	0	0	1	1	**0**	7	1	1	0	2	1	2	0
PV	0	0	1	0	2	**7**	2	1	1	0	2	5	2
PVA	0	0	2	1	1	3	**2**	0	0	0	0	0	0
RL	4	0	0	0	0	2	0	**9**	0	5	0	2	0
RQ	0	0	0	0	0	0	0	0	**0**	0	0	0	0
RR	0	0	0	0	3	1	0	0	0	**1**	0	0	0
RY	0	0	0	0	0	0	0	0	0	0	**1**	1	0
SX	1	1	0	0	1	3	1	0	0	0	5	**9**	0
TX	0	0	0	0	0	0	0	0	0	0	0	0	**0**
Ac. de classe	0.58	0.50	0.50	0.50	0.42	0.56	0.63	0.78	0.50	0.53	0.55	0.67	0.50
Val. Cruzada	0,57	Ac. geral	0,27										

*Ac. Classe: acurácia balanceada de classe; Val. Cruzada: acurácia pela validação cruzada; Ac. Geral: acurácia pelo conjunto de validação externa; CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

Sob outra ótica, a baixa acurácia do modelo gerado pelo conjunto de dados E3 também pode ser observado pela importância de covariáveis às classes de solo (Figura 15). Embora grande parte das covariáveis utilizadas pelo modelo E3 serem as mesmas dos demais vistos até aqui, chama atenção o fato da geologia ser a mais importante. Isso pode ser um reflexo da forma como os mapas legados, que deram origem aos dados do conjunto E3 foram gerados, visto que, na época, eram poucas as informações sobre o solo e nesse cenário o mapa geológico pode ter sido uma das ferramentas para a geração dos mapas convencionais.

Figura 15 - Importância das 10 covariáveis mais utilizadas pelo modelo E3 nas classes preditas.



*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

Outro aspecto que pode ser observado que a maioria das covariáveis tidas pelo modelo E3 como mais importantes para a predição possuem baixos valores de importância dentro de cada classe de solo. Classes como PV, PVA e RL, que nos outros modelos mostravam todas as covariáveis com mais de 20 % de importância, no conjunto E3 não apresentam mais que quatro covariáveis com valor superior a esse.

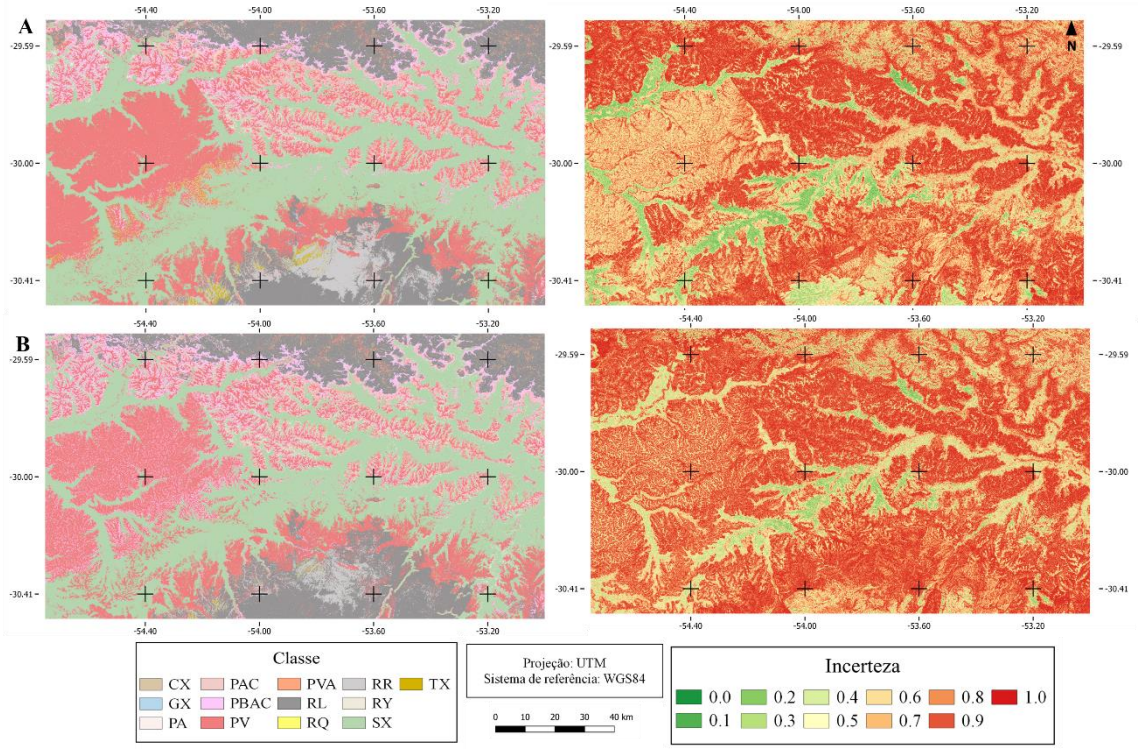
5.3.5 Estratégia E4 - Predição usando um banco de dados híbrido com as duas formas de obtenção de dados

