UFRRJ

INSTITUTO DE AGRONOMIA

PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA - CIÊNCIA DO SOLO

TESE

Mapeamento Digital de Atributos do Solo em Áreas Remotas sob Floresta Amazônica: Um Estudo de Caso na Formação Solimões

Ana Carolina de Souza Ferreira

2022

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE AGRONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA
CIÊNCIA DO SOLO

MAPEAMENTO DIGITAL DE ATRIBUTOS DO SOLO EM ÁREAS REMOTAS SOB FLORESTA AMAZÔNICA: UM ESTUDO DE CASO NA FORMAÇÃO SOLIMÕES

ANA CAROLINA DE SOUZA FERREIRA

Sob orientação do Professor
Marcos Bacis Ceddia

Tese submetida como requisito parcial para obtenção do grau de **Doutora**, no Programa de Pós-Graduação em Agronomia – Ciência do solo, Área de Concentração em Pedologia e Física do solo.

Seropédica, RJ
Fevereiro de 2022

Universidade Federal Rural do Rio de Janeiro
Biblioteca Central / Seção de Processamento Técnico

Ficha catalográfica elaborada
Com os dados fornecidos pelo(a) autor(a)

F383m

Ferreira, Ana Carolina de Souza, 1987-
Mapeamento digital de atributos do solo em áreas remotas sob Floresta Amazônica: um estudo de caso na Formação Solimões / Ana Carolina de Souza Ferreira. – Seropédica, 2022.
99 f. : il.

Orientador: Marcos Gervasio Pereira.
Tese (Doutorado). – – Universidade Federal Rural do Rio de Janeiro, Programa de Pós-Graduação em Agronomia Ciência do Solo, 2022.

1. Aprendizado de máquina. 2. Mapeamento digital de solo. 3. Área de referência. I. Ceddia, Marcos Bacis, 1968-, orient. II Universidade Federal Rural do Rio de Janeiro. Programa de Pós-Graduação em Agronomia-Ciência do Solo III. Título.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE AGRONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA - CIÊNCIA DO SOLO

ANA CAROLINA DE SOUZA FERREIRA

Tese submetida como requisito parcial para obtenção do grau de **Doutora,** no Programa de Pós-Graduação em Agronomia – Ciência do Solo, Área de Concentração em Pedologia e Física do Solo.

TESE APROVADA EM 28/02/2022.

_____
Marcos Bacis Ceddia. Dr. UFRRJ
(Orientador)

_____
Helena Saraiva Koenow Pinheiro. Dra. UFRRJ

_____
Waldir de Carvalho Júnior. Dr. Embrapa Solos

_____
Elias Mendes Costa. Dr. UFRRJ

_____
Diego Nunes Brandão Dr. CEFET/RJ

A Deus, minha família e amigos, por todas as palavras que me deram força para seguir em frente nos momentos difíceis, obrigada pelo apoio, paciência e confiança.

Ofereço.

À minha mãe, Maristela, pela minha vida, amor, paciência e dedicação durante minha jornada, me aconselhando e acreditando no meu sucesso. A ela eu devo todo meu amor e carinho.

Dedico.

AGRADECIMENTOS

Agradeço primeiramente a Deus, que está presente em todos os momentos da minha vida, iluminando minha jornada que por vezes se tornou árdua, mas ao meu lado ele me deu força, esperança, saúde e coragem para enfrentar todos os obstáculos.

A minha mãe e meu padrasto Maristela Guimarães de Souza e Maickel Barbosa Oliveira por todo o carinho, paciência e amor, que me deram educação e que nunca mediram esforços para que eu continuasse meus sonhos, sempre dispostos a me ajudar independentemente de qualquer situação.

A todos os meus familiares que de alguma forma contribuíram para que este sonho se tornasse realidade.

Ao meu namorado Daniel Costa, uma pessoa muito especial que esteve presente ao meu lado nas horas difíceis, sendo compreensível, me confortando, ajudando e sendo meu companheiro nos problemas surgidos no decorrer de todo o meu trabalho.

A Universidade Federal Rural do Rio de Janeiro, que me ofereceu a oportunidade de realizar esse curso, ao PPGA-CS e a todos os professores que puderam contribuir para meu aprendizado.

Ao Professor Marcos Bacis Ceddia pela orientação, incentivo, paciência, confiança e amizade, sempre solícito durante todo o trabalho.

Aos amigos que fizeram parte da minha jornada, pelos momentos divertidos, tristes, engraçados, que pudemos passar juntos e que de alguma forma fazem parte em nossas vidas.

Às amigas, Sandra Lima, Andréa Silva Gomes, Priscila pela confiança, amor carinho e apoio. Vocês são pessoas muito especiais que ficaram marcadas na minha vida.

Ao meu amigo Igor Leite, pela ajuda, amizade e carinho.

Ao meu amigo Elias, pela ajuda, amizade todo apoio no decorrer dessa caminhada.

A toda equipe LASA, as análises de laboratório, aos amigos, cada um teve um papel nesse trabalho.

Aos funcionários da secretaria do PPGA-CS, Marquinhos, Michele e Wagner pela atenção e gentileza.

Ao Programa de Pós-graduação em Agronomia – Ciência do Solo da UFRRJ, a CAPES pela concessão das bolsas de estudos. E a linda e encantadora UFRRJ por ter me dado a oportunidade de evoluir profissionalmente e como ser humano.

Ao povo brasileiro que custeou meus estudos.

Enfim a todos que contribuíram para conclusão desse trabalho.

Meus sinceros agradecimentos!

BIOGRAFIA

Ana Carolina de Souza Ferreira, filha de Ricardo Luiz Lopes Ferreira e Maristela Guimarães de Souza, nasceu em 29 de julho de 1987 na cidade de Volta Redonda, Estado do Rio de Janeiro. Ingressou na Universidade Federal Rural do Rio de Janeiro em 2006 no curso de Agronomia, formando-se em janeiro de 2013. Em 2013 foi bolsista de aperfeiçoamento em trabalhos na área de solos, atuando nos temas de geoprocessamento, sensoriamento remoto e pedologia. Em março de 2014 iniciou o mestrado no Curso de Pós-Graduação em Agronomia – Ciência do Solo pela Universidade Federal Rural do Rio de Janeiro concluindo-o em 2016. Durante o mestrado foi bolsista NOTA 10 da FAPERJ por 12 meses. Ingressou no Curso de Especialização em Estatística Aplicada (*latu sensu*) da UFRRJ em março de 2015 concluindo-o em setembro de 2016. Em 2018 iniciou o doutorado no Curso de Pós-Graduação em Agronomia – Ciência do Solo pela Universidade Federal Rural do Rio de Janeiro concluindo-o em fevereiro de 2022.

RESUMO GERAL

FERREIRA, Ana Carolina de Souza. **Mapeamento digital de atributos do solo em áreas remotas sob Floresta Amazônica: Um estudo de caso na Formação Solimões** 2022. 99f. Tese (Doutorado em Agronomia - Ciência do Solo). Instituto de Agronomia, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2022.

Atualmente a procura pelo mapeamento digital de solos (MDS) tem crescido consideravelmente nos estudos ambientais. Diante dessas demandas e a falta de informações sobre os solos em escala adequada, é mais que necessário o desenvolvimento de pesquisas, técnicas e métodos que permitam auxiliar no estudo de solos e seus atributos. Este estudo foi divido em dois capítulos, a saber: o primeiro capítulo teve como objetivo avaliar a aplicação de modelos de aprendizagem de máquina (AM) na predição de estoque de carbono do solo a 30cm (SOCS30) e 100cm (SOCS100) de profundidade e no segundo capítulo na predição da composição granulométrica do solo em superfície e subsuperfície. Os objetivos específicos foram: comparar duas formas de distribuição de dataset: área de referência (AR) e área total (AT), avaliar duas categorias de seleção de covariável: "método wrapper" e "seleção prévia de covariável" como etapa de pré-processamento, antes da calibração do AM e avaliar o desempenho de três algoritmos de AM: “Regression Tree” (RT), “Random Forest” (RF) e “Support Vector Machine” (SVM) na modelagem dos atributos do solo. O local do estudo foi dividido em três blocos, denominados blocos Urucu, Araracanga e Juruá. O conjunto de dados consistiu em 120 observações de estoque de carbono e 151 observações de composição granulométrica combinadas com 21 covariáveis (20 covariáveis de relevo e 1 índice derivado da banda P do radar) que foram abordadas em dois datasets diferentes: (AR) e (AT). A similaridade entre a AR e a paisagem dos blocos foi avaliada por meio do índice geral de Gower e a estatística descritiva das covariáveis. Os resultados mostraram que o uso da seleção de covariáveis, combinada com o uso de conjunto de dados da AR, permitiram desenvolver modelos mais precisos para prever a maior parte dos atributos mapeados. De acordo com o índice geral de Gower, a AR possui alta similaridade com os blocos Urucu, Araracanga e Juruá. Entretanto, as estatísticas mostraram que aumentando a distância da AR, algumas covariáveis de relevo são mais diferentes. No primeiro capítulo os modelos de predição desenvolvidos para predizer o SOCS100 apresentaram maior acurácia e transferibilidade do que aqueles desenvolvidos para predizer o SOCS30. O algoritmo RF gerou os mapas mais acurados de SOCS100 para os Blocos de Urucu e Juruá (R^2 = 0,70 e 0,51, respectivamente). Os valores de SOCS100 dos mapas gerados para a região do Bloco de Urucu variaram de 3,89 kg C. m^{-2} a 10,64 kg C .m^{-2}, enquanto para o bloco Juruá variaram de 5,03 kg C. m^{-2} a 10,42 kg C. m^{-2}. No capítulo 2 o melhor desempenho também foi obtido com o algoritmo RF na predição de silte em superfície e subsuperfície para os Blocos Urucu e Juruá (R^2 = 0,58 e 0,52, 0,51 e 0,56 respectivamente). Os valores de silte superficial e subsuperficial dos mapas gerados para a região do Bloco Urucu variaram de 208,97 g kg-1 a 576,68 g kg-1 e 215,32 g kg-1 a 517,06 g kg-1, enquanto para o bloco Juruá variaram de 236,10g kg-1 a 555,70 g kg-1 e 229,83 g kg-1 a 460,56 g kg-1, respectivamente. Apesar da baixa densidade de observação do conjunto de dados disponível, os resultados mostram não só a importância dos algoritmos de AM para mapear os atributos do solo, mas também do uso de conhecimento pedológico especializado gerado em uma AR para apoiar uma seleção de covariáveis antes de calibrar os algoritmos.

Palavras-chave: Aprendizado de máquina. Mapeamento digital de solo. Área de referência.

GENERAL ABSTRACT

FERREIRA, Ana Carolina de Souza. **Digital mapping of soil attributes in remote areas under the Amazon Forest: A case study in the Solimões Formation**. 2022. 99p. Thesis (Doctor Science in Agronomy-Soil Science). Agronomic Institute, Rural Federal University of Rio de Janeiro, Seropédica, RJ, 2022.

Currently, the demand for digital soil mapping (DSM) has grown considerably in environmental studies. Faced with these demands and the lack of information on soils on an adequate scale, it is more than necessary to develop research, techniques and methods that help in the study of soils and their attributes. This study was divided into two chapters: The first chapter aimed to evaluate the application of machine learning models (ML) in the prediction of soil carbon stock at 30cm (SOCS30) and 100cm (SOCS100) depth and in the second chapter in the prediction in surface and subsurface the soil texture composition. The specific objectives were; Compare two forms of dataset distribution; reference area (RA) and total area (TA), to evaluate two categories of covariate selection: "wrapper method" and "pre-selection of covariate" as a pre-processing step, before ML calibration, and to evaluate the performance of three ML algorithms: regression tree (RT), random forest (RF) and support vector machine (SVM) in the modeling of soil attributes. The study site was divided into three blocks, called Urucu, Araracanga and Juruá blocks. The dataset consisted of 120 carbon stock observations and 151 particle size composition observations combined with 21 covariates (20 relief covariates and 1 P-band-derived index from radar) that were addressed in the two different datasets: (RA) and (TA). The similarity between the RA and the landscape of the blocks was evaluated using the Gower general index and the descriptive statistics of the covariates. The results showed that the use of the previous covariates selection, combined with the use of an RA dataset, allowed the development of more accurate models to predict most of the attributes studied. According to Gower's general index, the RA has high similarity with the Urucu, Araracanga and Juruá blocks. However, statistics showed that increasing the distance from the RA, some relief covariates are more different. In the first chapter, the prediction models developed to predict SOCS100 showed greater accuracy and transferability than those developed to predict SOCS30. The RF algorithm generated the most accurate SOCS100 maps for the Urucu and Juruá Blocks (R^2 = 0.70 and 0.51, respectively). The SOCS100 values of the maps generated for the Urucu Block region ranged from 3.89 kg C. m-² to 10.64 kg C. m-², while for the Juruá block they ranged from 5.03 kg C. m -² to 10.42 kg C. m-². In chapter 2, the best performance was also obtained with the RF algorithm in the prediction of surface and subsurface silt for the Urucu and Juruá Blocks (R^2 = 0.58 and 0.52, 0.51 and 0.56 respectively). The surface and subsurface silt values of the maps generated for the Urucu Block region ranged from 208.97 g kg-1 to 576.68 g kg-1 and 215.32 g kg-1 to 517.06 g kg-1, while for the Juruá block they ranged from 236.10g kg-1 to 555.70 g kg-1 and 229.83 g kg-1 to 460.56 g kg-1, respectively. Despite the low observation density of the available dataset, the results show not only the importance of ML algorithms to map soil attributes, but also the use of specialized pedological knowledge generated in an RA to support a selection of covariates before calibrate the ML algorithms.

Keywords: Machine learning. Digital soil mapping. Reference area.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO GERAL

Atualmente, o interesse pelo mapeamento digital de solos (MDS) tem crescido consideravelmente em estudos ambientais. Diante dessa demanda e da falta de informações sobre os solos em escala adequada, é mais que necessário o desenvolvimento de pesquisas, técnicas e métodos que permitam auxiliar no estudo dos solos e seus atributos. Os mapas de atributos do solo têm um papel fundamental na avaliação das funções do solo. Além de auxiliar nos processos de tomada de decisão, principalmente no que se refere a propriedades do solo, proporciona ainda, ganho de conhecimento a respeito da importância do solo no auxílio e na recuperação de áreas impactadas, controle de erosão e, sobretudo, no planejamento e logística de diversas atividades.

O MDS, surge no contexto de uma época em que temos várias ferramentas digitais (hardware e software) e aprimoramento de pesquisas. Pesquisas estas, que tentam atingir os objetivos de produzir mapas com maiores acurácias, com menor tempo, custo, melhor representatividade e variabilidade espacial. Sendo o mais importante objetivo, apresentar modelos de predição que sejam passíveis de reprodução e avaliação por outros pesquisadores.

A Amazônia brasileira contém mais de um terço das florestas do mundo. A região compreende vasta extensão territorial e representa 59% do território do Brasil (BRASIL, 1978; PINTO et al., 2003). Apesar da importância ambiental, a região ainda é relativamente pouco conhecida e, dentre as grandes lacunas de informação, destaca-se a de inventário de dados de classes e atributos do solo. Segundo Ceddia et al. (2015), a maior parte dos dados disponíveis nesta região é proveniente da mesma fonte: o projeto RADAMBRASIL, que foi realizado entre os anos de 1973 e 1984. Os dados disponíveis e mapas utilizados por diversos autores têm sido publicados em escalas pouco detalhadas e os dados disponíveis para os cálculos dos atributos do solo são, em geral, muito escassos em toda a região amazônica. A baixa densidade de dados de solo nesta região, pode ser explicada pela presença de uma floresta densa, o que dificulta os levantamentos pedológicos. A floresta densa se torna um obstáculo para o acesso ao território, sendo possível somente por meio de barcos, helicópteros e aviões. Além disso, existem diversas ameaças naturais inerentes a esse bioma, como animais silvestres e doenças tropicais. O conjunto desses fatores demanda logística específica e aumenta os custos de levantamentos de solos nessa região (CEDDIA et al., 2017).

Uma grande diversidade de algoritmos tem sido utilizada para predizer propriedades e tipos de solos. Diferentes abordagens e técnicas são aplicadas para mapear esses atributos, tais como: modelos estatísticos lineares simples, geoestatísticos, técnicas híbridas a métodos avançados e complexos de aprendizado de máquina (LAMICHHANE et al., 2019).

A estrutura geoestatística por exemplo, tem sido comumente incorporada no trabalho do MDS para que um modelo estatisticamente sólido seja assumido para variação espacial. Além disso, explicitamente, tanto a autocorrelação espacial é modelada quanto as medidas de incerteza são associadas à predição (WADOUX & MCBRATNEY, 2021; WADOUX et al., 2020). Apesar dessas vantagens, o uso da geoestatística pode ser limitado quando os dados disponíveis para modelar a variação espacial têm baixa densidade e a localização espacial das unidades amostrais são desequilibradas sobre a área a ser mapeada. Nesse contexto, modelos de aprendizado de máquina (AM) têm sido utilizados com maior frequência na ciência do solo. Esses modelos não lineares são empregados principalmente para fins de mineração de dados e reconhecimento de padrões, e frequentemente usados para tarefas de regressão e classificação. A vantagem desses algoritmos, é que não estão condicionados a seguir nenhuma suposição estatística, além de poderem lidar com um grande número de covariáveis relacionadas entre si (colinearidade) como preditoras. No entanto, apesar da maioria dos modelos de AM serem

robustos à multicolinearidade entre covariáveis, há várias razões para selecionar um subconjunto de covariáveis para calibrar o modelo. A seleção de covariáveis é uma etapa importante pois, além de reduzir o número de covariáveis usadas para calibrar os algoritmos de AM, permite calibrar o modelo mais rapidamente, reduzir a complexidade, aumentar a acurácia da predição ou até mesmo evitar overtting do modelo. (WADOUX & MCBRATNEY, 2021; WADOUX et al., 2020). Muitos estudos apontam o potencial desses algoritmos para produção de mapas de atributos e classes de solo (ADHIKARI et al., 2013; ADHIKARI, 2014; BHERING et al., 2016; CHAGAS et al., 2016a; GUO et al., 2019; HENGL et al., 2015; KOVAČEVIĆ et al., 2010; MEIER et al., 2018; PINHEIRO, 2015; PINHEIRO et al., 2018; SREENIVAS et al., 2014; TAGHIZADEH-MEHRJARDI et al., 2014; VASCONCELOS & OLIVEIRA, 2018).