O conjunto de dados E4-1, que possui o maior número de dados para o treinamento do modelo, conseguiu mapear as mesmas 11 classes dos demais modelos. Da mesma forma que os mapas gerados pelos conjuntos de dados E1 e E3, o modelo E4-1 apresentou predominância das classes SX, PV e RL. Já o modelo E4-2, que utilizou o número de pontos que obteve o melhor resultado de acurácia das estratégias E1 e E2, apresenta uma distribuição de classes menos discrepante, tendo predominância de SX, PV, RL e PBAC.

A análise visual dos mapas deixa clara a generalização realizada no mapa E4-1 (Figura 16a), com predominância de SX e PV nas regiões leste e sudeste do mapa, justamente nos

mesmos locais onde não haviam pontos advindos do conjunto E0 e apenas pontos de mapas de classe legados (estratégia E1) realizavam a calibração do modelo. Por outro lado, o mapa gerado pelo conjunto E4-2, que obteve uma distribuição mais homogênea das classes na paisagem, consegue mapear de forma mais correta as classes na paisagem.

Figura 6 - Mapa de classe de solo e mapa de incerteza gerados pelos conjuntos (A) E4-1 e (B) E4-2.



*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

Isso se reflete nos valores de acurácia dos dois modelos. O modelo gerado com os dados do conjunto E4-1 obteve valores de acurácia de 0,55 na validação cruzada e 0,48 na validação externa, com incerteza geral de 0,73. O modelo gerado com os dados do conjunto E4-2, por sua vez, obteve o melhor desempenho entre todos os conjuntos testados, com acurácia de 0,55 na validação cruzada e 0,55 também na validação externa, com incerteza geral de 0,77. Resultado da melhor distribuição de classes e menor percentual de SX no conjunto E4-2, quando comparado ao E4-1 (Tabela 4), esse obteve a melhor capacidade de predizer as classes de solo na paisagem. Isso representa que a grande quantidade de dados do conjunto E4-1 de certa forma enviesa o modelo preditivo, principalmente em regiões onde só havia informações

de mapas de solo legados, que por sua vez apresentavam apenas uma ou outra unidade de mapeamento simples, não possibilitando ao modelo expressar a real variação do solo na paisagem.

O bom resultado do conjunto E4-2 pode ser também demonstrado pela matriz de confusão do conjunto (Tabela 9). Apresentando valor de acurácia balanceada maior em classes comumente de difícil predição na paisagem como PA (0,63) e RR (0,69), além de resultados elevados em CX, PBAC, PV e RL, o modelo gerado apresentou resultado satisfatório frente aos demais.

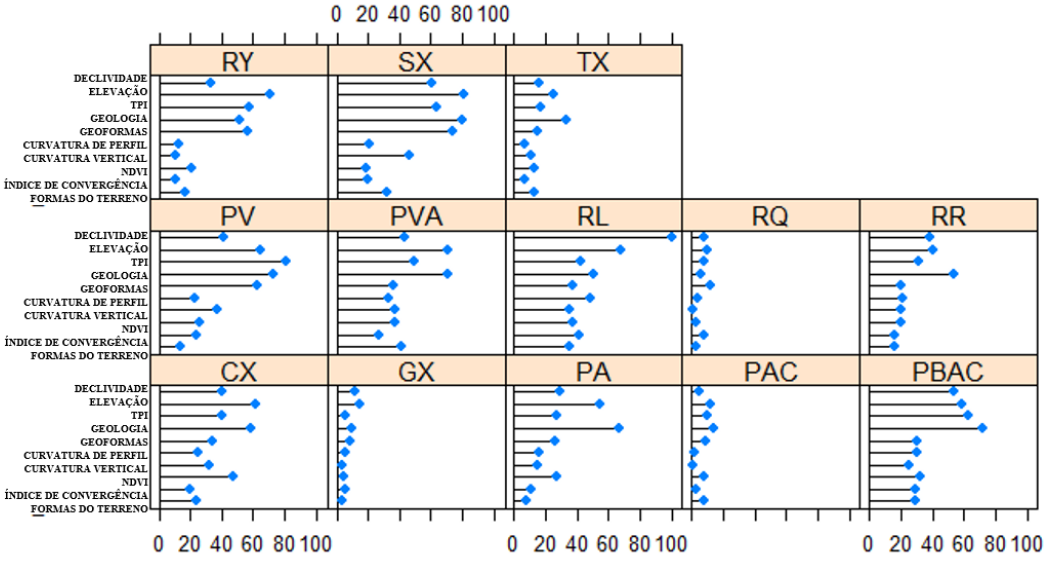
Tabela 9 - Matriz de confusão da validação externa do mapa gerado com o conjunto de dados E4-2.

	CX	GX	PA	PAC	PBAC	PV	PVA	RL	RQ	RR	RY	SX	TX
CX	**6**	0	0	0	0	0	0	0	0	0	0	0	0
GX	0	**0**	0	0	0	0	0	0	0	0	0	0	0
PA	0	0	**1**	0	0	0	0	0	0	0	0	0	0
PAC	0	0	0	**0**	0	0	0	0	0	0	0	0	0
PBAC	0	0	1	1	**13**	2	1	0	0	2	0	1	0
PV	0	0	2	1	1	**16**	4	1	1	0	0	7	1
PVA	0	0	0	0	1	1	**0**	0	0	0	0	0	0
RL	1	0	0	0	1	2	0	**12**	0	5	1	2	0
RQ	0	0	0	0	0	0	0	0	**0**	0	0	0	0
RR	0	0	0	0	0	0	0	0	0	**3**	0	0	0
RY	0	0	0	0	0	0	0	0	0	0	**1**	0	0
SX	0	1	0	0	0	2	1	0	0	0	7	**9**	0
TX	0	0	0	0	0	0	0	0	0	0	0	0	**1**
Ac. classe	0.93	0.50	0.63	0.50	0.87	0.75	0.49	0.90	0.50	0.65	0.56	0.69	0.75
Val. Cruzada	0,55	Ac. geral	0,55										

*Ac. Classe: acurácia balanceada de classe; Val. Cruzada: acurácia pela validação cruzada; Ac. Geral: acurácia pelo conjunto de validação externa; CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

As covariáveis ambientais de maior importância para o modelo E4-2 são apresentadas na Figura 17. Apenas com alteração na ordem, as covariáveis são as mesmas utilizadas pelo conjunto E0. Em ambos os modelos, a covariável declividade apresentou importância de 100% para a classe RL. Chama atenção também a importância da covariável geologia para ambos os modelos, sempre obtendo valores elevados de importância para diferentes classes.

Figura 17 - Importância das 10 covariáveis mais utilizadas pelo modelo E4-2 nas classes preditas.