Considerando esses aspectos, uma importante estratégia para criar e expandir áreas com mapas mais detalhados dos atributos do solo, é o uso de técnicas de MDS, associadas a algoritmos de machine learning extraindo e otimizando as informações de levantamentos detalhados de solos já existentes ao longo de uma região (Áreas de Referência - AR). Essa abordagem foi desenvolvida com base na hipótese de que é possível delimitar áreas (chamadas "pequenas áreas naturais") que englobam um número finito de classes de solos, que são recorrentes em associação entre si na paisagem formando um padrão que se repete e é identificável. Assim, uma área de referência propositalmente escolhida, seria suficiente para identificar todos os tipos de solos em uma área maior e determinar suas relações espaciais. Essa abordagem também pode ser testada para desenvolver modelos preditivos para estimar atributos do solo, como estoques de carbono e composição granulométrica do solo (ADHIKARI, 2014; ARRUDA et al., 2016; ARRUDA et al., 2013; GRINAND et al., 2008; HÖFIG et al., 2014; LAGACHERIE et al., 1995; LAGACHERIE et al., 2001; LAGACHERIE & VOLTZ, 2000; LIESS et al., 2012; MCNICOL et al., 2019; SCULL et al., 2005; SILVA et al., 2016; VOLTZ et al., 1997; WANG et al., 2017; WOLSKI et al., 2017).

A hipótese principal deste estudo é de que é possível gerar mapas de variabilidade espacial dos atributos do solo a partir da abordagem de área de referência da base de Urucu para dimensões de áreas mais extensas da formação Solimões.

O objetivo geral é avaliar a transportabilidade de modelos de predição para mapear atributos do solo na formação Solimões.

Os objetivos específicos do presente estudo foram: i) Comparar duas formas de distribuição de dataset (Área de Referência - AR e Área Total - AT) para prever SOCS em profundidades de 30 (SOCS30) e 100 cm (SOCS100) e composição granulométrica do solo em superfície e subsuperfície; ii) avaliar duas categorias de seleção de covariável: "método wrapper", todas as covariáveis são selecionadas na entrada do modelo, o analista não tem controle sobre as covariáveis que estão entrando para poder julgar a plausabilidade do modelo e "seleção prévia de covariável" como etapa de pré-processamento, o analista tem o controle das covariáveis inseridas antes da calibração do AM; iii) avaliar a transferibilidade e o desempenho de três algoritmos de AM: "Regression Tree" (RT), "Random forest (RF) e "Support Vector Machine" (SVM).

O estudo foi divido em dois capítulos, identificados a seguir:

Capítulo I - Predicting soil carbon stock in remote areas of the central amazon region using machine learning techniques; e,

Capítulo II - Predição da composição granulométrica do solo em áreas remotas da amazônia central usando técnicas de aprendizagem de máquinas.

2. REVISÃO DE LITERATURA GERAL

2.1 Contexto Ambiental da Região

Desde 1988 a Petrobrás explora gás natural e petróleo na província petrolífera de Urucu. A unidade, denominada Base de Operações Geólogo Pedro Moura (BOGPM/UN-BSOL), está situada no município de Coari, distando 650 km a sudoeste de Manaus. Até o ano de 2007, os dados disponíveis de solos eram basicamente provenientes do projeto RADAMBRASIL e alguns outros estudos locais de fertilidade e recuperação de áreas degradas, porém sem uma sistematização maior.

Este trabalho está inserido em um projeto de pesquisa que teve início em 2008, intitulado: Petrossolos Amazônicos, cujo principais objetivos eram organizar uma base de dados sistemática georreferenciada com diferentes temas, mapear classes e atributos do solo da BOGPM e gerar protocolos de exploração e recuperação ambiental. Os levantamentos das atividades na Amazônia iniciaram-se em 2008 na área da UN-BSOL, a qual passou a ser considerada como área de referência para desenvolvimento de modelos de predição para áreas no entorno. De 2008 a 2013 foram desenvolvidos os primeiros mapeamentos digitais e gerada a primeira base de dados de solos que, posteriormente, contribuiu para diversos artigos publicados, com destaque para Ceddia et al. (2015, 2017), referente aos atributos que foram modelados nesse estudo. Em 2018, um novo trabalho de campo visitou novas clareiras remotas (até 100 km da AR), o qual gerou mais perfis em outras regiões fora da AR, totalizando o banco de dados utilizado para mapear os atributos de solo abordados com a metodologia proposta por este estudo.

A área de estudo é de aproximadamente 13.440 km², está situada entre os paralelos 4°0’ e 6°0’S e 67°0’ e 64° 00’W localizada entre os municípios de Carauari e Coari, no Estado do Amazonas (Figura 1). Durante os períodos Pleistoceno e Holoceno as flutuações climáticas exibiram estações úmidas a secas bem definidas, culminando no clima atual. O clima atual, segundo a classificação de Köppen e Geiger, é Af (equatorial, do mês mais frio acima de 20°C, sem períodos secos pronunciados e precipitação média anual igual ou superior a 2500 mm). A região é remota, sendo possível o acesso somente por transporte aéreo e fluvial.

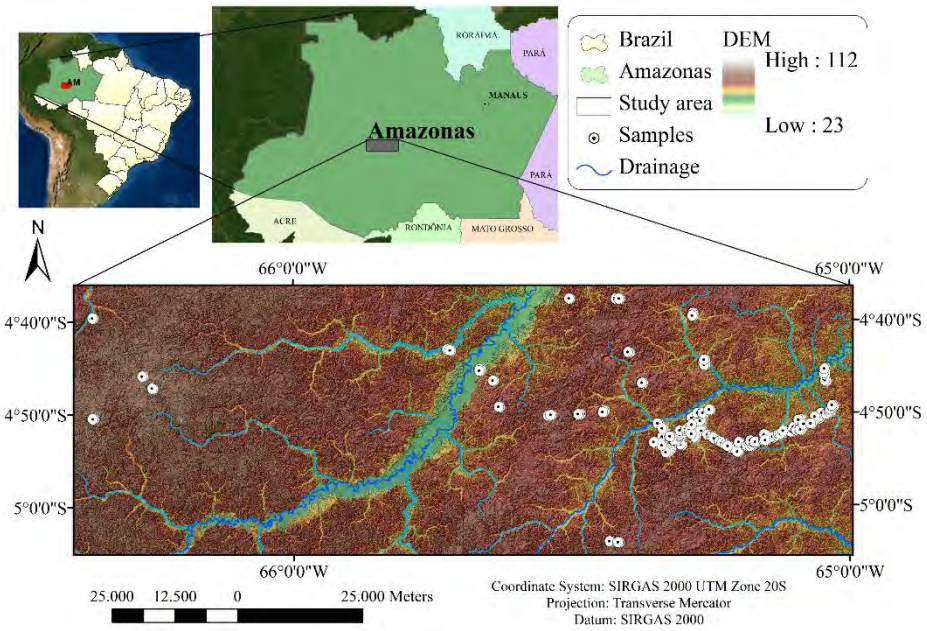


Figura 1. Mapa de localização da área de estudo. (Fonte: ArcGIS, elaborada pela Autora).

A região de estudo está inserida na província de coberturas Fanerozóicas, mais especificamente a parte da Formação Solimões entre os arcos de Iquítos e Purus (Figura 2).

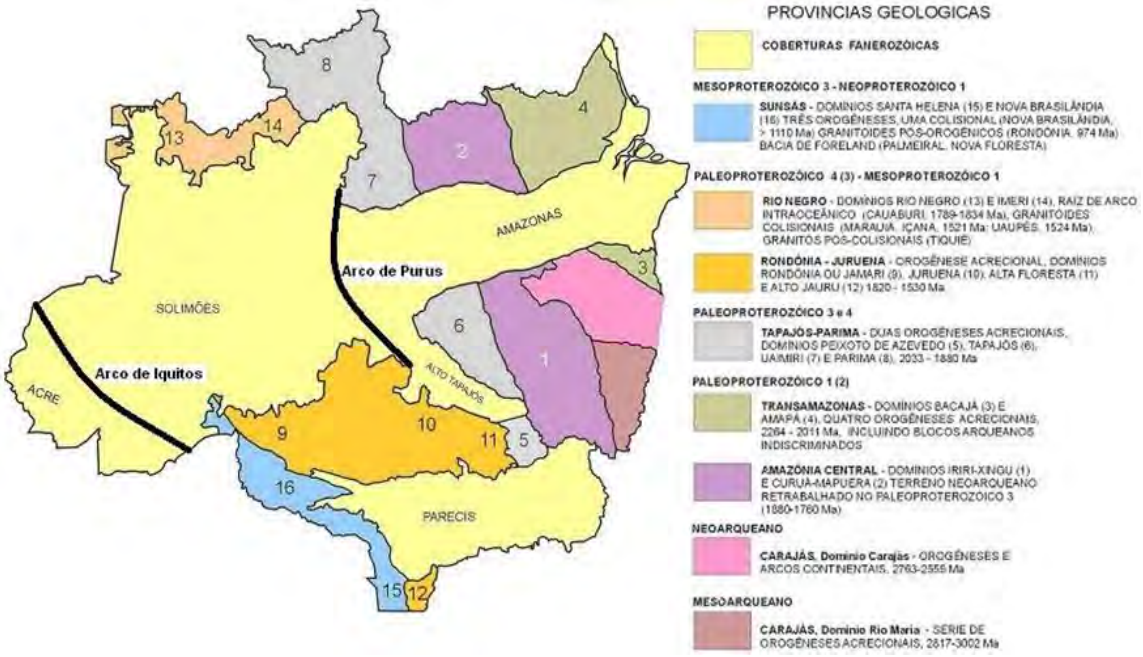


Figura 2. Províncias geológicas da região da Amazônia Legal. (Fonte: BRASIL, 1978).

A partir da consolidação de informações e hipóteses levantadas no projeto RADAMBRASIL (1978), bem como do estudo de Ceddia et al. (2015), verifica-se que o material de origem, relevo, vegetação e clima, atuando de forma interdependente, explicam a relação dos tipos de solo e seus respectivos atributos. Para entendermos melhor os solos que compõe a região de estudo, podemos observar essa relação solo-relevo-vegetação representada pela topossequência da área de referência (Figura 4). Os solos da região de estudo são formados a partir de sedimentos da formação Solimões (composta por argila, silte e areia muito fina) durante o período Terciário-Quaternário. Nesse ambiente, as formas de relevo encontradas são classificadas em APf (planícies fluviais), C11 (áreas secas em topos planos), T21 (interflúvios tabulares) e EP2 (superfícies biplanícies-planícies. Basicamente, são encontradas quatro unidades de mapeamento (MU1, MU2, MU3 e MU4) que podem fornecer informações para melhor definir as covariáveis usadas (CEDDIA et al., 2015).

MU1- Complexo/Associação: PVAa+CXba. Essa unidade, representa as regiões de encostas com declividades mais acentuadas e com boa drenagem, geralmente mais próximas das grandes redes de drenagem. Constata-se o predomínio de solos classificados como ARGISSOLO VERMELHO-AMARELO Alumínico típico - PVAa e CAMBISSOLO HÁPLICO Tb Alumínico típico - CXba. A diferença básica entre os Argissolos vermelho amarelo e os Cambissolos háplicos é a mudança textural para atender os critérios de horizonte diagnóstico subsuperficial B textural para enquadrar na ordem Argissolo.

MU2- Complexo/Associação: GXvd + CXbd + GXbd, unidade que se estende por todas as áreas de baixada. As áreas de baixadas podem estar associadas às calhas dos grandes rios e igarapés e terraços no entorno do curso d'água principal (vales em formato U) bem como de vales mais encaixados (vales em formato V) das regiões de encostas. Diversos solos foram encontrados nessas baixadas, no entanto, predominam os classificados como GLEISSOLO HÁPLICO Ta Distrófico típico A moderado – GXvd, CAMBISSOLO HÁPLICO Tb Distrófico

típico A moderado - CXbd e GLEISSOLO HÁPLICO Tb Distrófico cambissólico A moderado - GXbd. Basicamente os solos apresentam cores acinzentadas típicas de ambientes hidromórficos com horizonte subsuperficial diagnóstico Gley dentro dos primeiros 50 cm de profundidade ou B incipiente.

MU3- Complexo/Associação: PAd+CXba+PAa. Áreas de encostas mais suaves e topos menos dissecados que os da unidade de mapeamento PVAa+CXba (Vermelhos-Amarelos). O ambiente onde ocorrem tem boa drenagem, predomínio de solos distróficos e com declividade média em torno de 5%. Predomínio de solos classificados como ARGISSOLO AMARELO Distrófico típico A moderado – PAd, CAMBISSOLO HÁPLICO Tb Alumínico típico A moderado – Cxba e ARGISSOLO AMARELO Alumínico típico A moderado – PAa. Basicamente os solos apresentam cores com predomínio de cor amarela (7,5YR e 10YR), inclusive os cambissolos (CXba), os quais se diferenciam dos demais cambissolos da MU1 por serem mais amarelos.

MU4- Complexo/Associação: GXbd2 + PACd (Planaltos Hidromórficos). Ocorre nos divisores de água das bacias hidrográficas. O ambiente é denominado de Planaltos de Altitude, pois está geralmente associado a um ambiente de altitude relativamente maior (70 a 80 metros acima do nível do mar) com relevo bastante plano a suave ondulado (declividade média de 3.9%). Nessas paisagens, a elevada precipitação associada ao relevo plano dos topos amplos de elevação, gera um ambiente de hidromorfismo. Predomínio de solos classificados como GLEISSOLO HÁPLICO Tb Distrófico argissólico A moderado - GXbd e ARGISSOLO ACINZENTADO Distrófico típico A moderado – PAd (CEDDIA et al., 2015).

Outra característica que se destaca nestas diferentes paisagens, são as vegetações específicas que estão associadas a estes tipos de solos. Considerando que o clima é praticamente o mesmo em toda a área de estudo e que comumente a precipitação é alta e relativamente homogênea ao longo do ano, os tipos de vegetação ao longo da área são fortemente afetados pela condição de drenagem dos solos e classificadas da seguinte forma: floresta tropical aberta de terras altas (fdb), comumente encontrada nos planaltos hidromórficos; floresta tropical densa de terras altas (fda) predominante nas áreas mais declivosas e com boa drenagem e as florestas tropicais abertas de várzea inundada (Fac), típicas de áreas inundadas e mal drenadas (Figura 3) (CEDDIA et al., 2015).

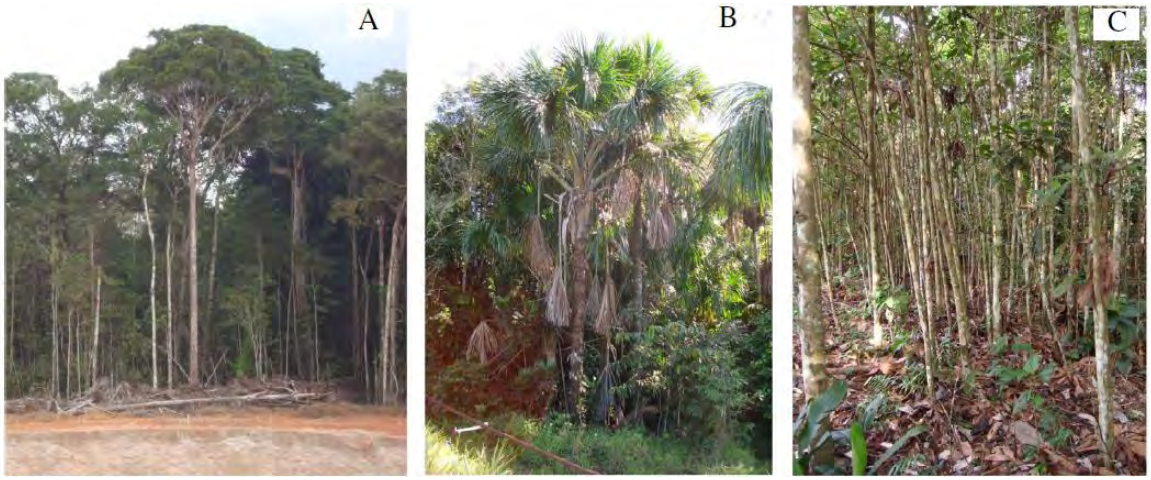
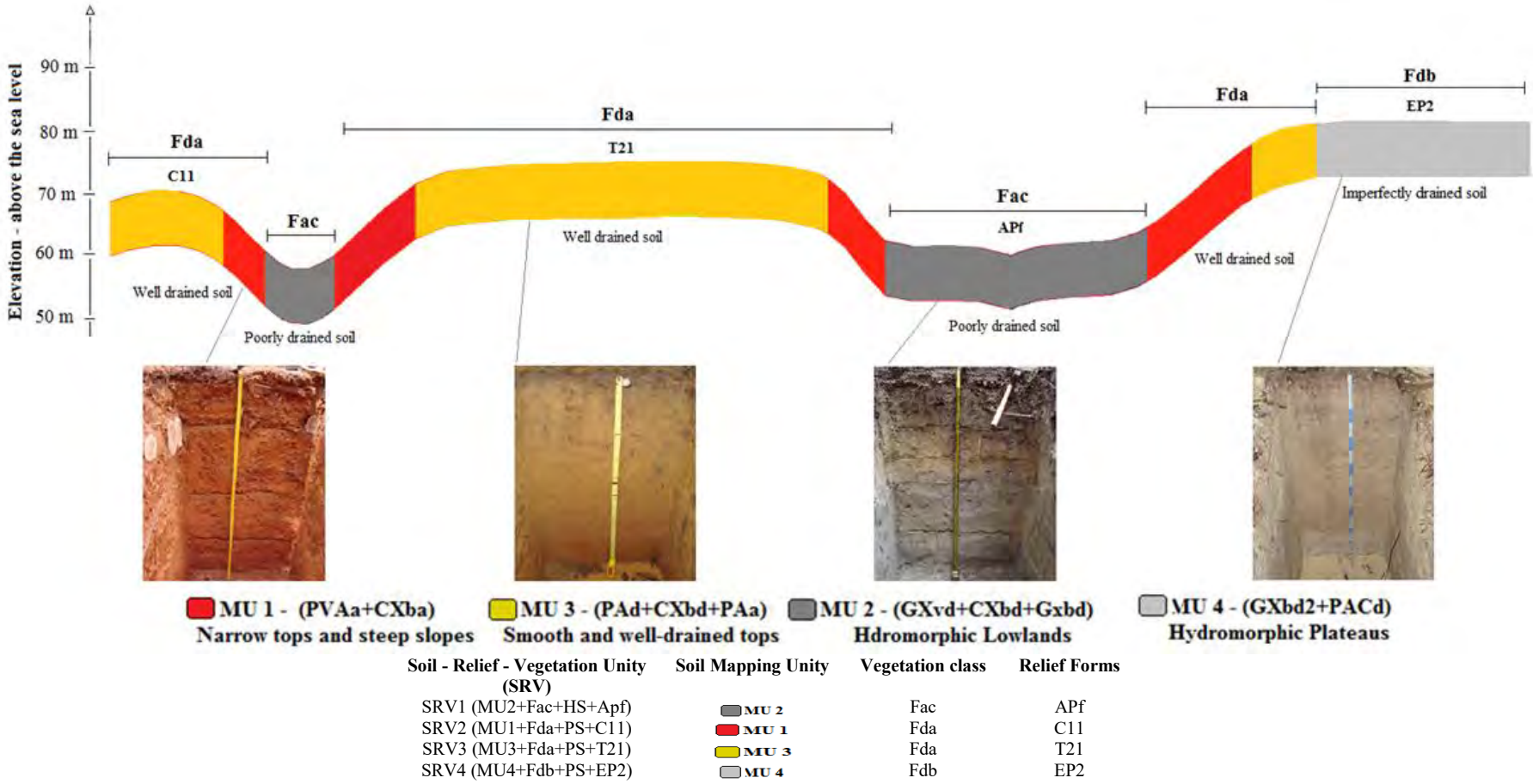


Figura 3. Tipo de vegetação na área de estudo. A) Fda – Floresta Tropical Densa de Terras Altas, B) Fac – Floresta Tropical Aberta de Baixada Inundada e C) Fdb – Floresta Tropical Aberta de Terras Altas. (Fonte: CEDDIA et al., 2015).