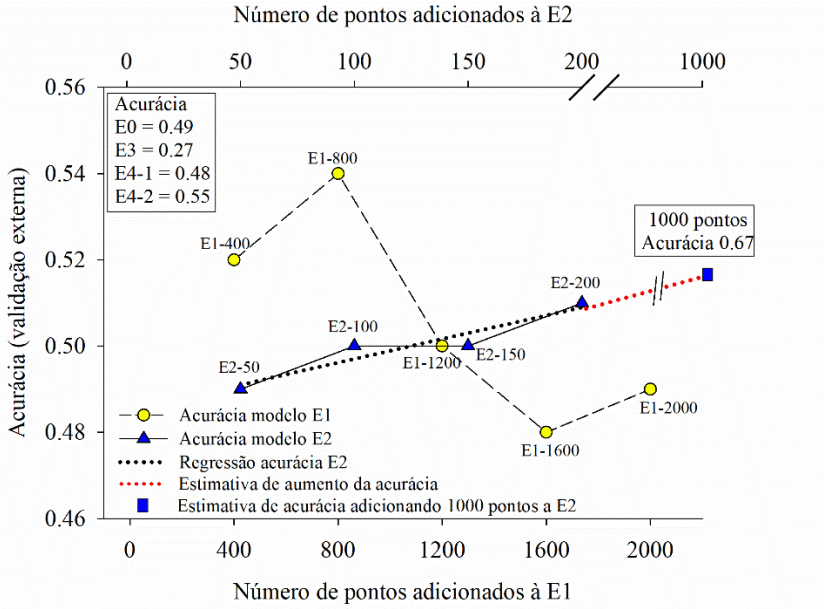
*CX: Cambissolo Háplico; GX: Gleissolo Háplico; PA: Argissolo Amarelo; PAC: Argissolo Acinzentado; PBAc: Argissolo Bruno-Acinzentado; PV: Argissolo Vermelho; PVA: Argissolo Vermelho-Amarelo; RL: Neossolo Litólico; RQ: Neossolo Quartzarênico; RR: Neossolo Regolítico; RY: Neossolo Flúvico; SX: Planossolo Háplico; TX: Luvissolo Háplico.

O conjunto E4-2 confirma que não somente a quantidade de dados, mas sim a qualidade deles é que tornarão os modelos preditores mais acurados para a predição de classes de solo. A imperfeita cobertura que o banco de dados E0 tinha sobre a área de estudo não permitia melhores resultados de acurácia. Contudo, foi demonstrado que a inclusão de pontos amostrais gerados sobre mapas legados deve ser utilizada com certa cautela. Esses, além de não possuírem escala detalhada e permitirem até 20% de inclusões em suas unidades de mapeamento, não foram gerados para esse fim. A inclusão de informações de mapas legados se torna benéfica ao modelo até certo ponto, mas a inserção de muitos dados de certa forma impõe ao modelo muitos dados com baixa precisão em locais específicos, não permitindo a correta interpretação do modelo da real ocorrência do solo na paisagem.

Como o intuito de demonstrar o comportamento da adição de dados às estratégias testadas, a Figura 18 apresenta de forma resumida os valores de acurácia obtido pelos modelos gerados pelos diferentes conjuntos de dados testados. A adição de um maior número de dados na estratégia E1 apresentou crescimento na acurácia até o conjunto E1-800, tendo decréscimo com a adição de mais dados. Já o conjunto E2 apresentou comportamento linear nos valores de acurácia com a adição de mais dados ao modelo. Devido a essa linearidade, foi calculada a partir desses valores de acurácia uma regressão linear com intuito de estimar o

número de pontos necessários para o aumento considerável de acurácia. Como resultado, a adição de 1000 pontos reamostrados com base na incerteza (estratégia E2) resultaria, com base em uma estimativa sobre os resultados obtidos, em um mapa com acurácia de 0,67.

Figura 18 - Acurácia obtida pela validação externa a partir da adição de pontos aos conjuntos das estratégias E1 e E2 e estimativa de acurácia com a adição de 1000 pontos à estratégia E2.



Salienta-se a grande margem de erro que possivelmente esteja embutida nesse resultado, visto que a complexidade da distribuição das classes de solo a campo está relacionada com fatores que podem ir além da compreensão do modelo, estando condicionado até mesmo à obtenção de covariáveis com maior detalhamento. Contudo, os resultados demonstram o potencial da reamostragem guiada pela incerteza como forma de obtenção de mais dados para a predição de classes de solo utilizando dados legados. A obtenção de novos dados a partir da reamostragem guiada pela incerteza, apesar de mais onerosa e trabalhosa, consegue capturar de melhor forma as variações do solo na paisagem, trazendo ao modelo informações de maior qualidade. A união de estratégias, contudo, se mostrou eficiente para predizer classes de solo na área estudada, principalmente por permitir obter informações de forma rápida sobre uma grande área em que não haviam dados.