Soil - Relief - Vegetation Unity (SRV)	Soil Mapping Unity	Vegetation class	Relief Forms
SRV1 (MU2+Fac+HS+Apf)	MU 2	Fac	APf
SRV2 (MU1+Fda+PS+C11)	MU 1	Fda	C11
SRV3 (MU3+Fda+PS+T21)	MU 3	Fda	T21
SRV4 (MU4+Fdb+PS+EP2)	MU 4	Fdb	EP2

Figura 4. Representação esquemática da relação solo-relevo-vegetação ao longo da RA. Fac- Floresta Tropical Aberta de Planície Inundada; Fda- Floresta Tropical Densa de Terras Altas; Fdb- Floresta Tropical Aberta de Planalto; APf- Planícies fluviais; C11- Áreas bem drenadas em topo plano; T21- Interflúvios Tabulares; EP2 - Superfícies biplanícies- planícies. H.S.—Sedimentos Holocênicos; P.S.—Sedimentos Pleistoceno. (Fonte: CEDDIA et al., 2015).

2.2 Área de Referência

O método de AR baseia-se na caracterização da cobertura de solos de regiões topograficamente e geologicamente identificáveis, denominadas "pequenas áreas naturais". O primeiro passo deste método, consiste em mapear detalhadamente uma pequena área que seja representativa, uma pequena região natural a qual é denominada AR. Nesta AR, caracterizam-se as principais classes de solos de toda a região e se estabelecem as regras de mapeamento. Esta primeira etapa facilita e acelera a etapa seguinte, que consiste na condução de novos trabalhos de mapeamento em áreas circunvizinhas. Nestes novos trabalhos, são levantados novos pontos de observação onde se esperam encontrar as mesmas classes de solos identificadas e mapeadas na AR. Também se espera poder fazer o delineamento dos limites de unidades de mapeamento a partir das regras de mapeamento pré-estabelecidas. Sendo assim, a metodologia consiste na utilização de mapas, cartas pedológicas, informações gerais existentes ou na geração de informações dessas "pequenas áreas naturais" para mapear áreas maiores que a mapeada originalmente ou áreas adjacentes, desde que condizem com os mesmos fatores de formação do solo (LAGACHERIE et al., 1995; VILLELA, 2013).

Quando se desenvolve mapeamento tendo como base o método de AR, o que está sendo testado é a hipótese de que é possível delimitar áreas que englobem um número finito de classes de solos, as quais são recorrentes em associação umas com as outras na paisagem, formando um padrão repetitivo e identificável. Desta forma, uma AR propositalmente escolhida seria suficiente para identificar todos os tipos de solos de uma área maior e determinar suas relações espaciais, ou seja, as ARs são utilizadas como base para a extrapolação das relações solo-paisagem, o que possibilita o mapeamento das áreas vizinhas com características semelhantes (VILLELA, 2013).

Embora existam alguns estudos sobre AR, sobretudo na França, a hipótese assumida no método da AR, ainda precisa ser experimentalmente confirmada, demonstrando que o método é satisfatório para aquela área de estudo específica (LAGACHERIE et al., 1995). Ainda segundo Lagacherie et al. (1995), os resultados obtidos na França demonstraram que os trabalhos efetuados utilizando AR têm reduzido a necessidade de trabalhos de campo, e que em geral menos de 10% das unidades de solos encontradas nas novas áreas são diferentes daquelas estabelecidas na AR.

A desvantagem do método da AR é não poder reproduzir o raciocínio que um pedólogo experiente realizou durante o mapeamento, limitando sua transferência a outros pedólogos. No entanto, se a hipótese da AR estiver correta, a partir de levantamentos detalhados em áreas devidamente selecionadas é possível ampliar o mapeamento com custos relativamente menores, além de reduzir a necessidade de trabalhos de campo (exceto para validação dos resultados). Isso se torna extremamente importante, principalmente em estudos onde as regiões mapeadas são muito remotas e a logística torna-se mais complexa.

Uma primeira avaliação experimental da hipótese de representatividade da AR, foi realizada por Favrot (1989) em cinco diferentes áreas em torno das quais áreas representativas foram delineadas manualmente por topógrafos. Observou-se nessas áreas representativas, que os tipos de solos já identificados nas ARs foram encontrados em mais de 90% da área. A partir dos primeiros estudos desenvolvidos na França, outros pesquisadores também aplicaram a abordagem da AR no MDS (BAGATINI et al., 2016; GRINAND et al., 2008; HÖFIG et al., 2014; LAGACHERIE et al., 1995; LAGACHERIE et al., 2001; LAGACHERIE & VOLTZ, 2000,; VOLTZ et al., 1997; WOLSKI et al., 2017).

Do ponto de vista metodológico, a aplicação da abordagem AR enfrenta dois aspectos fundamentais: 1- Como escolher e delimitar uma AR; 2- Qual a representatividade dessa AR (qual a abrangência territorial que os modelos de predição de classes e atributos dos solos são transferíveis com acurácia aceitável a partir da AR). Nesse estudo, por se tratar de uma área

remota e por questões de logística de acesso, não existe a opção de escolher e delimitar uma AR. A AR é previamente estabelecida como sendo a BOGPM (Base de Urucu), pois, é a única onde se tem um mapa de solo detalhado desenvolvido. Nesse estudo é avaliado o segundo desafio, ou seja, qual a sua representatividade territorial e a transferabilidade dos algoritmos desenvolvidos na AR para áreas maiores e distantes (blocos de Urucu, Araracanga e Juruá).

2.3 Uso de Imagens de Radar em Áreas de Floresta

Um radar imageador é considerado um sistema ativo, pois, possui sua própria fonte de energia para obter uma imagem, onde uma determinada quantidade de energia eletromagnética emitida pelo sensor é retroespalhada pelos alvos e registrada pelo sistema. Sendo assim, os radares são capazes de operar tanto de dia quanto à noite e, em razão do maior comprimento de onda utilizado, são de grande utilidade em regiões com condições climáticas e ambientais adversas. A extensão da penetração depende da umidade, da densidade da vegetação, bem como do comprimento de onda. Comprimentos de onda menores interagem com as camadas superficiais da vegetação, enquanto os comprimentos de onda mais longos com as camadas inferiores da vegetação. Podendo, em alguns casos, até mesmo interagir com o solo ou mesmo com o subsolo (INPE, 2008).

Um radar de imageamento transmite e recebe pulsos situados na faixa das micro-ondas, entre as bandas P e K, a intervalos regulares e confinados a curto intervalo de tempo. Os pulsos são transmitidos através de uma antena e atingem a superfície do terreno, e a interação faz com que a energia seja espalhada em várias direções. Uma parcela desta é espalhada em direção ao próprio sensor, o que é chamado de retroespalhamento ("backscattering") e é recuperada à medida em que o sensor se desloca (HENDERSON & LEWIS, 1998).

Às imagens de radar, possuem diversas vantagens tais como: fonte de iluminação controlável (como mencionado), sendo o seu imageamento independente de condições atmosféricas e de iluminação solar, maior capacidade de penetração nos alvos que as imagens de sensores ópticos, observação noturna, penetração através de nuvens e da chuva, alta resolução de 3 a 10m e diferentes feições são registradas ou discriminadas quando comparadas com sensores ópticos. Esses aspectos, as tornam muito interessantes principalmente no estudo de regiões de clima tropical, já que este tipo de imagem sofre pouca influência das condições atmosféricas e podem obter informações da cobertura e superfície do solo (RENNÓ, 2003).

A Região da Amazônia Legal Brasileira representa um grande desafio para o desenvolvimento de estudos sistemáticos de mapeamento de solos. Além de cobrir um território imenso, tem grande parte deste coberto por uma floresta amazônica densa. Destaca-se também a baixa densidade de estradas, sendo a maior parte do território acessado apenas por barcos e transporte aéreo. Associado a isso, o clima equatorial e a constante presença de nuvens dificultam o uso de imagens de satélite e de fotos aéreas obtidas por sistemas passivos. Esse conjunto de características faz com que os sistemas ativos como o radar, sejam alternativas com grande potencial para desenvolvimentos de diversos estudos na região, servindo como suporte ao desenvolvimento de mapeamento sistemático de diferentes planos de informação (por exemplo: solos, geologia e geomorfologia).

No mapeamento topográfico de regiões de floresta densa, especificamente o comprimento da banda "P" (72 cm de comprimento de onda), possibilita a penetração no dossel da floresta e a interação da onda com a superfície do terreno, gerando informações das feições existentes ao nível do solo da floresta tropical densa. A banda P torna-se um objeto de muito interesse pois, por ter um comprimento de onda maior, a mesma possui peculiaridades em relação a outros comprimentos de onda. Uma dessas características é a capacidade de penetrar pelas copas das árvores e gerar reflexões suficientemente fortes do terreno abaixo, sendo mais sensível às variações de biomassa do que outras bandas como X, C e L. Além disso, através das

polarizações VV, HH, pode-se gerar índices através de razões e cruzamentos dessas polarizações, que poderiam discriminar aspectos da vegetação e da superfície do solo.

A maior parte das pesquisas encontradas na literatura com dados de radar referem-se a estudos florestais. No entanto, recentemente, houve um crescimento da aplicação na área de solos principalmente nos estudos de umidade (BLUMBERG et al., 2002; DU et al., 2015; LIN et al., 1994; MOGHADDAM et al., 1997; ZRIBI et al., 2016b). Isso torna-se interessante pois, o comportamento dielétrico do solo também é afetado pela distribuição dos tamanhos dos grãos e pela quantidade de água. A variação da quantidade de água do solo pode ser detectada por sensores de micro-ondas e está diretamente relacionada a capacidade de retenção de água, que por sua vez, está associada com a textura do solo, podendo ter uma relação indireta com a composição granulométrica do solo (SRIVASTAVA et al., 2006).

Parte dos trabalhos encontrados na literatura, apresentaram melhores coeficientes de correlação para a biomassa dos galhos, troncos e folhas na banda P comparado a outros comprimentos de onda. Os autores justificam uma melhor e maior resposta da banda P nas polarizações VV e HH provavelmente devido à alta penetração nas copas das árvores e a consequente ocorrência dos mecanismos de double-bounce nos troncos das árvores, principalmente na polarização HH. Assim sendo, a banda P tem um enorme potencial para aplicações na estimação de biomassa, na obtenção de modelos digitais de elevação principalmente em áreas com cobertura densa (interferometria), na detecção e exploração e degradação da floresta e estudos relacionados ao solo (SAMBATTI et al., 2012; GAMA et al., 2009; LIN et al., 1994; MOGHADDAM et al., 1997; SAATCHI et al., 2011; ZRIBI et al., 2016a, 2016b) . A Figura 5 mostra um exemplo de aplicação de imagens SAR relativo ao sistema InSAR da Orbisat. O radar utilizou duas frequências para gerar produtos interferométricos (bandas X e P), possibilitando tanto a medida de altura da copa das árvores como a do solo sob a vegetação (modelo digital de elevação da área de estudo).

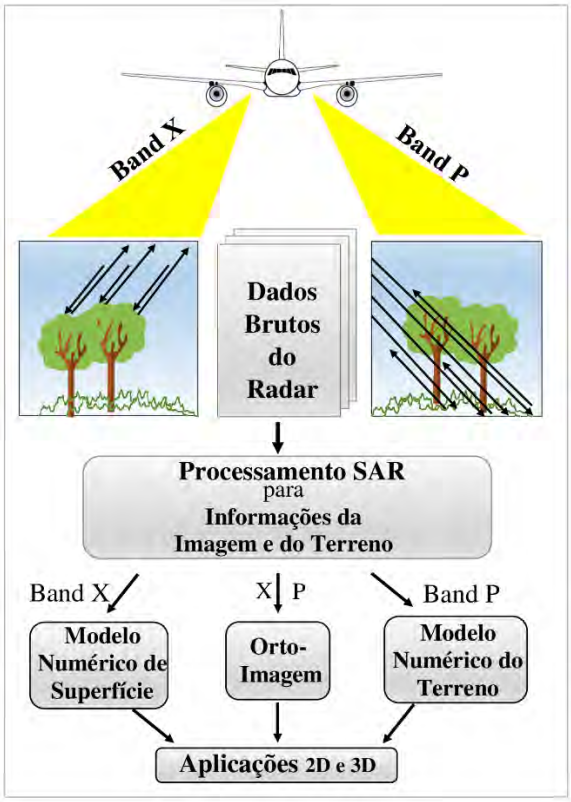


Figura 5. Exemplo de aplicação de imagens SAR. (Fonte: modificado ROSA, 2009).

2.4 Algoritmos de Aprendizado de Máquina

As técnicas de AM, referem-se a uma grande classe de algoritmos orientados a dados não lineares, empregados principalmente para fins de mineração de dados e reconhecimento de padrões. Tem como base, o aprendizado com dados e a identificação de padrões com o mínimo ou, sem nenhuma intervenção humana. São frequentemente usados para tarefas de regressão e classificação em várias áreas científicas.

Por serem modelos não paramétricos, tornam-se muito interessantes, já que nenhuma suposição é feita em relação à distribuição das variáveis. Também realizam seleção automática de subconjunto de variáveis e podem lidar com dados quantitativos e dados categóricos, permitindo a relação de variáveis tanto qualitativas quanto quantitativas, por exemplo; atributos do terreno com índices de sensoriamento remoto, geologia ou classes categóricas de cobertura de solo.

Muitos desses algoritmos possuem técnicas robustas e processam grandes números de dados. No entanto, ainda existe uma interpretação limitada dos resultados dessas técnicas. Uma vez, que as relações entre os preditores e as respostas não podem ser examinadas individualmente com mais clareza, muitas vezes são considerados como “caixas pretas”. Apesar disso, muitos estudos têm obtido resultados satisfatórios, fazendo com que aprendizagem de máquina (*machine learning*) ganhe cada vez mais espaço no MDS.

Existe uma grande diversidade de algoritmos que compõem o grupo denominado Machine Learning, nesse estudo utilizamos três algoritmos de AM amplamente utilizados em MDS: (Arvore de decisão - AD, Random Forest-RF e Support Vector Machine-SVM) suas vantagens e desvantagens podem ser observadas na Tabela 1.

2.4.1 Modelos baseados em árvores

a) Árvore de decisão/regressão (AD/R)

Em mineração de dados a árvore de decisão (AD) é um algoritmo de aprendizado de máquina supervisionado que é utilizado tanto para classificação quanto para regressão. Quando usada para predição categórica (aprendizagem de classificação), é denominada árvore de classificação como por exemplo seu uso no mapeamento de classes de solos e, quando usada para predição numérica, é denominada de árvore de regressão (RT) como na aplicação de mapeamento de atributos do solo (estoque de carbono, composição granulométrica dentre outros) (BREIMAN et al., 1984).

As árvores de decisão (ADs), representam um conjunto de regras sobre uma sequência hierárquica com a finalidade de particionar os dados. A mais importante função é sua capacidade de converter processos de decisões complexos em uma série de decisões simples (KHEIR et al., 2010). O objetivo das ADs é separar observações em grupos cada vez menores e homogêneos em relação ao resultado de interesse, como por exemplo: classe ou atributos de solo. (BREIMAN et al., 1984).

Os modelos de AD são apresentados em uma estrutura constituída de um nó “raiz”, de nós “internos” que apresentam a indicação de qual variável ou de quais valores das variáveis preditoras foram divididos. Os nós que apresentam a predição da variável resposta, são chamados de “folhas”, “nó terminal" ou “nó decisão", a raiz da árvore será o atributo que tiver maior ganho de informação (GAMA, 2004). A Figura 6 mostra as decisões realizadas por uma árvore de classificação (Figura6A) e regressão (Figura6B) aplicada ao MDS. Na Figura 6 o nó raiz é representado por um círculo, os nós internos são representados como retângulos e as folhas são indicados como triângulos que é o resultado da predição. Na Figura 6A podemos observar as decisões tomadas para as predições das classes de solos e Figura 6B atributo do solo (silte).

Dentre as principais vantagens, podemos dizer que as ADs possuem uma teoria mais simples do que outras técnicas, gerando modelos mais facilmente interpretados e com um tempo computacional reduzido, além de realizar seleção de variáveis, ou seja, se uma variável não é importante ela não a utiliza. (COMLEY & DOWE, 2005).

O maior problema da AD é a forte tendência ao sobreajuste ou também chamado de “overfitting”, que é o ajuste demasiado dos dados de treinamento e também um dos principais desafios enfrentados ao se modelar árvores de decisão. Para prevenir o sobreajuste pode-se definir restrições no tamanho da árvore e/ou podar a árvore (BREIMAN et al., 1984).

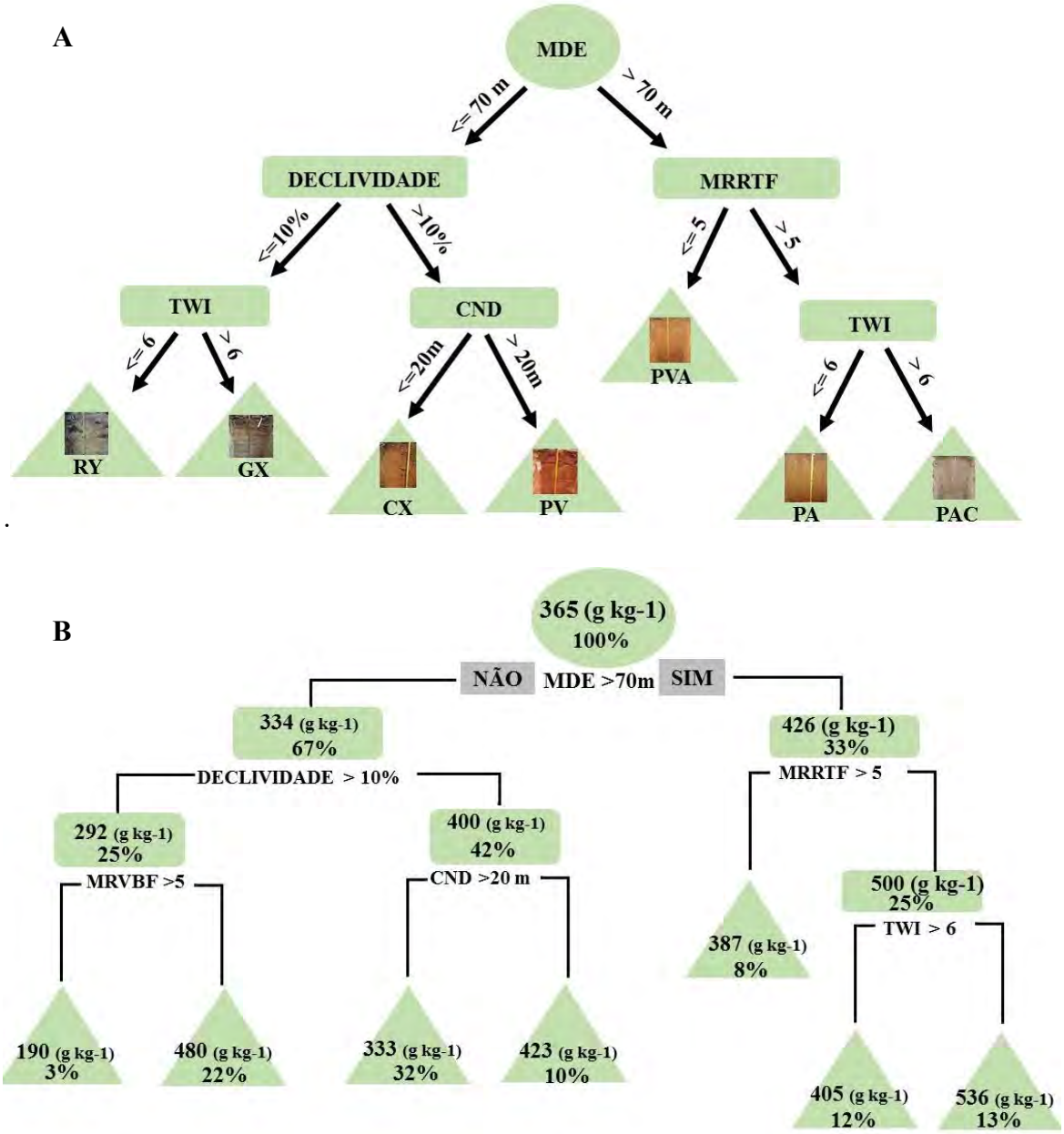


Figura 6. Exemplos de árvores de decisão. (A) Exemplo de árvore de classificação na predição de classes de solo. (B) Exemplo de árvore de regressão para predição de silte. (Fonte: Ilustração elaborada pela Autora).

Alguns estudos apontam para o uso de AD como ferramentas preditoras que apresentam bons resultados no MDS.

Crivelenti et al. (2009) para mapear classes de solos, utilizaram modelo digital de elevação (MDE) que foram extraídos os atributos morfométricos e aplicado o modelo de AD. A acurácia geral do modelo para mapeamento de classes de solos foi de 61%.

Pinheiro et al. (2015) modelaram unidades de mapeamento de solos em uma bacia hidrográfica, no Estado do Rio de Janeiro, que apresentava grande variação de condições de paisagem. Utilizaram como covariáveis atributos morfométricos derivados do MDE e índices derivados de imagem Landsat TM 5. A modelagem realizada por AD obteve um valor Kappa de 0,83.

Taghizadeh-Mehrjardi et al. (2014) utilizaram modelos de AD para predição espacial de classes de solo em uma área localizada na região árida no centro do Irã. Os autores utilizaram

covariáveis de MDE, e da imagem Landsat ETM além de um mapa de geomorfologia. Os resultados mostraram que algumas variáveis auxiliares influenciaram na predição de classes de solos tais como: índice de umidade topográfica, mapa geomorfológico, MRVBF, elevação e algumas bandas das imagens do Landsat ETM. A acurácia do modelo foi 67,5%.

Mehrabi-GoharI et al. (2019), usando RT obtiveram bons resultados no mapeamento de textura do solo em regiões áridas do Irã. Os autores mapearam os atributos em 5 profundidades diferentes (0-5, 5-15, 15-30, 30-60 e 60-100 cm). Os resultados de R^2 para a argila, areia e silte em todas as profundidades foram acima de 0,51. Sendo a melhor predição de RT em areia a 5 cm de profundidade com R^2 de 0,70.

Emadi et al. (2020) utilizaram RT com diversas covariáveis de relevo, derivados de dados climáticos, dados de solo, índices de sensores remotos e obtiveram valores de R^2 de 0,53 para RT cubista.

Pinheiro et al. (2018) encontraram bom desempenho na aplicação do modelo de RT em conjunto de dados harmonizados, com valores de 0,52 para argila (camada 0-0,05 m), 0,69 para silte (camada de 0,05-0,15) e R^2 maiores que 0,52 em todas as profundidades de areia.

b) *Random forest* (RF)

Tratando - se do modelo RF ou também chamado de Floresta Aleatória, sua forma de aprendizado e hierarquização baseia-se no conjunto de árvores de classificação, para variáveis categóricas, ou regressão (variáveis contínuas) randomizados (BREIMAN, 2001). Em ambos os casos um grande número de árvores é gerado dentro do algoritmo, e então são agregados para se obter um único valor ou classe de predição. (BREIMAN et al., 1984).

Na Figura 7 podemos observar uma ilustração simplificada de como o algoritmo RF funciona para uma modelagem. Na primeira etapa o algoritmo seleciona aleatoriamente amostras de um determinado conjunto de dados (Treinamento T), assim constrói uma árvore de decisão para cada amostra aleatória e obtém-se a predição de cada árvore de decisão. Cada árvore criada irá apresentar o seu resultado (em problemas de classificação o resultado que mais vezes for apresentado será o escolhido). Após essa etapa, realiza-se a votação para cada resultado predito e por fim seleciona o resultado da predição mais votada como resultado final.

É bem provável que as árvores de decisão criadas sejam diferentes, pois tanto na seleção das amostras, quanto na seleção das variáveis, o processo acontece de maneira aleatória. Pode-se construir quantas árvores desejar, no entanto, devemos considerar que quanto mais árvores forem criadas, maior será o custo operacional e melhor serão os resultados do modelo até determinado ponto, onde uma nova árvore não conseguirá levar a uma melhora significativa no desempenho do modelo.

Para tarefas de classificação, a saída do RF é a classe selecionada pela maioria das árvores, já para tarefas de regressão, a média das árvores individuais é retornada.

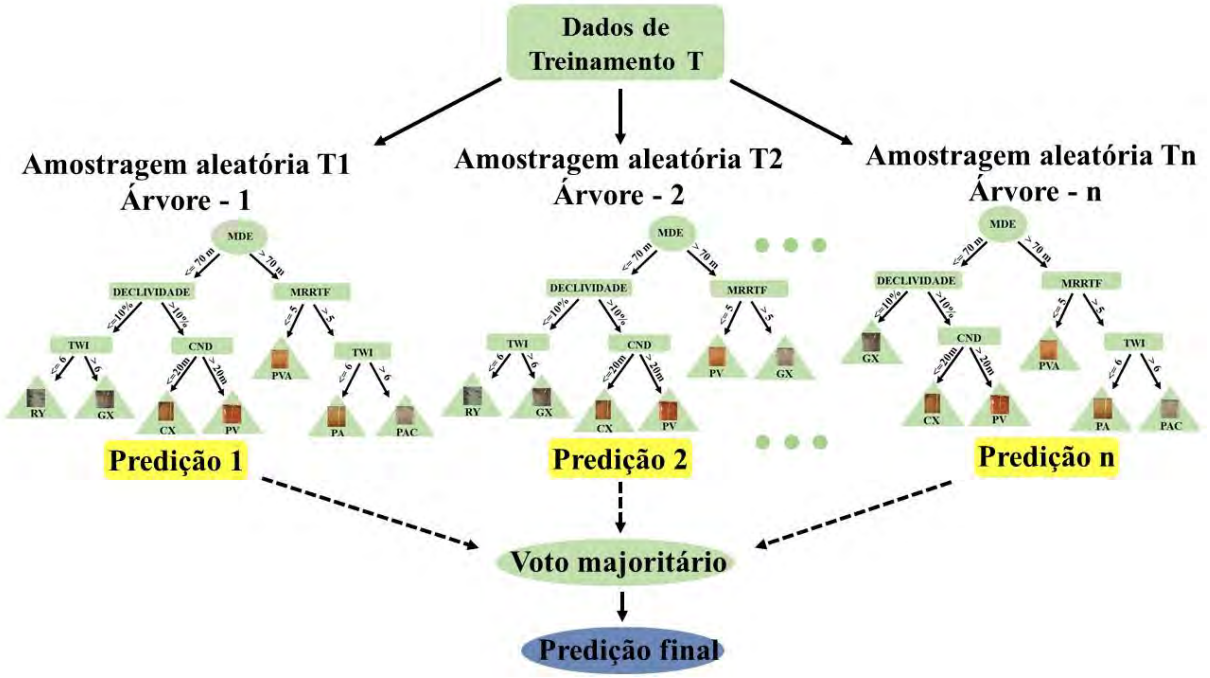


Figura 7. Disposição geral do Random Forest. (Fonte:Ilustração elaborada pela Autora).

O RF depende, basicamente, de três parâmetros definidos pelos usuários: o número de árvores na floresta, o número mínimo de pontos de dados em cada nó terminal, e o número de variáveis usadas para produzir cada árvore.

As principais vantagens do RF é ser robusto para o superajustamento, uma vez que cada árvore é treinada em uma subamostra de inicialização original dos dados. A natureza aleatória de construção de cada árvore minimiza esse sobreajuste "overfitting". Além disso, outra característica importante é a capacidade de fornecer medidas de importância das covariáveis na predição (ARUN & LANGMEAD, 2005; BREIMAN et al., 1984; HEUNG et al., 2014).

A principal desvantagem do RF, é a capacidade de interpretação limitada, sendo muitas vezes chamado de abordagem de "caixa preta", uma vez que a relação entre preditores e a resposta não pode ser examinada individualmente para cada árvore na floresta (BREIMAN, 2001; GISLASON et al., 2006; GRIMM et al., 2008).

Alguns estudos realizados com algoritmo RF no MDS podem ser citados:

Bhering et al. (2016) utilizam RF na modelagem de atributos do solo: areia, argila e carbono orgânico (CO). Como covariáveis ambientais derivaram índices de imagem Landsat 5 e atributos morfométricos de MDE em duas resoluções (30m e 90m). Os autores relataram que o uso de covariáveis preditoras, como atributos morfométricos derivados do MDE, dados do sensor TM do Landsat 5 e a litologia da área, aliado à abordagem de RF, mostraram potencial para estimar valores de areia, argila e CO do solo com uso de uma reduzida base de dados de solos. O R^2 quanto à predição de areia, argila e CO foi respectivamente de 0,44, 0,40 e 0,33 para resolução de MDE 30 m e de 0,45, 0,46 e 0,33 para resolução de MDE 90 m.

Chagas et al. (2016a) utilizaram RF na modelagem de areia, silte e argila e como covariáveis ambientais utilizaram índices e razões de bandas como NDVI, Clay minerals, razões de bandas, b3/b2 dentre outros derivados da imagem Landsat 5. A validação do modelo RF, explicou 63%, 56% e 25% da variabilidade espacial de areia, argila e silte respectivamente.

Chagas et al. (2016b) aplicaram o algoritmo de RF para modelar areia, silte, argila, capacidade de campo e ponto de murcha permanente a 0 – 20 cm de profundidade, utilizando covariáveis e índices obtidos de imagens do satélite Landsat. Os resultados de R^2 do modelo

foram: moderado para a areia = 0,64, argila = 0,56, CC = 0,52 e valor de R^2 fraco para PMP = 0,42 fraco e silte = 0,27.

Heng et al. (2015) mapearam propriedades do solo da África a uma resolução de 250 m através dos modelos de RF e Regressão Linear. Os autores utilizaram conjuntos de amostras pontuais em combinação com um grande número de covariáveis. Os resultados da validação cruzada demonstraram que o algoritmo RF superou consistentemente o algoritmo de regressão linear, com decréscimos médios de 15 a 75% no erro médio quadrático (RMSE) entre as propriedades e profundidades do solo. A porcentagem de variação explicada variou entre 40 - 85% para RF e entre 10 - 45% para regressão linear.

Sreenivas et al. (2014) realizaram mineração de dados para avaliação espacial de estoque carbono orgânico do solo (SOC) com modelo RF. As covariáveis clima, NDVI, cobertura do solo, tipo de solo e topografia foram usadas como base para a modelagem de SOC a 30 cm. Os resultados dos experimentos indicaram que das várias variáveis de entrada usadas para modelar SOC, o uso e cobertura de terra foi o fator mais significativo que influenciou o SOC com um escore de importância distinta de 34,7 seguido por NDVI com uma pontuação de 12,9. O valor de R^2 para o modelo RF foi de 0,86.

Tesfa et al. (2009) usaram modelos estatísticos para predição da profundidade do solo em uma microbacia montanhosa na região semiárida, baseados na relação entre profundidade do solo com atributos topográficos e cobertura de terra. Os atributos topográficos foram derivados de um MDE, os atributos de cobertura da terra de imagens de sensoriamento remoto Landsat TM e fotografias aéreas de alta resolução. O modelo RF explicou cerca de 50% na predição da profundidade do solo ao longo da bacia.

Stum et al. (2010) realizaram modelagem de classes de solos em uma bacia hidrográfica árida e semiárida do oeste de Utah, utilizando amostras coletadas em campo, imagens Landsat 7 e covariáveis derivados de um MDE de 10m de resolução espacial, sendo estas, as que apresentaram uma maior importância de predição. Para o autor, o RF funcionou como um poderoso algoritmo para predição de classes de solo e seus resultados facilitaram ainda mais a compreensão das relações solo-paisagem.

Pinheiro et al. (2015) modelaram unidades de mapeamento de solos em uma bacia hidrográfica, no Estado do Rio de Janeiro, que apresentava grande variação de condições de paisagem. A modelagem realizada por RF obteve valor de Kappa 0,96. Os autores utilizaram covariáveis derivadas do MDE tais como; altimetria, declividade, curvatura, índice topográfico composto, distância euclidiana da hidrografia, e índices derivados de sensoriamento remoto tais como: clay minerals, iron oxide e índice de vegetação da diferença normalizada (NDVI).

Wiesmeier et al. (2011) avaliaram o modelo RF para predição de SOC, carbono total (Ctot), nitrogênio total (Ntot) e enxofre total (Stot) para uma bacia hidrigráfica localizada na Mongólia interior da China, ao norte de Pequim. A acurácia das predições da modelagem de RF e os mapas gerados foram aceitáveis, com R^2 de 0,67, 0,78, 0,75 e 0,74 para Stot, Ntot, Ctot e SOC respectivamente.

2.4.2 *Support vector machine*

O *Support Vector Machine* (SVM) pode ser considerado uma técnica computacional de aprendizado de máquina supervisionado, cujo enfoque é o reconhecimento de padrão, podendo ser usado para tarefas de classificação e regressão. O algoritmo tem como objetivo determinar limites de decisão onde haja separação ótima entre as classes com minimização dos erros, permitindo dessa forma a maximização dos limites do hiperplano de separação ótima (optimal hyperplane). Os pontos próximos ao limite de separação são definidos vetores de suporte (support vectors) e a separação ótima entre classes ocorre por meio de um hiperplano condicional (L), tal que este plano é orientado para maximizar a margem (distância entre as

bordas, L1 e L2) e pelo ponto mais próximo de cada classe Figura 8 (FARIA & FERNANDES FILHO, 2013; NASCIMENTO et al., 2009).

Em outras palavras, o que um SVM faz é encontrar uma linha de separação, mais comumente chamada de hiperplano entre dados de duas classes. Essa linha busca maximizar a distância entre os pontos mais próximos em relação a cada uma das classes linearmente separáveis Figura 8.

No entanto, se os dados não forem linearmente separáveis, o algoritmo conta com uma técnica chamada truque de kernel. Utilizando as funções de kernel, o algoritmo transforma os problemas não lineares em um espaço de mais alta dimensão, possibilitando que os dados não separáveis linearmente, tornem-se separáveis no espaço de características. O algoritmo possui quatro funções de kernel: linear, sigmoide, polinomial e radial (SANTOS, 2002).

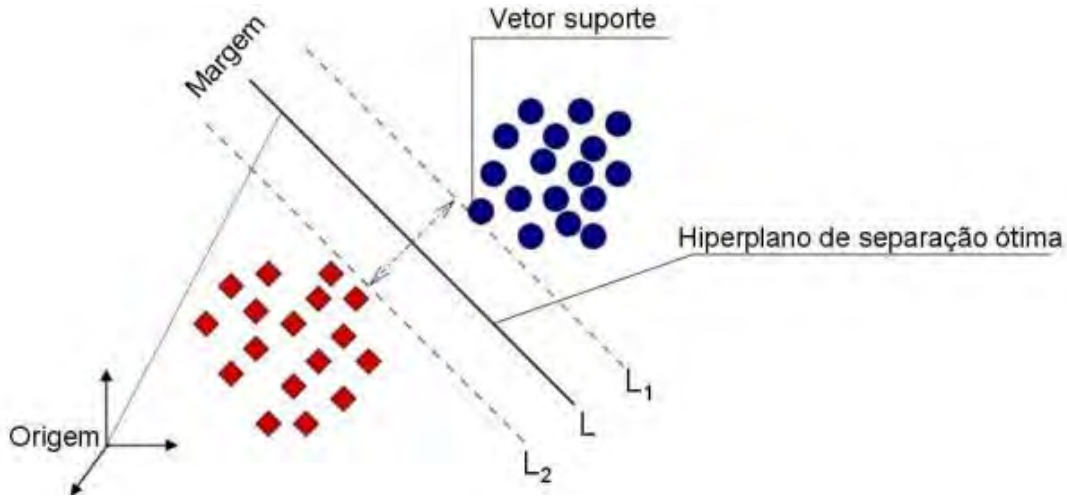


Figura 8. Esquema de classificação por meio do Support Vector Machines. (Fonte: imagem gerada no programa RStudio, modificado de HUANG et al., 2002; MELGANI & BRUZZONE, 2004).

Uma das vantagens do SVM, é a sua capacidade de processar rapidamente conjuntos de alta dimensão. Contudo, é necessário definir um bom Kernel. Como desvantagem, o resultado do SVM é dificilmente interpretável e, conforme o tamanho do dataset vai aumentando, o tempo necessário para fazer os cálculos cresce muito rapidamente e a interpretabilidade diminui mais rápido ainda.

Padarian et al. (2014) na Austrália, utilizaram técnicas de modelagem como Regressão, Cubist e SVM para modelar água disponível do solo, usando diferentes combinações de informações ambientais: topográficas, climáticas, solos, imagens de landsat e raios gama espectrometria. Em geral, dentre os modelos testados o SVM produziu a melhor acurácia. Tanto no limite superior de drenagem quanto no limite inferior da cultura nos cinco intervalos de profundidade (0 a 5; 5 a 15; 15 a 30; 30 a 60 e 60 a 100 cm) os valores de R^2 do SVM em profundidade variaram na média de 0,49 a 0,69 para valores de R^2 menores em profundidades maiores.

Meier et al. (2018) avaliaram o desempenho de oito algoritmos de aprendizado de máquina entre eles SVM com função linear e polinomial para o mapeamento de solos em uma área tropical montanhosa de um assentamento rural na região da Zona da Mata de Minas Gerais no Brasil. Foram utilizados como covariáveis atributos morfométricos gerados a partir de um MDE, juntamente com imagens de satélite Landsat-8 e mapas climáticos. O SVM mostrou desempenho similar aos outros algoritmos sem diferença estatística significativa (Kappa 0,45 para SVM com função linear e Kappa 0,42 para SVM polinomial).