5.4 CONCLUSÕES

Os dados legados, embora com uma distribuição irregular na área e com um grande número de classes, apresentaram resultado satisfatório ao mapear uma área com grandes proporções, obtendo acurácia de 0,49 e incerteza geral de 0,84.

A reamostragem guiada pela incerteza apresentou potencial para melhoria do mapa, aumentando a acurácia para 0,51 e reduzindo a incerteza geral para 0,81, mesmo com um pequeno número de amostras adicionais incorporadas ao modelo.

A obtenção de pontos amostrais em unidades de mapeamento de mapas legados, apesar de ter trazido benefícios ao modelo preditivo, demonstrou inconsistências devido a sua escala pouco detalhada, de modo que o modelo gerado apenas pelo conjunto de dados legados do mapa, além de apresentar baixa acurácia, foi incapaz de capturar as diferenças entre as classes de solo na paisagem.

Esses resultados credenciam o MDS a servir de apoio ao PronaSolos, uma vez que foi possível mapear classes de solo de uma grande área apenas com dados legados. Além disso, a inserção de dados de uma forma rápida e de baixo custo proporcionou melhorias na qualidade dos mapas.

5.5 REFERÊNCIAS

ADHIKARI, K.; MINASNY, B.; GREVE, M. B.; GREVE, M. H. Constructing a soil class map of Denmark based on the FAO legend using digital techniques. **Geoderma**, v. 214, p. 101-113. 2014.

ALVARES, C. A. et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711-728, 2013.

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

BARTHOLD, F. K.; WIESMEIER, M.; BREUER, L.; FREDE, H. G.; WU, J.; BLANK, F. B. Land use and climate control the spatial distribution of soil types in the grasslands of Inner Mongolia. **Journal of Arid Environments**, v. 88, p. 194-205, 2013.

BRUNGARD, C. W.; BOETTINGER, J. L.; DUNIWAY, M. C.; WILLS, S. A.; EDWARDS JR, T. C. Machine learning for predicting soil classes in three semi-arid landscapes. **Geoderma**, v. 239, p. 68-83, 2015.

BURROUGH, P. A.; VAN GAANS, P. F.; HOOTSMANS, R. Continuous classification in soil survey: spatial correlation, confusion and boundaries. **Geoderma**, v. 77, n. 2-4, p. 115-135, 1997.

CONGALTON, R. G. A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. **Remote Sensing Environment**, v. 37, n. 1, p. 35-46, 1991.

CONRAD, O.; BECHTEL, B.; BOCK, M.; DIETRICH, H.; FISCHER, E.; GERLITZ, L.; WEHBERG, J.; WICHMANN, V.; BÖHNER, J. System for automated geoscientific analyses (SAGA) v. 2.1.4. **Geosci Model Dev**, v. 8, p. 1991-2007, 2015

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

HARTEMINK, A.E., et al. GlobalSoilMap.net — A New Digital Soil Map of the World. In: Boettinger, J.L., et al. (Ed.), **Digital Soil Mapping**. Springer, Dordrecht, p. 423 – 427, 2010.

HEUNG, B.; BULMER, C. E.; SCHMIDT, M. G. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. **Geoderma**, v. 214, p. 141-154, 2014.

HEUNG, B.; HO, H. C.; ZHANG, J.; KNUDBY, A.; BULMER, C. E.; SCHMIDT, M. G. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. **Geoderma**, v. 265, p. 62-77, 2016.

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**, v. 290, p. 51-68, 2017.

HOUNKPATIN, K. O.; SCHMIDT, K.; STUMPF, F.; FORKUOR, G.; BEHRENS, T.; SCHOLTEN, T.; AMELUNG, W.; WELP, G. Predicting reference soil groups using legacy data: A data pruning and Random Forest approach for tropical environment (Dano catchment, Burkina Faso). **Scientific reports**, v. 8, n. 1, p. 9959, 2018.

JAFARI, A.; AYOUBI, S.; KHADEMI, H.; FINKE, P. A.; TOOMANIAN, N. Selection of a taxonomic level for soil mapping using diversity and map purity indices: a case study from an Iranian arid region. **Geomorphology**, v. 201, p. 86-97, 2013.