Vasconcelos et al. (2018) avaliaram a eficiência de AD, SVM e kNN na classificação de solos, no 1º nível categórico do Sistema Brasileiro de Classificação de Solos (SiBCS). Os três algoritmos de AM apresentaram excelentes resultados para a classificação das 13 classes de solo, entretanto, o algoritmo SVM apresentou, uma acurácia levemente superior aos demais. O modelo de classificação baseado no algoritmo SVM obteve uma acurácia acima de 90% para todas as 13 ordens de solos do SiBCS.

Tabela 1. Vantagens e desvantagens dos algoritmos de aprendizagem de máquinas utilizados na modelagem das classes e atributos do solo.

Modelos	Vantagens	Desvantagens	Referências
AD/AR	Técnica robusta; fácil compreensão; processam grande números de dados em curto espaço de tempo; Dados requerem o mínimo de preparação; Seleção de variáveis; Novas opções podem ser adicionadas às árvores existentes; relaciona variáveis categóricas e nominais;	Podem se tornar excessivamente complexas; Sobreajuste ("overfitting")	Adhikari, (2014), Comley & Dowe, (2005), Crivelenti et al., (2009), Giasson et al. (2013); Pinheiro, (2015), Ten Caten et al. (2011).
RF	Natureza não paramétrica, alta precisão de classificação e capacidade de determinar a importância da variável preditora; variáveis preditoras podem ser tanto contínuas como categóricas; robusto ao ruído em preditores e assim, não necessita de pré-seleção de covariáveis; robusto para o superajustamento; considerado acurado e robusto devido ao número de arvores no processo	Limitada interpretação dos resultados, uma vez que as relações entre os preditores e as respostas não podem ser examinadas individualmente para cada árvore na floresta.	Bhering et al. (2016), Breiman, (2001), Chagas et al. (2016b, 2016a), Hengl et al. (2015), Pinheiro, (2015), Rodriguez-Galiano et al. (2012).
SVM	Funciona muito bem com margem de separação clara. É eficaz nos casos em que o número de dimensões é maior que o número de amostras. Usa um subconjunto de pontos de treinamento na função de decisão (chamados de vetores de suporte), portanto, também é eficiente em termos de memória. Versátil: diferentes funções de Kernel podem ser especificadas para a função de decisão. Kernels comuns são fornecidos, mas também é possível especificar kernels personalizados.	Necessário definir um bom Kernel.	Kovačević et al. (2010), Padarian et al. (2014).

AD- Árvore de Decisão, AR- Árvore de Regressão, RF- *Random Forest*; SVM- *Support Vector Machine*.

3. CHAPTER I

PREDICTING SOIL CARBON STOCK IN REMOTE AREAS OF THE CENTRAL AMAZON REGION USING MACHINE LEARNING TECHNIQUES

3.1 RESUMO

Estimativas do estoque de carbono orgânico do solo (SOCS, inglês) de regiões sob a Floresta Amazônica são essenciais para subsidiar estudos do aquecimento global, visto que este bioma desempenha um papel importante no sequestro e liberação de SOCS. O uso de covariáveis derivadas de sensores remotos combinados com algoritmos de aprendizado de máquina (AM) tem se mostrado promissor para mapear tipos de solo e seus atributos em grandes áreas. Este estudo explora a viabilidade de usar o conhecimento existente de SOCS derivado de uma densidade relativamente baixa e conjunto de dados irregulares para mapear uma grande área de 13.440 km², localizada em uma região remota sob a Floresta Amazônica. Os objetivos deste estudo foram: 1- avaliar dois tipos diferentes de abordagem de amostragem (Área de Referência - AR e Área Total - AT) para prever SOCS em profundidades de 30 (SOCS30) e 100 cm (SOCS100); 2- avaliar dois métodos de seleção de covariável: "método wrapper", que se baseia na inferência feita por um modelo AM calibrado, e "seleção prévia de covariável" como etapa de pré-processamento, antes da calibração do AM; 3- avaliar a transferibilidade e o desempenho de três algoritmos de AM: "Regression Tree" (RT), "Random Forest" (RF) e "Support Vector Machine" (SVM). O local do estudo foi dividido em três blocos, denominados blocos Urucu, Araracanga e Juruá. O conjunto de dados consistiu em 120 observações de SOCS30, SOCS100 e 21 covariáveis (20 covariáveis de relevo e banda P do radar) que foram abordadas em dois tipos diferentes de dataset: AR e AT. Usando o conjunto de dados da AR, 96 observações foram usadas para treinar os algoritmos e 24 para validação, enquanto usando o conjunto de dados da AT, 90 observações foram usadas para treinamento (75%) e 30 para validação (25%). A similaridade entre a paisagem dos blocos e a da AR foi avaliada por meio do índice geral de Gower e estatística descritiva das covariáveis. Os resultados mostram que o uso da seleção prévia das covariáveis, combinada com o uso de um conjunto de dados da AR, permite desenvolver modelos mais acurados para predizer SOCS30 e SOCS100. De acordo com o índice geral de Gower, a AR possui alta similaridade com os blocos Urucu, Araracanga e Juruá. Entretanto, as estatísticas mostraram que, aumentando a distância da AR, algumas covariáveis de relevo são mais diferentes. Embora a dissimilaridade aumente proporcionalmente à distância em relação a AR, o desempenho geral dos modelos e a validação dos blocos Urucu e Juruá, separadamente, foi melhor do que o conjunto de dados de validação total e o bloco Araracanga. Os modelos desenvolvidos para predizer o SOCS100 apresentaram maior acurácia e transferibilidade do que aqueles desenvolvidos para predizer o SOCS30. O mapa do SOCS30 foi gerado apenas para o bloco Urucu e o melhor desempenho foi obtido usando o algoritmo RT (R2 = 0,32). O mapa SOCS30 apresentou uma faixa de 2,29 kg C. m-² a 4,04 kg C. m-². O algoritmo de RF gerou os mapas mais precisos de SOCS100 para os blocos de Urucu e Juruá (R2 = 0,70 e 0,51, respectivamente). Os valores de SOCS100 dos mapas gerados para a região do bloco de Urucu variam de 3,89 kg C. m-² a 10,64 kg C .m-², enquanto para o bloco Juruá variam de 5,03 kg C. m-² a 10,42 kg C. m-². Apesar da baixa densidade de observação do conjunto de dados disponível, os resultados mostram não só a importância dos algoritmos de AM para mapear o SOCS, mas também o uso de conhecimento pedológico especializado gerado em uma AR para apoiar uma prévia seleção de covariável antes de calibrar os algoritmos de AM.

Palavras-chave: Área de referência. Índice de Gower. Regression tree. Random forest. Support vector machine.

3.2 ABSTRACT

Estimates of the soil organic carbon stock (SOCS) of regions under the Amazon Rainforest are essential to support studies of global warming, as this biome plays an important role in the sequestration and release of SOCS. The use of covariates derived from remote sensors in combination with machine learning (ML) algorithms has been shown to be promising for mapping soil types and their attributes in large areas. This study explores the feasibility of using the existing knowledge of SOCS derived from a relatively low density and irregular dataset to map a large area of 13,440 km² located in a remote region under the Amazon Rainforest. The objectives of this study were: 1- to evaluate two different types of sampling approach (Reference Area - RA and Total Area - TA) to predict SOCS at depths of 30 (SOCS30) and 100 cm (SOCS100); 2- to evaluate two categories of covariate selection: "wrapper method", which rely on the inference made by a calibrated ML model, and "previous covariate selection" as a pre-processing step, before calibrating the ML; 3- to evaluate the transferability and the performance of three ML algorithms: regression tree (RT), random forest (RF) and support vector machine (SVM). The study site was divided into three blocks, called Urucu, Araracanga and Juruá blocks. The dataset consisted of 120 observations of SOCS30, SOCS100 and 21 covariates (20 relief covariates, and radar P-band) that were addressed in two different types: RA and total area TA. Using the RA dataset, 96 observations were used for training the algorithms and 24 for validation, while using the TA dataset, 90 observations were used for training (75%) and 30 for validation (25%). The similarity between the landscape of the blocks and that of the RA was evaluated using the general Gower index and descriptive statistics of the covariates. The results show that the use of previous covariates selection, combined with the use of a dataset of the RA, allows to develop more accurate models to predict SOCS30 and SOCS100. According to the general Gower index, the RA has high similarity with the Urucu, Araracanga and Jurua blocks, however, the statistics show that increasing the distance from the RA, some relief covariates are more different. Although the dissimilarity increases proportionally to the distance from the RA, the overall performance of the models and the validation of the Urucu and Juruá blocks, separately, was better than the total validation dataset and the Araracanga block. The prediction models developed to predict SOCS100 presented both higher accuracy and transferability than those developed to predict SOCS30. The map of SOCS30 was only generated to Urucu Block and the best performance was achieved using RT algorithm (R2=0.32). The SOCS30 map presented a range from 2.29 kg C. m-² to 4.04 kg C. m-². The RF algorithm generated the most accurate maps of SOCS100 for the Urucu and Juruá Blocks (R2=0.70 and 0.51, respectively), The SOCS100 values of the maps generated to Urucu Block region range from 3.89 kg C. m-² to 10.64 kg C. m-², while for the Jurua block range from 5.03 kg C. m-² to 10.42 kg C. m-². Despite the low observation density of the dataset available, the results show not only the importance of ML algorithms to map SOCS, but also the use of specialized pedological knowledge generated in a RA to support a previous covariate selection before calibrating the ML algorithms.

Keywords: Reference area. Gower Index. Regression tree. Random forest. Support vector machine.

3.3 INTRODUCTION

Soil organic matter contains more organic carbon than global vegetation and the atmosphere combined. In numerical terms, the global soil carbon pool (3,500 to 4,800 Pg C) is around 8 and 5 times larger than the vegetation (420 to 620 Pg C) and the atmospheric (829 Pg C) pools, respectively. This superiority of the global soil carbon pool means that even a small proportion of carbon contained in soil organic matter can cause quantitatively relevant variations in the atmospheric concentrations of greenhouse gases (LEHMANN & KLEBER, 2015).

Although the estimative of soil organic carbon stocks (SOCS) represents a baseline to support environmental policy, in some regions, like Brazilian Amazon Rainforest, the estimates remain insufficiently detailed. For example, according to Batjes & Dijkshoorn (1999), the SOCS was 25 Pg C and 46.5 Pg C at the soil depth up to 30 and 100 cm, respectively. Similar results were also found for the soil depth of 1 meter (47 Pg of SOCS in the first 1 m, and 44% of this carbon is in the top 0.2 m) by Moraes et al. (1995). According to Ceddia et al. (2015), most of the available data in this region comes from the same source, the RADAMBRASIL project, which generates very sparse soil data. The limited soil data is due to the presence of a dense forest that is only accessed by boats, helicopters and airplanes. These factors demands specific logistics and increases the costs of work in this region (CEDDIA et al., 2017).

In recent years, the soil science area has witnessed a significant increase in research on Digital Soil Mapping (DSM). During this time, a great diversity of algorithms has been used to predict quantitative and categorical soil data (properties and types). Digital mapping approaches to SOCS range from simple linear statistical models, geostatistical, hybrid techniques to advanced and complex machine learning (ML) ones (LAMICHHANE et al., 2019). ML techniques refer to a large class of data-driven algorithms employed primarily for data mining and pattern recognition purposes and are now frequently used for regression and classification tasks in all fields of science. The advantage of using ML is that it is not conditioned to follow any statistical assumptions and the algorithms can also handle a large number of cross-correlated covariates (collinearity) as a predictor (WADOUX et al., 2020).

Considering these aspects, one important strategy to create and expand more detailed SOCS maps in poorly accessed areas is the use of DSM techniques (especially ML), extracting and optimizing the information from already existing detailed soil survey along the region (Reference Areas - RA). A RA purposely chosen would be sufficient to identify all soil types and attributes in a larger area and determine their spatial relationships.

Considering the potential of ML and remote sensing data to assist in DSM in remote and densely vegetated regions, we hypothesize that SOCS (0-30 cm and 0-100 cm) estimative in Central Amazon Rainforest can be improved by adding relief and radar covariate on robust ML algorithms. Thus, the objectives were to evaluate: 1- two different types of sampling approach (Reference Area - RA and Total Area - TA) to predict SOCS at depths of 30 (SOCS30) and 100 cm (SOCS100); 2- two categories of covariate selection: "*wrapper method*", which rely on the inference made by a calibrated ML model, and "*previous covariate selection*" as a pre-processing step, before calibrating the ML; 3- the transferability and the performance of three ML algorithms: regression tree (RT), random forest (RF) and support vector machine (SVM).

3.4 MATERIAL AND METHODS

3.4.1 Study area

The study was carried out in a region under the Amazon Rainforest between the municipalities of Carauari and Coari, approximately 640 km from Manaus, capital of the Amazonas State. The study area is approximately 13,440 km², between the parallels 4°0' and 6°0'S and 67°0' and 64°00'W (Figure 9A). According to Köppen, the climate is classified as Af (equatorial, with the coldest month temperature above 20°C, average annual rainfall of 2500 mm and dry period not pronounced). The region is remote, being only possible to access by air and river transports.

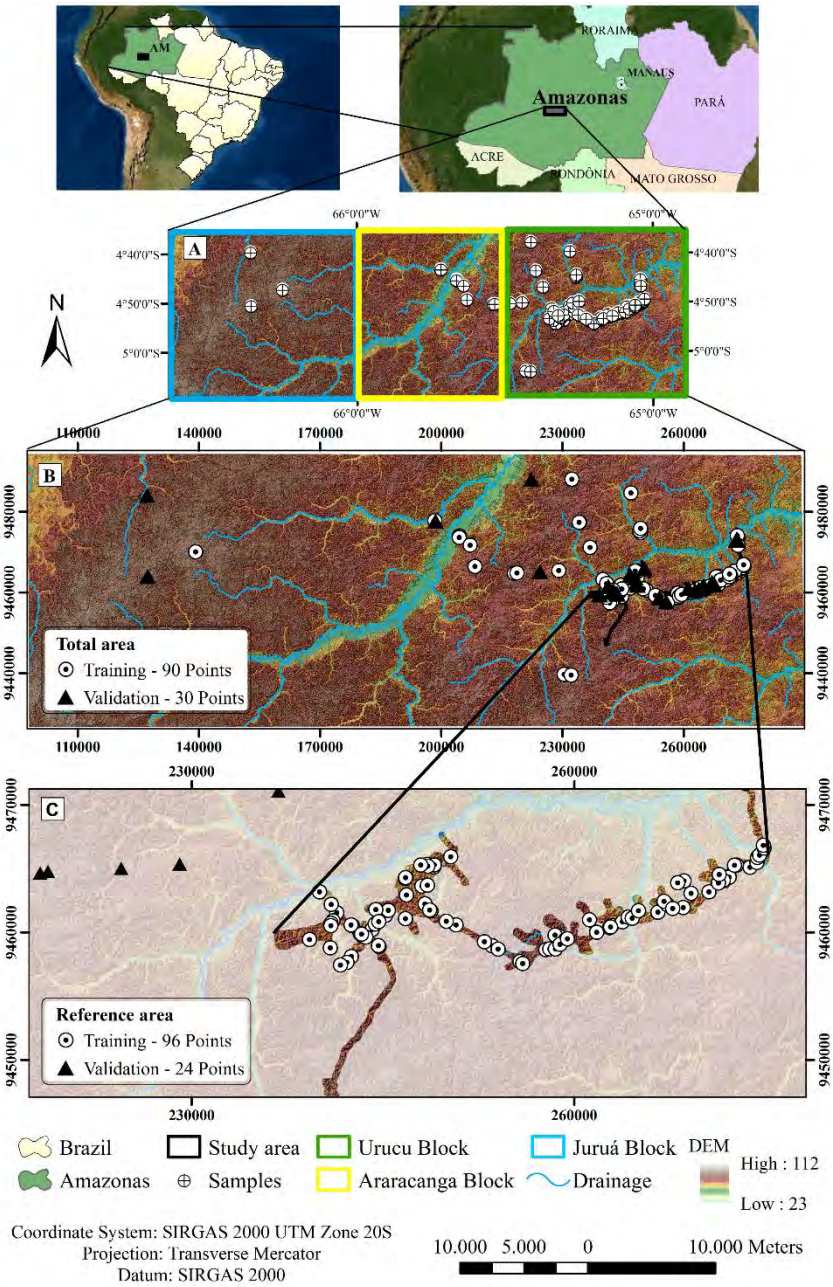


Figure 9. Study Area Map. (A) The location of the study area in the Central Amazon, Brazil. (B) Sampling based on the total area, 75% training 25% validation. (C) Sampling based on reference area approach.

3.4.2 The approaches to model SOCS and the datasets used

Primarily, the challenge was to explore existing knowledge of SOCS to map a territory of 13.440 km². This entire territory covered by the original Amazon Forest, was divided into three blocks (Urucu, Araracanga and Juruá) for the purpose of analyzing the transferability of models. The criterion for dividing the blocks, shown in the upper part of Figures 9 (Figure 9A), was based on the way in which relief data and radar images were generated and made available by Petrobras (Figure 9A).

This work uses the soil data collected in two moments. The first set of data was generated in 2010, as part of a detailed soil survey in an area of 80 km² located around the Urucu River - Coari Municipality/AM (Base de Operações Geólogo Pedro de Moura BOGPM - Petrobras-BR). This area is called the Reference Area (RA -Figure 9C), which resulted in a database of 96 soil profiles containing SOCS up to the depths of 30 and 100 cm (CEDDIA et al., 2015). In 2018, a new fieldwork visited 16 remote clearings (up to 100 km far from the RA) using a helicopter. A total of 41 new soil profiles were described with soil sample collections. Out of 41 profiles, in 24 the SOCS was also calculated up to depths of 30 and 100 cm.

Considering the two field works (2010 and 2018), a dataset with 120 SOCS sample (96+24) were used. The methodological strategy to predict SOCS for each soil depth (30 and 100 cm) is presented in the flowchart of Figure 10. The first dataset followed the RA approach. In this case, the 96 data of SOCS of the RA (80km²) was used to train the ML prediction models (dataset 1). The 24 new SOCS sample points of the remote clearings were used to validate de models (external validation - triangles in black color - Figure 9C) which is an area outside the RA. The second dataset (dataset 2) refers to the random division of the total dataset, 120 soil samples, that covers the TA inside and outside RA. In this case, a random draw was carried out to make the subdivision considering 75% of the data for model training (90 samples) and 25% for validation (30 samples - Figure 9B).

The soils were classified according to the Brazilian Soil Classification System (SANTOS et al., 2018). The soil mapping units of the study site, as well as the number of profiles and frequency, are shown in Table 2. Most soils have low base content, high aluminuim content and medium to high sand content. Some soils in the region have characteristics of hydromorphism, especially those close to the floodplain of water courses and flattened tops.

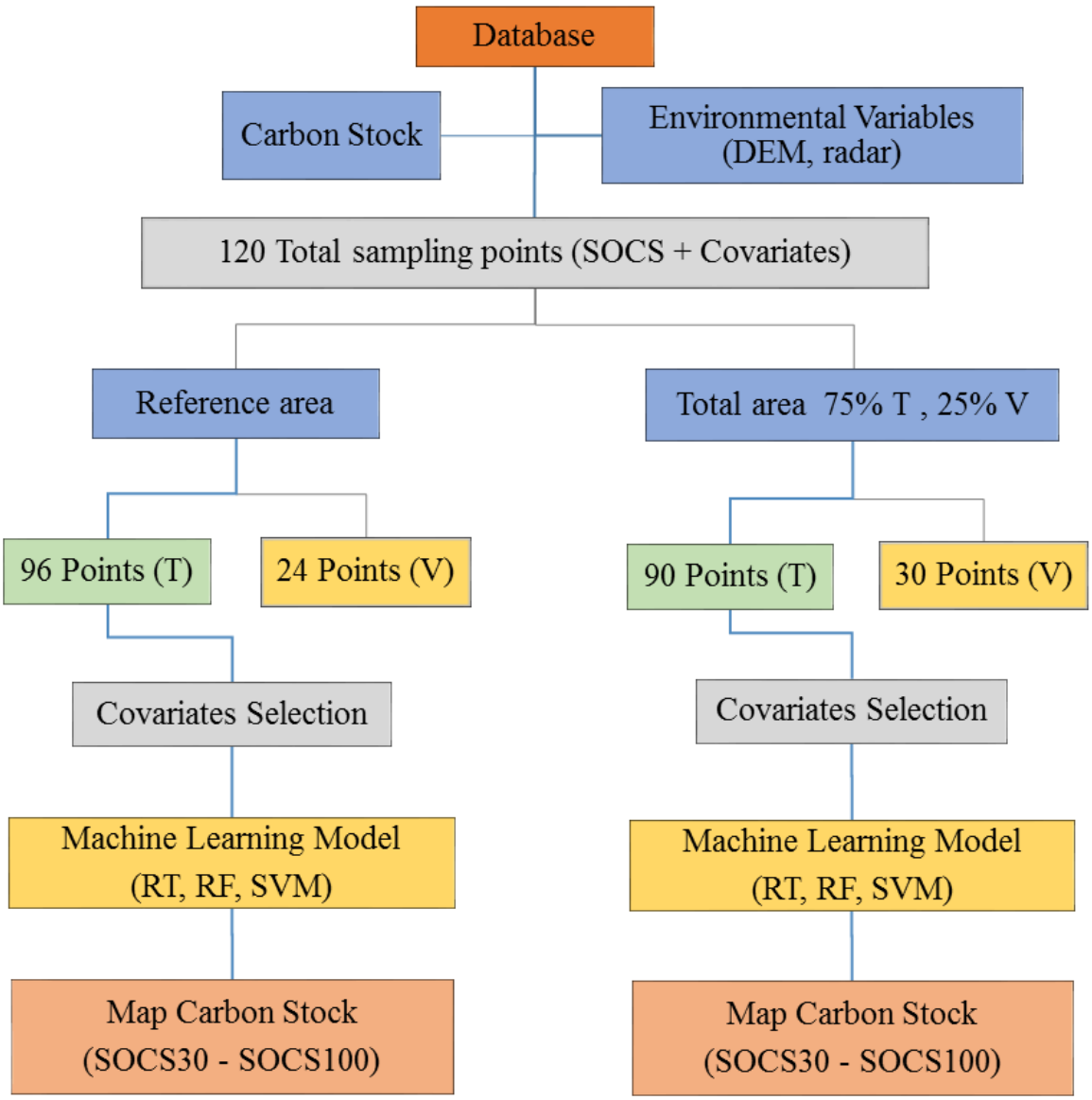


Figure 10. Flowchart with the methodological strategy for mapping carbon stocks up to 0.30 (SOCS30) and 1.0 m (SOCS100); T-Training; V-Validation; RT- Regression Tree, RF-Random Forest; SVM-support vector machine.

Table 2. Number of soil profiles (n) and frequency of soil in the visited sites.

*SiBCS[a]	**Soil Taxonomy	**WRB	SOCS30	SOCS100	n	Frequency (%)
Argissolo Amarelo	*Ultisols*	*Acrisols; Lixisols*	3.34 ± 0.93	7.91 ± 2.69	34	28.3
Argissolo Vermelho	*Utisols (Typic Rhodustults)*	*Acrisols; Lixisols*	3.57 & 4.21	8.05 & 8.83	2	1.66
Argissolo Vermelho Amarelo	*Ultisols*	*Acrisols; Lixisols*	3.25 ± 0.77	7.34 ± 2.03	24	20.0
Argissolo Acizentado	*Ultisol (Hapludult)*	*Haplic Lixisol*	2.14 ± 3.81	4.12 ± 8.45	2	1.70
Cambissolo Háplico	*Inceptisols*	*Cambisols*	3.57 ± 1.04	8.37 ± 2.42	40	33.3
Cambissolo Flúvico	*Entisols (Fluvents)*	*Fluvisols*	3.01 & 3.37	6.32 & 9.49	2	1.70
Espodossolos Humilúvicos	*Humiluvic spodosol.*	*Podzols*	3.84	12.30	1	0.83
Espodossolos Ferri- Humilúvicos	*Ferrohumiluvic spodosol.*	*Podzols*	2.12 ± 1.02	7.68 ± 5.29	3	2.5
Neossolo Quartzarênico	*Entisols (Quartzipsamments)*	*Arenosols*	2.95	7.05	1	0.83
Neossolo Flúvico	*Entisols (Fluvents)*	*Fluvisols*	2.20	5.34	1	0.83
Planossolo Háplico	*Ultisols (Albaquults)*	*Planosols*	3.46 & 5.35	6.78 & 7.60	2	1.70
Gleissolos Háplicos	*Entisols (Aquents)*	*Gleysols; Stagnosols*	2.41 ± 0.79	5.52 ± 2.22	7	5.83
Gleissolos Melânicos	*Entisols (Fluvaquentic Humaquepts)*	*Umbric Gleysols*	6.22	12.47	1	0.83
Total					120	100

[a] Brazilian Soil Classification System. **Note: This is partial equivalence between soil classes at high categorical level in SiBCS, WRB and Soil Taxonomy. SOCS (kg C m^{-2}).

3.4.3 Soil organic carbon stock

For each of the 120 soil profiles, the calculation of the SOCS was performed in the 0–30 and 0–100 cm depth layers. The calculation methodology followed the same applied by Ceddia et al. (2015). The classical way of calculating carbon densities (carbon mass per area) for a given depth consists of summing carbon stocks by horizon, determined as a product of Bulk density (BD), SOC concentration, and horizon thickness, according to Bernoux et al. (2002) Eq. (1):

$$SOCS = (SOC*BD*T) \quad (1)$$

Where:

SOCS soil organic carbon stock (kg C m−2);
SOC soil organic carbon (g kg−1);
BD soil bulk density (Mg m−3);
T horizon thickness (m).

The SOC was measured by wet combustion, according to the methodology proposed by Walkley and Black (1934). In the soil survey, the soil profiles were divided into horizons A-, B- and C. In most cases, the calculations involved two horizons, where the first horizon was totally above 30 cm, and the second crossed at this depth 30 cm or 100 cm. When a horizon crossed the 30 or 100 cm limit, only the portion of the horizon that was above that depth was used to calculate its SOCS (CEDDIA et al., 2015).

3.4.4 Remote sensing covariate

In view of the dense coverage of the Amazon Rainforest in the area, to obtain covariates of the relief and soil surface with the potential to make SOCS prediction, data generated by the OrbiSAR-1 radar was used, corresponding to an Airborne Synthetic Aperture Radar (SAR), developed by *Orbisat da Amazônia Indústria e Aerolevantamento SA*. This system works in the frequency range of 9.35 to 9.75 GHz and has a bandwidth of 400 MHz. Only images referring to the P band of the radar were used, a total of 84 images were needed to cover the entire area.

The P-band microwave radar images (72.0 cm wavelength) of the sensor were used to derive the backscatter coefficient (σ°) of the polarization (HH), which indicates the amount of microwave energy reflected by the target which returns to the sensor antenna per unit area in decibels (dB). All calibration and radiometric correction processing were performed using ENVI software. Reflectors points in the ground were for radiometric calibration. The equation used for calculating the zero sigma of triedral reflectors Eq. (2).

$$\sigma° = 4\pi.a4 / 3\lambda2 \quad (2)$$

For calibration in one scene Eq. (3):

$$\sigma° = (\sigma/\text{pixel area}) = \sigma/(\rho az \,.\, \rho r) = (4\pi.a4) /(3\lambda2.\rho az \,.\, \rho r) \quad (3)$$

fc = (σ°.senθι)/(DN reflector peak amplitude)
DN= digital number
fc= calibration factor
Where:
σ0(i,j) = (DN amplitude (i,j) * fci)
σ0(i,j)dB = 20 log (DN amplitude (i,j) * fci)

$\sigma 0(i,j) = (DN\ intensity\ (i,j) * fci)$
$\sigma 0(i,j) = 10\ log\ (DN\ intensity\ (i,j) * fci)$

3.4.5 Digital elevation model and topography attributes

The digital elevation model (DEM) was obtained from the airborne radar interferometric process, all appropriate treatment was carried out, aiming a model without interpolation failures, resulting in a hydrologically consistent DEM with a 20 m final resolution. Terrain attributes and other indexes were calculated using SAGA GIS version 2.3.2, their descriptions, acronyms and unit can be seen in Table 3.

Table 3. Environmental covariates extracted from the digital elevation model.

Covariate (Unity)	Description	COD	References
Digital elevation model (m)	Represents the elevation of each cell in the model.	DEM	Hutchinson & Gallant (2000), Moore et al. (1991).
Convergence index (d)	Calculates the convergence / divergence index in relation to runoff.	CI	Conrad O. (2012)
Topographic Wetness Index (d)	Function of declivity and the contributing area per orthogonal width unit towards flow direction.	TWI	Boehner et al. (2002), Moore et al. (1993).
Relative Slope Position (0–1)	Relative slope position based on the base channel network. Defined as the position of one point relative to the ridge and valley of a slope, with a value of 0 for the bottom of the valley and 1 for the top of the ridge	RSP	Boehner &.Conrad (2008); Nguyen et al (2006).
Channel Network Distance (m)	Distance from the channel level of the local network to the terrain. It is similar to elevation and is defined as the elevation difference between the cell and the closest channel network	CND	Grimaldi et al (2007).
Channel Network Base Level (m)	Vertical distance to the base level of the channel network. It is a grid in which the value of each cell is the spatially interpolated elevation of the channel networks. It is different from sea level and water level for which there is only one value for the whole study area, whereas CNBL varies with location.	CNBL	Grimaldi et al (2007).
LS factor (d)	Attribute equivalent to the topographic factor of the Revised Universal Soil Loss Equation (RUSLE).	LSf	Boehner, J., Selige, T. (2006).
Multiresolution Index of Valley Bottom Flatness (d)	Identifies valley bottoms using a slope classification constrained to convergent areas. Indicate flat surfaces on valley bottom.	MRVBF	Gallant & Dowling (2003).
Multiresolution index of the ridge top flatness (d)	Indicate flat positions on high elevation areas.	MRRTF	Gallant & Dowling (2003).
Convexity index (d)	Terrain surface convexity.	CXI	Conrad O. (2012).
Aspect (°)	Represents exposure faces, values in degrees (0 to 360°)	ASP	Carvalho Júnior (2005).
Landforms (d)	It represents the landforms of the area.	LF	Jasiewicz & Stepinski (2013).
Profile curvature (m^{-1})	Slope profile curvature. A curvature of a normal section of the land surface by a plane. The shape of the hillside on the vertical plane (concave, rectilinear or convex).	ProfC	Hall & Olson, (1991), Gessler et al. (1995), Figueiredo (2006).

To be continued...

Continuation of **Table 3.**

Covariate (Unity)	Description	COD	References
Plan curvature (m^{-1})	Contour curvature. This section is orthogonal to the section of profile curvature at a given point on the land surface. The shape of the hillside on the horizontal plane (concave, rectilinear or convex).	PlanC	Hall e Olson, (1991), Gessler et al. (1995), Figueiredo (2006).
Valley depth (m)	Vertical distance of a base level channel network.	VD	Conrad O. (2012).
Slope Height (m)	Vertical distance from the base of the slope to the crest, or line of intersection of the two slope planes.	SH	Boehner & Conrad (2008). Gokceoglu & Aksoy(1996).
Mid Slope Positon (%)	Relative vertical distance to the mid slope valley or crest directions. The higher the relative vertical distance to the mid slope in valley or crest directions the higher this value	MSP	Bohner & Antonic (2009), Häring et al. (2012).
Slope (%)	Gradient or rate of change of elevation between neighboring cells.	S	Thompson et al. (2001), Wilson, & Gallant (2000).
Melton Ruggedness (d)	A simple flow accumulation related index, calculated as difference between maximum and minimum elevation in catchment area divided by square root of catchment area size.	MR	Marchi, L. & Fontana, G.D. (2005).
Flow Accumulation (d)	Areas where the flow is concentrated and allows the identification of water paths or flow.	FC	Mathias et al. (2020).

* COD: Code used to represent the covariate in modeling; d: dimensionless.

3.4.6 Prediction models

Three ML algorithms were applied, including RT, RF and SVM for SOCS prediction at 30 and 100 cm depth. Each algorithm can find complex relationships between SOCS and environmental covariates in different ways. A brief description of the ML techniques used in this study is presented in the sequence. The performance of an ML model is impacted by the values of its parameters. Table 4 summarizes the hyperparameters of each ML algorithms used in this study.

Table 4. Hyperparameters of used machine learning algorithms.

Algorithms	Hyperparameters	Definition	Tuning
RT	cp	A non-negative number for complexity parameter.	0.001-0.01
	method	anova	anova
RF	mtry	number of variables used to produce each tree.	1– 10
	ntree	the number of trees.	100 –1000
	nodesize	the minimum number of data points in each terminal node.	5
SVM	Kernel type	the kernel function.	polynomial
	type	svm can be used as a classification machine, as a regression machine, or for novelty detection. Depending on whether y is a factor or not, the default setting for type is C-classification or eps-regression, respectively, but may be overwritten by setting an explicit value.	'nu-regression' or 'eps-regression'
	degree	parameter needed for kernel of type polynomial (default: 3)	2-3
	cost	The cost of predicting a sample within or on the wrong side of the margin.	0 -10
	gamma	parameter needed for all kernels except linear (default: 1/(data dimension))	1
	coef0	parameter needed for kernels of type polynomial and sigmoid (default: 0)	0
	tolerance	tolerance of termination criterion (default: 0.001)	0.001

RT: regression tree; RF: random forest; SVM: support vector machine.

3.4.7 Similarity of soil environmental conditions between reference area and the blocks of Urucu, Araracanga and Juruá

According to Wu (2004) the scale effects on spatial pattern analysis may occur in each of the following three situations: a) changing grain size or resolution only; b) changing extent only, and; c) changing both grain and extent. In the context of this study case, the main challenge is the transfer or application of a SOCS prediction model developed using a relatively few number of SOCS data, from a restricted area (RA), to a wider one (situation b, changing extent). The adequacy of applying SOCS models developed using different approaches (RA e

TA datasets) relies on the similarity of the environmental conditions. To examine the constraining effect of soil environmental conditions on the transfer of the SOCS models developed, the similarities among the field in RA and the three blocks (Urucu, Araracanga e Juruá) were characterized by the relief factor (R) of the SCORPAN model (MCBRATNEY et al., 2003). The Gower similarity coefficient proposed by Gower (1971) as outlined by Mallavan et al. (2010), was employed to measure the similarity among fields Eq. (4):

$$Sij = \frac{1}{p}\sum_{k=1}^{p}\left(1 - \frac{|x_{ik} - x_{jk}|}{range\ k}\right) \quad (4)$$

where Sij is the Gower similarity coefficient between sites i and j; k represents the relief factor variables (R); p is the number of variables; range k is the value range of variable k in the whole study area. Thus, Sij ranges between 0 and 1; a value of 1 means that the two individuals differ in no character whereas 0 means they differ maximally in all their characters. In literature, the Gower index of similarity (Sij) is generally used in its inverted form (1-Sij), or the Gower index of dissimilarity. In this case, the interpretation is the opposite of the one presented above, that is, values of 1-Sij equal to 0 means that the two individuals differ in no character whereas 1 means they differ maximally in all their characters. In this work, the Gower dissimilarity index (1-Sij) was used. Important variables that were included in the dissimilarity analysis are shown in Table 3.

3.4.8 Exploratory analysis and covariates selection

One of the reasons for employing ML techniques is that they are not only capable of capturing complex non-linear interactions between variables but also because there is no need to follow any statistical assumptions. Besides, the algorithms can also handle a large number of cross-correlated covariates (collinearity) as a predictor. However, commonly the modeling process using ML is quite automated, resulting in that the exact path between model input and output is unknown (Black Box). The poor knowledge may result in a disconnected model to the reality, not resemble the process described by the existing knowledge (WADOUX & MCBRATNEY, 2021).

According to Wadoux et al. (2020) only one third of the works that use ML adopt a covariate selection criterion as a data pre-processing step before making them available for the algorithms to train a model. In this study, two ways of developing the models were tested, called wrapper method" and "previous covariate selection". In the first case (wrapper method), all the covariates were made available for the algorithms to develop the training, that is, a more automatic process. In the second case (previous covariate selection), considering that the spatial distribution of the SOCS in the region has a hypothesis of a soil-relief-vegetation (SRV) relationship, proposed by Ceddia et al. (2015), the selection of input covariates followed three steps of pre-processing, prior to calibrating the ML models: a) exploratory analysis for evaluating anomalous data that could wrongly influence the results; b) evaluation of the Pearson correlation between SOCS and relief covariates to better understand the data and the pedological and environmental relationship and; c) multicollinearity assessment. Covariates highly correlated with each other were discarded. This evaluation was important because collinearity and multicollinearity can dilute the covariates importance in the model.

For the assessment of multicollinearity, Garson (2005) suggests the use of the Variance Inflation Factor (VIF), which assesses the increase in variance due to the presence of multicollinearity Eq. (5).

$$VIF = \frac{1}{1-R^2} \quad (5)$$

According to Gujarati (2000), the limit value of the VIF to establish whether a variable is not collinear is ≤ 4, and if this value is greater than 10, the variable is highly collinear. Therefore, multicollinearity was minimized by removing variables with variance inflation factors (VIF) > 10.

3.4.9 Assessing model's accuracy

The accuracy of the models was assessed using the following parameters: coefficient of determination R^2, mean absolute error (MAE) and root mean square error (RMSE), Eq. (6) and (7), respectively. MAE and RMSE are among the best general measures of performance of the model, since they summarize the average difference in the units of observed and predicted values. Compared to MAE, RMSE is more sensitive to extreme values (WILLMOTT, 1982).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |P_i - O_i| \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (O_i - P_i)^2} \quad (7)$$

Where: Oi and Pi are the observed and predicted values, respectively.