JAFARI, A.; FINKE, P. A.; VANDE WAUW, J.; AYOUBI, S.; KHADEMI, H. Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. European Journal of Soil **Science**, v. 63, n. 2, p. 284-298, 2012.

JEUNE, W.; FRANCELINO, M. R.; SOUZA, E.; INÁCIO, E. Multinomial Logistic Regression and Random Forest Classifiers in Digital Mapping of Soil Classes in Western Haiti. **Revista Brasileira de Ciência do Solo**, v. 42, e0170133, 2018.

KEMPEN, B.; BRUS, D. J.; HEUVELINK, G. B.; STOORVOGEL, J. J. Updating the 1: 50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. **Geoderma**, v. 151, n. 3-4, p. 311-326, 2009.

KLAMT, E.; DALMOLIN, R.S.D; CABRAL, D.R. **Solos do Município de São João do Polêsine: classificação, distribuição geográfica e aptidão de uso**. Santa Maria: Centro de Ciências Rurais, Departamento de Solos, 1997. 93p.

KROL, B. G. C. M. Towards a data quality management framework for digital soil mapping with limited data. In: AHRENS, R. J. **Digital soil mapping with limited data**, p. 137-149. Springer, Dordrecht, 2008.

KUHN M. Caret: classification and regression training. R package version 6.0-76; 2017. Available from: https://CRAN.R-project.org/package=caret.

MEIER, M.; SOUZA, E. D.; FRANCELINO, M. R.; FERNANDES FILHO, E. I.; SCHAEFER, C. E. G. R. Digital Soil Mapping Using Machine Learning Algorithms in a Tropical Mountainous Area**. Revista Brasileira de Ciência do Solo**, v. 42, 2018.

MINASNY, B.; BISHOP, T. F. A. **Analysing uncertainty**. Guidelines for surveying soil and resources. p. 383-393, 2008.

ODGERS, N. P.; SUN, W.; MCBRATNEY, A. B.; MINASNY, B.; CLIFFORD, D. Disaggregating and harmonising soil map units through resampled classification trees. **Geoderma**, v. 214, p. 91-100, 2014.

OMUTO, C.; NACHTERGAELE, F.; ROJAS, R. V. **State of the art report on global and regional soil information: Where are we? Where to go?** Global Soil Parthership Technical Report. Roma: FAO, 2013. 69p.

PAHLAVAN-RAD, M. R.; TOOMANIAN, N.; KHORMALI, F.; BRUNGARD, C. W.; KOMAKI, C. B.; BOGAERT, P. Updating soil survey maps using random forest and conditioned Latin hypercube sampling in the loess derived soils of northern Iran. **Geoderma**, v. 232, p. 97-106, 2014.

PEDRON, F.A.; DALMOLIN, R.S.D.; AZEVEDO, A.C.; POELKING, E.L. & MIGUEL, P. Utilização do sistema de avaliação do potencial de uso urbano das terras no diagnóstico ambiental do município de Santa Maria - RS. **Ciência Rural**, v. 36, p. 468-477, 2006.

PELEGRINO, M. H. P.; SILVA, S. H. G.; MENEZES, M. D. D.; SILVA, E. D.; OWENS, P. R.; CURI, N. Mapping soils in two watersheds using legacy data and extrapolation for similar surrounding areas. **Ciência e Agrotecnologia**, v. 40, n. 5, p. 534-546, 2016.

PINHEIRO, H. S. K.; CARVALHO JUNIOR, W. D.; CHAGAS, C. D. S.; ANJOS, L. H. C. D.; OWENS, P. R. Prediction of Topsoil Texture Through Regression Trees and Multiple Linear Regressions. **Revista Brasileira de Ciência do Solo**, v. 42, 2018.

POLIDORO JC et al. **Programa Nacional de Solos do Brasil (PronaSolos)**. Rio de Janeiro: Embrapa Solos, 2016, 54 p.

QGIS DEVELOPMENT TEAM et al. 2018. QGIS geographic information system. Open Source Geospatial Foundation Project. Available at: http://qgis.osgeo.org. Accessed on July 12, 2018.