3.5 RESULTS AND DISCUSSIONS

3.5.1 Descriptive statistics

Table 5 presents the statistics of SOCS data at depths 30 and 100 cm for the two dataset approaches used (Dataset 1- RA and dataset 2 - TA). For purposes of organization and comparison, the data coding presents some differences according to the approach used. For example, in the first part of the table, the codes represent the terms associated with the RA approach (Dataset-1). In this case, the data are presented as follows: whole 120 data (W), the 96 data from the RA (T-training) and 24 soil profiles that were used as external validation, which were collected in the remote clearings (V). Also in this first part of Table 5, the statistics of the 24 validation data are presented separately, following the division of the study area into blocks as shown in Figure 9A, namely: Urucu Block (VU - 11 observations), Araracanga Block (VA - 10 observations) and the Juruá Block (VJ - 3 observations). The statistics of the 21 observations from the Urucu and Araracanga blocks (VUA) and the 14 observations from the Urucu and Juruá blocks (VUJ) are also presented. With these divisions, we seek to better understand the variation of SOCS data in regions located at different distances from the RA and with different patterns of relief covariates. This division also helps in understanding the performance of ML models used in SOCS prediction. Finally, the second part of Table 5 presents the data statistics when the TA approach is adopted (dataset 2). In this case, in addition to the dataset with whole 120 observations (W), the statistics of 90 training data (T - 75%) and 30 validation data (V-25%) are shown, which were randomly selected Figure 9B.

Table 5. Descriptive statistics of target soil variables.

Variables	Set	n	Min	Max	Mean	Median	SD	Sk	k	CV%
Reference area (dataset 1)										
SOCS30 (kg C.m^{-2})	**W**	**120**	1.29	6.57	3.35	3.21	1.00	0.62	0.37	29.85
	T	**96**	1.67	5.74	3.22	3.03	0.83	0.48	-0.001	25.78
	V	**24**	1.29	6.57	3.87	3.83	1.42	0.00	-0.89	36.69
	VU	**11**	2.45	6.57	4.24	4.41	1.18	0.33	-0.95	27.83
	VA	**10**	1.29	6.22	3.75	3.78	1.71	-0.13	-1.53	45.60
	VUA	**21**	1.29	6.57	4.01	3.85	1.44	-0.17	-0.82	35.91
	VUJ	**14**	2.37	6.57	3.95	3.83	1.23	0.43	-0.83	31.14
	VJ	**3**	2.37	3.84	2.88	2.44	0.82	0.38	-2.33	28.47
SOCS100 (kg C.m^{-2})	**W**	**120**	2.59	15.36	7.82	7.68	2.55	0.53	0.14	32.61
	T	**96**	3.26	11.93	7.35	7.48	2.00	0.07	-0.46	27.21
	V	**24**	2.59	15.36	9.68	9.71	3.53	-0.20	-1.14	36.47
	VU	**11**	4.43	15.36	10.59	10.60	3.30	-0.24	-1.18	31.16
	VUA	**21**	2.59	15.36	9.91	9.94	3.54	-0.32	-0.98	35.72
	VA	**10**	2.59	14.59	9.16	9.24	3.82	-0.23	-1.38	41.70
	VUJ	**14**	4.43	15.36	10.05	10.04	3.40	-0.08	-1.43	33.83
	VJ	**3**	5.71	12.30	8.08	6.23	3.66	0.37	-2.33	45.30

To be continued...

Continuation of **Table 5.**

Total area (dataset 2)										
Variables	**Set**	**n**	**Min**	**Max**	**Mean**	**Median**	**SD**	**Sk**	**k**	
SOCS30 (kg C.m^{-2})	**W**	**120**	1.29	6.57	3.35	3.21	1.00	0.62	0.37	29.85
	T	**90**	1.29	6.57	3.33	3.25	0.96	0.62	0.87	28.83
	V	**30**	1.73	5.74	3.40	3.01	1.14	0.57	-0.79	33.53
SOCS100 (kg C.m^{-2})	**W**	**120**	2.59	15.36	7.82	7.68	2.55	0.53	0.14	32.61
	T	**90**	2.59	15.36	7.77	7.68	2.54	0.53	0.16	32.69
	V	**30**	3.26	14.59	7.96	7.96	2.59	0.47	-0.12	32.54

W: Whole dataset; T: Training dataset; V: Validation dataset; VU: Urucu block validation data set; VA: Araracanga block validation data set; VJ: Jurua block validation data set; VUA: Urucu/Araracanga block validation data set; n: number of observations; Min: minimum; Max: maximum; SD: standard deviation; Sk: Skwewness; K: Kurtosis; SOCS30: carbon stock at 0-30 cm; SOCS100: carbon stock at 0-100 cm.

Comparing the statistic of 120 data (W) with that of 96 training data restricted to RA (T), a greater range of variation is found, showing that the 24 new data obtained in the more distant remote clearings present not only lower values minimum, but also higher maximum SOCS values at depths of 30 and 100 cm. It is also noted that the maximum values occur in the Urucu block while the minimum values are found in Araracanga. In other words, when combining the data obtained in the Urucu and Araracanga blocks (VUA), the amplitude is equal to that of the total dataset (W). In the remote areas (V), the smallest SOCS found at a depth of 30 cm (1.29 Kg Cm-2) is 30% smaller than the lowest value observed in the reference area (T - 1.67 Kg C. m-2). The highest value found in remote clearings reached 6.57 Kg C.m-2, being approximately 15% higher than that found in the RA (5.74 Kg C. m-2). On average, the SOCS30 in the remote clearings used for external validation (3.87 Kg C. m-2) was approximately 20% higher than that found in the RA (3.22 Kg C. m-2). A similar pattern was also observed with SOCS100. In this case, the lowest carbon stock value in remote clearings (2.59 Kg C.m-2) is 26% lower than that observed in the RA (3.26 Kg C. m-2). On the other hand, the highest value found in remote clearings reached 15.36 Kg C.m-2, approximately 30% higher than that observed in the RA (11.93 Kg C. m-2). These differences result that, on average, the soils of the remote clearings have a SOCS100 32% greater than that surveyed in the RA at 100 cm depth.

Although there are differences between the data observed in RA and in remote clearings, the fact is that the SOCS found are in agreement with previous studies carried out in the Amazon region. Moraes et al. (1995), working with data from the RADAMBRASIL project, found mean SOCS100 values equivalent to 10.3 kg C.m-2, for the entire Legal Amazon. The studies by Batjes and Dijkshoorn, (1999), also covering the entire Legal Amazon, found mean SOCS30 and SOCS100 values of 5.05 kg C.m-2 and 9.8 kg C.m-2, respectively.

3.5.2 The correlation between SOCS and covariates

The correlation matrices between covariates and the values of SOCS30 and SOCS100, using both RA and TA datasets are shown in Figure 11. It is noted that the correlations are low (less than 0.4), for easy viewing, only the covariates with a correlation greater than 0.10 are highlighted in Figure 12.

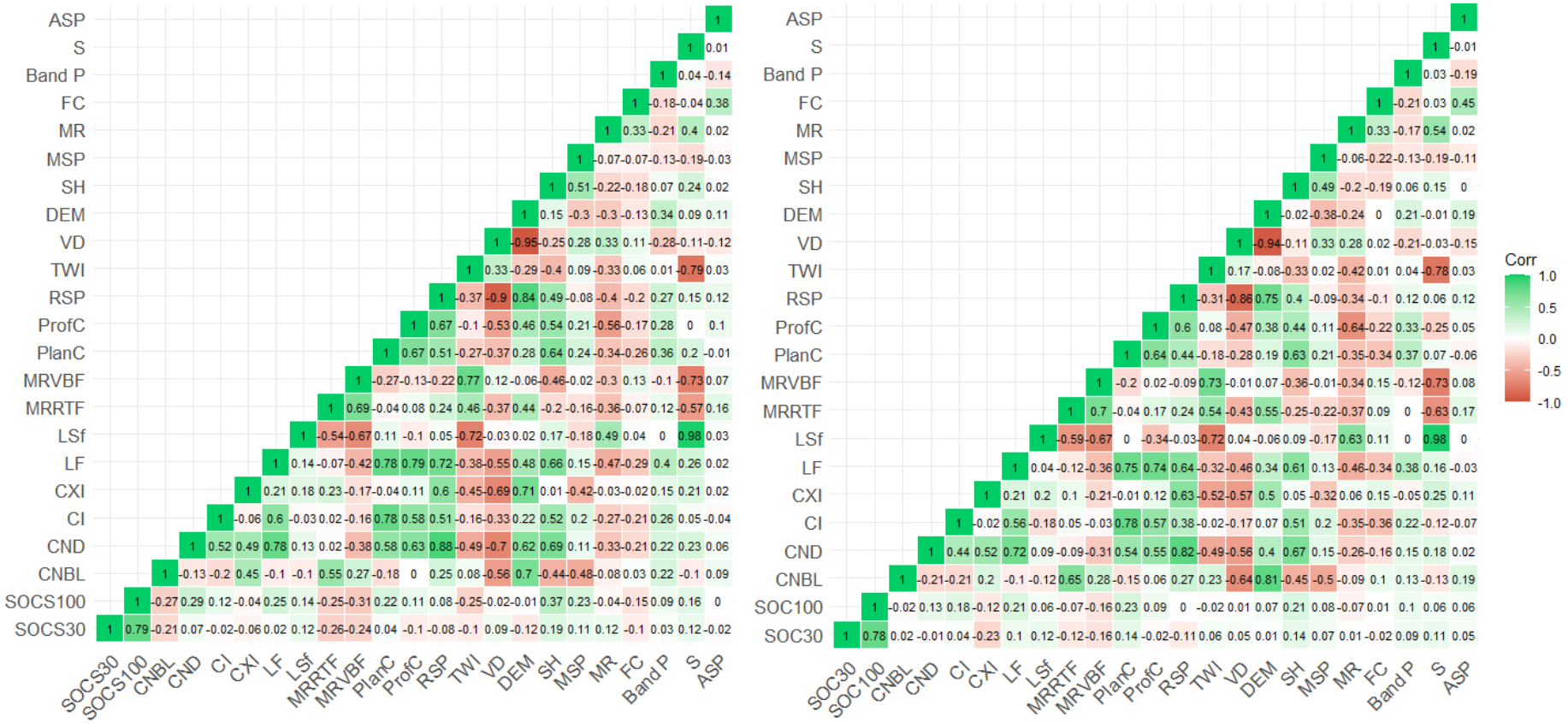


Figure 11. Correlation matrix of environmental covariates. (A) Covariables correlated with carbon stock data at 30cm and 100cm at the reference area. (B) Covariables correlated with carbon stock data at 30cm and 100 cm at the total area. (image generated in RStudio program).

In the data set referring to RA, the covariates that have a higher correlation coefficient directly proportionally with SOCS30 presented the following order of importance (Figure 12A): SH > ASP = MR = LSf = S > MSP. On the other hand, other covariates have an inversely proportional relationship with SOCS30, such as MRRTF > MRVBF > CNBL > DEM. Analyzing the SOCS100 (Figure 12B), there is a greater number of covariates with a correlation coefficient ≥ |0.10|. In this case, in addition to the covariates SH, LSf, S and MSP, it is also directly proportionally to the SOCS100, the covariates LF, PlanC and ProfC. It is also observed that the following relief covariates, according to a decreasing order of correlation, MRVBF > CNBL > MRRTF > TWI > FC, present an association inversely proportional to SOCS100.

In the TA Figure 12 (C) and (D), it was observed that correlation coefficients with SOCS are smaller (at 30 and 100 cm) than when compared to the RA approach (r less than |0.25|). In this case, at a depth of 30 cm, the covariates PlanC = SH > LSf > S > LF and CXI > MRVBF > MRRTF > RSP were highlighted. The first covariates being directly proportional and the last covariates are inversely proportional. When considering the SOCS100, all covariates reduce the correlation coefficient (less than |0.25|). The covariates that present a correlation coefficient directly proportional to the stocks are, in order of importance: PlanC> LF= SH > CI> CND> Band P and as for the SOCS30 depth the covariates with high correlation coefficient (inversely proportional) were MRVBF > CXI.

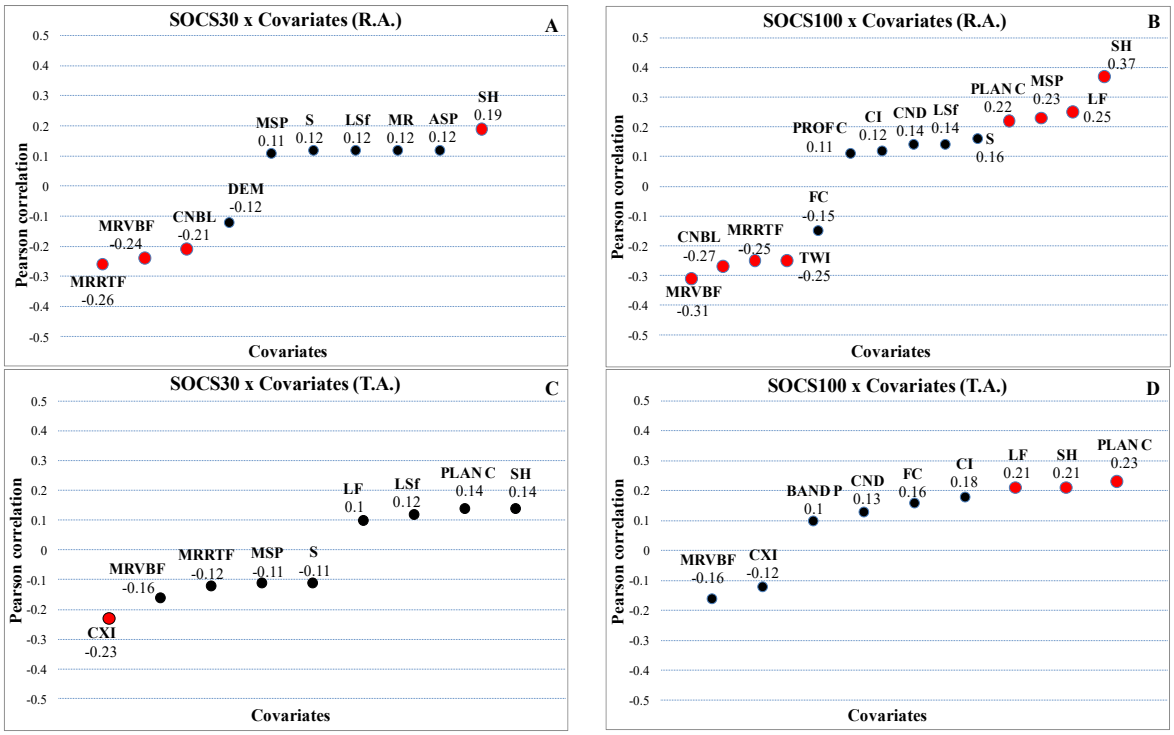


Figure 12. Covariates with higher correlation (r≥ |0.10|) with SOCS at 30 and 100 cm. Red dots are those covariates that the correlation coefficient value is very close to or greater than 0.20.

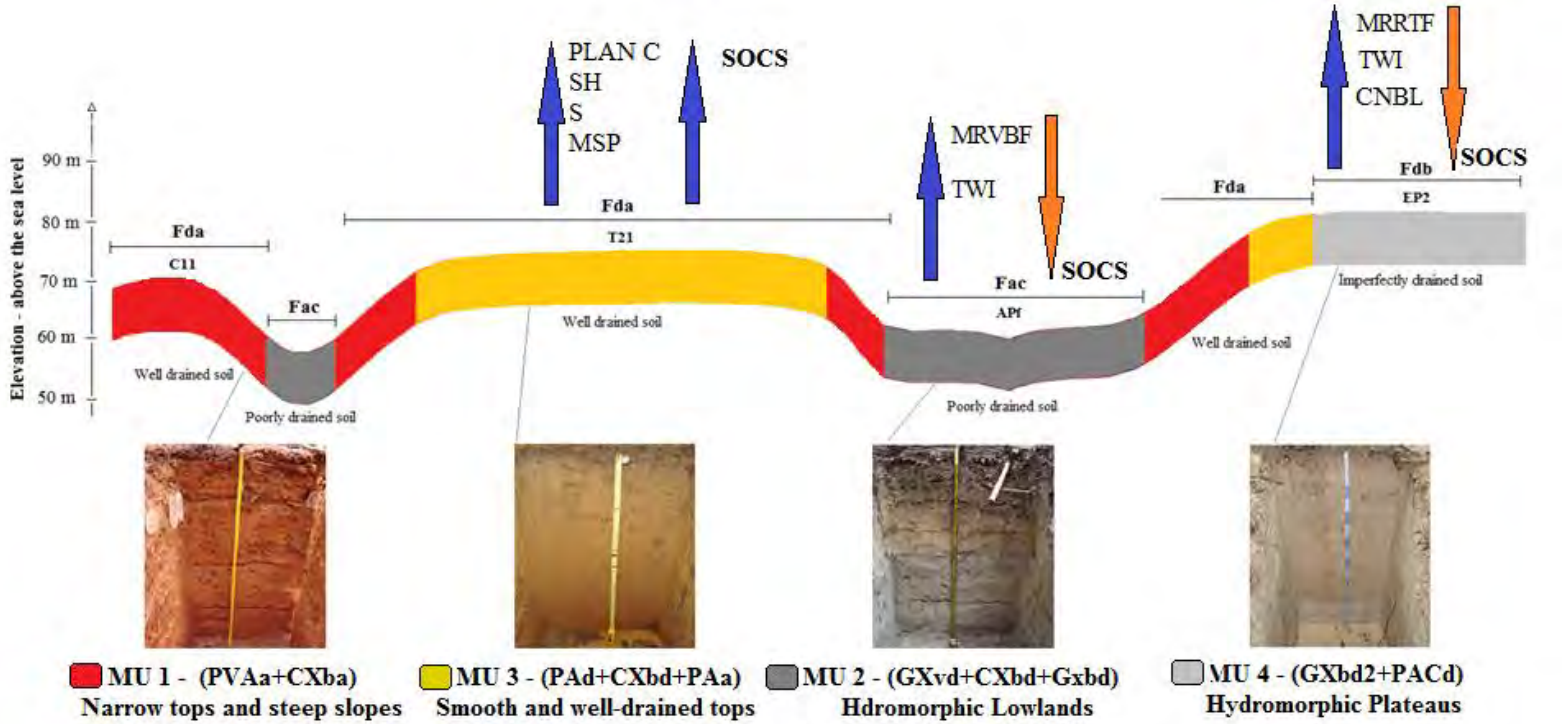
The change of the relationship between the SOCS data and the covariates according to the dataset used (changing from RA to TA), is associated with the introduction of more dispersed data of both SOCS and the covariates. In both datasets, the number of data used for training is almost the same (96 for RA and 90 for TA). The main change observed is in the amplitude values of the TA dataset, which is higher than the observed in RA dataset (Table 5). Since interpretations of the relationships between relief covariate values and carbon stock change with changing datasets, it may be questioned which of these interpretations are more

realistic. In addition, the data belonging to dataset 2 (TA), considering a greater amplitude in the training data, could be closer to what actually occurs in reality. This issue can be contextualized by highlighting the importance of exercising the challenge of introducing into DSM studies, especially using ML techniques, the existing knowledge about soils and their attributes in a given area.