R CORE TEAM. R: A language and environment for statistical computing. Vienna: 2016.

ROBINSON D. A.; PANAGOS, P.; BORRELLI, P.; JONES, A.; MONTANARELLA, L.; TYE, A.; OBST, C. G. Soil natural capital in Europe; a framework for state and change assessment. **Scientific Reports**, v. 7, n. 1, p. 6706, 2017.

SANTOS, H. G. et al. **Sistema brasileiro de classificação de solos**. 3ª Ed. rev. e ampl. Brasília, DF: Embrapa, 2013. 353 p.

SARTORI, P. L. P. Geologia e geomorfologia de Santa Maria. **Ciência e Ambiente**, v. 38, p. 19-42, 2009.

SOIL SURVEY STAFF, NATURAL RESOURCES CONSERVATION SERVICE, UNITED STATES DEPARTMENT OF AGRICULTURE. **Web Soil Survey**, 2019. Disponível em: https://websoilsurvey.nrcs.usda.gov/. Acesso em 08 de janeiro de 2019.

STRECK, E. V.; KÄMPF, N.; DALMOLIN, R. S. D.; KLAMT, E.; NASCIMENTO, P. C.; GIASSON, E.; PINTO, L. F. S. **Solos do Rio Grande do Sul**. Porto Alegre: UFRGS, EMATER/RS-ASCAR, 3ª Edição, 2018. 251 p.

STUMPF, F.; SCHMIDT, K.; BEHRENS, T.; SCHÖNBRODT-STITT, S.; BUZZO, G.; DUMPERTH, C.; SCHOLTEN, T. Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. **Journal of Plant Nutrition and Soil Science**, v. 179, n. 4, p. 499-509, 2016.

STUMPF, F.; SCHMIDT, K.; GOEBES, P.; BEHRENS, T.; SCHÖNBRODT-STITT, S.; WADOUX, A.; SCHOLTEN, T. Uncertainty-guided sampling to improve digital soil maps. **Catena**, v. 153, p. 30-38, 2017.

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**, v. 213, p. 334-345, 2014.

SULAEMAN, Y.; MINASNY, B.; MCBRATNEY, A. B.; SARWANI, M.; SUTANDI, A. Harmonizing legacy soil data for digital soil mapping in Indonesia. **Geoderma**, v. 192, p. 77-85, 2013.

TEN CATEN, A.; DALMOLIN, R. S. D.; RUIZ, L. F. C. Digital soil mapping: strategy for data pre-processing. **Revista Brasileira de Ciência do Solo**, v. 36, n. 4, p. 1083-1092, 2012.

TERAMOTO, E. R.; LEPSCH, I. F.; VIDAL-TORRADO, P. Relações solo, superfície geomórfica e substrato geológico na microbacia do ribeirão Marins (Piracicaba-SP). **Scientia Agricola**, v. 58, n. 2, p. 361-371, 2001.

VALERIANO M. M. **Topodata: guia para utilização de dados geomorfológicos locais**. São José dos Campos: INPE; 2008.

VINCENT, S.; LEMERCIER, B.; BERTHIER, L.; WALTER, C. Spatial disaggregation of complex Soil Map Units at the regional scale based on soil-landscape relationships. **Geoderma**, v. 311, p. 130-142, 2018.

WILSON, J. P.; GALLANT, J. C. Digital terrain analysis. In: Wilson JP, Gallant JC, Eds. **Terrain analysis: principles and applications**. New York: John Wiley; 2000. p. 1-28.

WOLSKI, M. S.; DALMOLIN, R. S. D.; FLORES, C. A.; MOURA-BUENO, J. M.; TEN CATEN, A.; KAISER, D. R. Digital soil mapping and its implications in the extrapolation of soil-landscape relationships in detailed scale. **Pesquisa Agropecuária Brasileira**, v. 52, n. 8, p. 633-642, 2017.

ZERAATPISHEH, M.; AYOUBI, S.; BRUNGARD, C. W.; FINKE, P. Disaggregating and updating a legacy soil map using ART, fuzzy c-means and k-means clustering algorithms in Central Iran. **Geoderma**, v. 340, p. 249-258, 2019.