As argued by Wadoux et al. (2020) the ML model predicting a number based on relationships between covariates that are unknown in the view of existing knowledge, should not be taken with the same seriousness as a number predicted by mechanistic steps or an established theory based on the current hypothesis or prior knowledge on the soil spatial variation for an area. In short, the question is whether the correlations between the covariates of relief and SOCS follow an existing knowledge along the study case region. In general, both the correlation coefficient values found and the covariates that are most associated with SOCS30 and SOCS100 coincide with those observed by Ceddia et al. (2015) and Ceddia et al. (2017), who studied the relationship between carbon stock and relief covariates in the RA. Also, in Figures 12 (A) and (B), those covariates whose correlation coefficient value is very close to or greater than 0.20 are highlighted in red dots. These same covariates were contextualized in the Soil-Relief-Vegetation (SRV) model of Figure 13, which represents an adaptation to that presented by Ceddia et al. (2015).

In Figure 13 the four SRV classes are organized showing the environmental condition that governs not only the soil types but also its attributes, like SOCS. The regions classified as SRV1 and SRV4 have in common the hydromorphic conditions of the soils, once due to the relief forms the water is preferentially stagnated. In both cases (SRV1 and SRV4) the main vegetation types covering the soils are Fac and Fdb, which are composed of species more adapted to soil aeration restriction, mainly at the subsurface. Consequently, the carbon stocks are lower, since both the input of carbon from leaves, tree trunks and roots are lower and concentrated at surface layers. In the lowlands (SRV1), near rivers and watercourses (APf - River plains), the highest density of palm trees are common and the relief's attributes MRVBF and TWI are highlighted once they represent regions with water stagnation in bottom valleys, which in turn is associated with lower carbon stocks (wide and narrow lowland areas with dark grey color, Figure 13). In the SRV4 unit (with the combination of relief forms EP2 - Biplained surfaces/flatlands, and vegetation type Fdb -Upland Open Tropical Rainforest) the covariates of the relief MRRTF, TWI and CNBL stand out, which also show a negative correlation with the carbon stock (top right of Figure 13).

The regions classified as SRV2 and SRV3 are associated with terrain types C11 (Dried out areas on flat-topped) and T21 (Tabular Interfluves), respectively. In both cases, the associated vegetation type is classified as Fda (Upland Dense Tropical Rainforest), which has a higher plant density, and also larger height and trunk diameter. In this case, the input of carbon is higher not only for the leaves and tree trunks but also due to roots growing into deeper layers of the soil. The soils in these environments, due to the better drainage condition, have reddish (SRV2) and yellowish (SRV3) colors. The relief covariates SH, S, MSP and PlanC are associated with these environments and show a positive correlation with the carbon stock.



Soil - Relief - Vegetation Unity (SRV)	Soil Mapping Unity	Vegetation class	Relief Forms	Carbon Stock
SRV1 (MU2+Fac+HS+Apf)	MU 2	Fac	APf	Lower CS and concentrated in the surface layers. Carbon source: leaves and tree trunks, and surface roots.
SRV2 (MU1+Fda+PS+C11)	MU 1	Fda	C11	Higher CS throughout the soil profile. Carbon source: leaves and tree trunks, and roots distributed from the surface to the deeper layers
SRV3 (MU3+Fda+PS+T21)	MU 3	Fda	T21	Higher CS throughout the soil profile. Carbon source: leaves and tree trunks, and roots distributed from the surface to the deeper layers
SRV4 (MU4+Fdb+PS+EP2)	MU 4	Fdb	EP2	Lower CS and concentrated in the surface layers. Carbon source: leaves and trunks of treetops and surface roots.

Figure 13. A schematic representation of the relationship between SOCS and Covariates along the RA. Fac- Flooded Lowland Open Tropical Rainforest; Fda- Upland Dense Tropical Rainforest; Fdb- Upland Open Tropical Rainforest; APf- River plains; C11- Dried out areas on flat-topped; T21- Tabular Interfluves; EP2 - Bi-plained superficies- flatlands. CS-Carbon Stocks. H.S.—Holocene Sediments; P.S.—Pleistocene Sediments. (Source: modified CEDDIA et al., 2015).

3.5.3 Similarity between RA and the Urucu, Araracanga and Juruá blocks

Comparing the descriptive statistics of the training area (RA) with the other regions to be mapped (Urucu, Araracanga and Juruá blocks) we can observe that some covariates in the Jurua region have very different minimum, maximum, average and median values of RA, which may significantly differentiate one landscape from another (Table 6). The covariates CNBL, CND, MRRTF and MRVBF stand out, which are more different in the Juruá region in relation to RA.

Table 6. Descriptive statistics of the covariates in the study area by blocks.

	RA					U				
Covariates	**Mean**	**Median**	**SD**	**Min**	**Max**	**Mean**	**Median**	**SD**	**Min**	**Max**
DEM	67.78	68.34	6.89	46.27	83.67	68.53	69.88	7.55	23.03	64.69
CI	0.03	0.59	16.80	-94.51	96.07	-0.0002	0.54	16.41	-98,08	98.91
TWI	7.66	7.56	1.06	4.61	12.30	8.07	7.98	1.23	4.332	12.54
RSP	0.48	0.51	0.30	0	1	0.44	0.45	0.30	0	1
CND	6.40	6.15	4.01	0	25.39	5.41	4.88	3.95	0	29.64
CNBL	61.72	61.16	5.95	46.56	79.59	63.47	64.07	7.16	23.03	83.16
LSf	0.63	0.46	0.60	0	10.46	0.51	0.30	0.59	0	95.76
MRVBF	5.73	9.38	4.52	0	9.98	6.69	9.82	4.33	0	9.98
MRRFT	2.84	1.97	2.67	0	7.93	4.02	4.76	3.09	0	7.99
CXI	51.34	52.41	7.63	0.15	69.19	50.29	51.85	8.89	0	73.19
ASP	177.10	175.22	106.81	0	360	173.78	171.04	107.03	0	360
LF	5.32	5.00	2.41	1.00	10.00	5.18	5.00	2.11	1.00	10.00
ProfC	-0	-0	0	-0.009	0.01	-0	0	0	-0.013	0.011
PlanC	0.0	3.40	0.0	-0.007	0.01	0	0	0	-0.010	0.013
VD	7.07	5.92	5.32	-2.18	26.08	7.324	5.708	5.93	-2.180	50.380
SH	4.08	3.55	1.85	1.47	18.94	3.84	3.36	1.79	1.13	25.51
MSP	0.27	0.25	0.17	0.00	0.82	0.25	0.23	0.16	0.00	0.85
S	6.23	5.15	4.87	0.00	48.86	5.16	3.70	4.77	0.00	67.20
MR	0.25	0.16	0.29	0.00	2.49	0.21	0.10	0.27	0.00	2.95
FC	2451	2996	3090	400	81207	2347	1449	2956	400	14170
BandP	0.43	0.43	0.07	0	0.99	0.44	0.44	0.06	0	0.90
	A					**J**				
Covariates	**Mean**	**Median**	**SD**	**Min**	**Max**	**Mean**	**Median**	**SD**	**Min**	**Max**
DEM	69.53	72.22	10.26	33.87	94.79	80.04	82.19	8.82	49.41	112.02
CI	0	0.49	16.45	-98.78	99.01	0.00	0.78	18.10	-99.21	99.40
TWI	7.92	7.72	1.41	4.36	12.37	7.58	7.38	1.28	3.86	12.01
RSP	0.41	0.41	0.31	0	1	0.35	0.32	0.29	0	1
CND	6.01	5.32	4.83	0	33.92	4.45	3.42	4.08	0	40.50
CNBL	63.93	65.28	8.85	34.16	85.97	76.03	77.62	8.40	49.88	95.63
LSf	0.60	0.36	0.67	0	9.39	0.53	0.30	0.67	0	25.63
MRVBF	4.96	4.77	4.13	0	9.96	3.70	3.89	2.82	0	9.65
MRRFT	3.37	2.67	3.15	0	9.73	6.53	9.36	4.19	0	9.98
CXI	48.32	50.92	11.07	0	73.40	39.58	41.13	8.27	0	63.48
ASP	171.04	168.26	109.06	0	360	168.08	166.38	109.74	0	360
LF	5.26	5.00	2.32	1.00	10.00	5.32	5.00	2.03	1.00	10.00
ProfC	-0.0	0.0	0.0	-0.011	0.012	-0.0	-0.0	0	-0.014	0.016
PlanC	0.0	0.0	0	-0.012	0.011	0.0	0.0	0	-0.013	0.018
VD	9.21	7.06	7.33	-9.16	36.62	7.98	6.04	6.55	-0.04	37.43
SH	4.18	3.59	2.11	1.16	27.33	3.62	3.12	1.73	1.14	32
MSP	0.31	0.29	0.20	0	0.88	0.22	0.18	0.16	0	0.89
S	5.81	4.25	5.34	0	50.21	5.39	4.02	5.24	0	76.92
MR	0.24	0.11	0.32	0	3.01	0.18	0.00	0.27	0	4.23
FC	2332	1421	2993	400	13304	1609	1059	1735	400	6948
BandP	0.45	0.45	0.11	0	0.93	0.43	0.43	0.10	0	0.94

RA: Referencea area; U: block Urucu; A: block Araracanga; J: block Jurua; n: number of observations; Referencea area: (n=199.167); Urucu: (n=11.209.198); n Araracanga: (n=93.64993); n Jurua: (n=11.730.902); Min: minimum; Max: maximum; SD: standard deviation.

If we interpret the statistical data of the covariates individually, we reach the conclusion that there are important differences between the Jurua and RA region that apparently can cause variations and limitations in the transferability of the models. However, to get a more comprehensive assessment, the results of the general Gower index are presented (Figure 14).

Considering that one of the approaches of this study is to evaluate the validity of using RA (with 80 km²) to develop prediction models for carbon stock in a much larger area (13,440 km².- 168 times larger), it is essential to evaluate the representativeness of this area in relation to the different study blocks (Urucu, Araracanga and Juruá blocks). The evaluation using covariates used to predict SOCS is a way to assess the similarity between different regions.

Figures 14A, 14B and 14C show the spatial distribution maps of the Gower index of dissimilarity (1-Sij) of the blocks, while Figures 14 (D), (E) and (F) represent the contribution of each covariate in the final value of the general Gower index. Basically, considering the values found (below Figures 14A, 14B and 14C) it can be said that there is little dissimilarity between the RA and the Urucu, Araracanga and Juruá blocks (values of 0.155, 0.164 and 0.171, respectively). The dissimilarity increases departing from the Urucu block towards Juruá (farthest from the reference area). In addition, the areas with the highest dissimilarity index values (1-Sij) are those associated with lowland areas (hydromorphic lowlands- black arrows on maps) and higher regions located at watershed boundary divisors (pixels with more discrete values highlighted with blue arrows on maps).

In Figures 14 (D), (E) and (F), graphs are presented with the general Gower index (red bars) and the same index for each covariate (blue bars). Also, in each of these figures, the covariates that contributed with the greatest dissimilarity (Gower index of each covariate > 0.25) were highlighted in the right corner (shaded region with yellow color). Note that in the three blocks Figures 14 (D), (E) and (F) the covariates that most contributed to differentiate the blocks in relation to RA are: MRVBF, MRRTF and RSP. In the case of the Juruá block (Figure 14D), in addition to these, the covariate CNBL also stands out. As seen in Figures 14A and 14B, the covariates MRVBF, MRRTF and CNBL are those that presented the highest correlation coefficients with the SOCS values in RA. The highest values of the MRRTF and CNBL covariates are more associated with regions with broad tops (hydromorphic tops - SRV4), while the covariate MRVBF is associated with lowland regions (hydromorphic lowlands - SRV1), which also identifies hydromorphic environments (Figure 13). In both cases, the broad tops (SRV4) and the hydromorphic lowlands (SRV1), as noted by Ceddia et al. (2015), are areas where the lowest SOCS values are commonly found, and also where most of the SOCS is concentrated in the first 30 cm of soil depth. Another covariate that stands out as a differentiator between RA and the other blocks is the RSP attribute. In this case, no correlation was found with the SOCS values at the two depths studied (Figure 11). This means that not all relief covariates that are important to characterize the environmental differences between the blocks and the RA will necessarily correlate with the attribute of interest of the prediction model (in this case SOCS).

On the other hand, higher values of the covariates SH, S, MSP and PlanC are associated with the regions with the greatest SOCS in the RA (SRV2 and SRV3, Figure 13). Among these, MSP also stands out in Figures 14D, 14E and 14F, since it presents Gower index values between 0.20 and 0.25 (greater than the global Gower index in the three blocks).

Also, with respect to the results seen in Figure 14, it is possible to emphasize that the Gower index can also be used to support the choice of new or changes in previously available RA. In this study case, as the RA is imposed because it is the only option with the easiest access, there is no sense in this use. However, it is possible to conjecture that if we were to change the RA, this change should be in the sense of including regions that expand the expression of the covariates that most differentiated the blocks in relation to the RA (in this case MRVBF, MRTFF, CNBL and RSP, for example).

Probably, the areas close to the blue arrow in Figure 14C (broad tops with higher values of the covariate MRRTF and CNBL) and also in the lowland regions with higher values of MRVBF (in this case, mainly associated with the Urucu river - areas close to the black arrow in Figure 14C, could be expanded. It is important to highlight that these areas are the most difficult to access and cause the most undersampling in these environments. Comparing the results of SOCS in the RA with the other blocks, it is noted that if the field surveys covered greater density in these new regions, there would be the possibility of finding not only more data for training the models, but would also open up the possibility of finding greater amplitude in the carbon stock results. It could make the training samples within the RA likely more similar to what was found in the remote clearings. It is also noteworthy that no major differences were found in the types of soils observed in the RA (2010 field campaign) and in remote clearings (2018 field campaign).

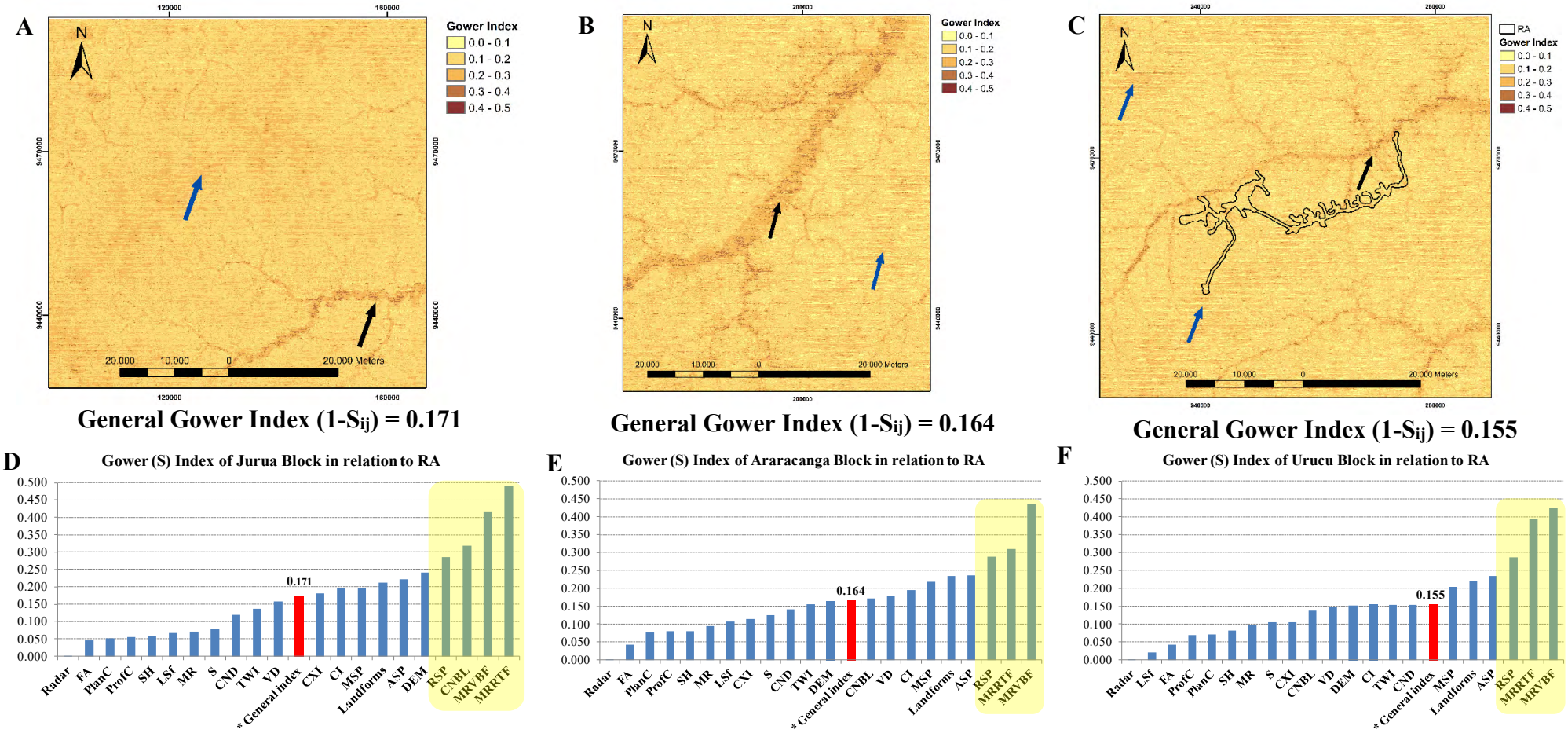


Figure 14. Gower index maps for Juruá, Araracanga and Urucu (A, B and C, respectively). Gower index by covariate and general Gower index for Juruá, Araracanga and Urucu (D, E and F, respectively).

3.5.4 Comparison of predictive models

Considering the three algorithms used, only the tree-based models (RT and RF) provide an indication of the most important covariates for SOCS prediction. The importance of the covariates to predict SOCS at 30 cm and 100cm soil depth is presented in Table 7 (RT) and Figure 15 (RF), respectively. As can be seen, the same covariates considered more important for RT were those for RF, changing only the order in some cases. The decreasing importance of the covariates to predict SOCS at 30 cm soil depth, using RT algorithm, is presented as follow: MRVBF > SH > MRRTF > CNBL > MR for RT (Table 7). On other hand, using RF algorithm, the decreasing order of importance is: SH > MRVBF > MRRTF > MR > CNBL Figure 15A. With the exception of the MR covariate, all of them was already highlighted in Figure 11A, due to the higher correlation coefficient with SOCS30. In SOCS30 the most important covariates were SH and MRVBF Table 7 and Figure 15A.

Table 7. Importance of the covariates in RT models for SOCS30 and SOCS100.

Covariate	SOCS30 (%)	Covariate	SOCS100 (%)
MRVBF	34	MRVBF	22
SH	20	CND	17
MRRTF	18	MRRTF	13
CNBL	16	SH	12
MR	11	CNBL	11
-	-	LF	9
-	-	MSP	8
-	-	PlanC	7
Total	100	Total	100

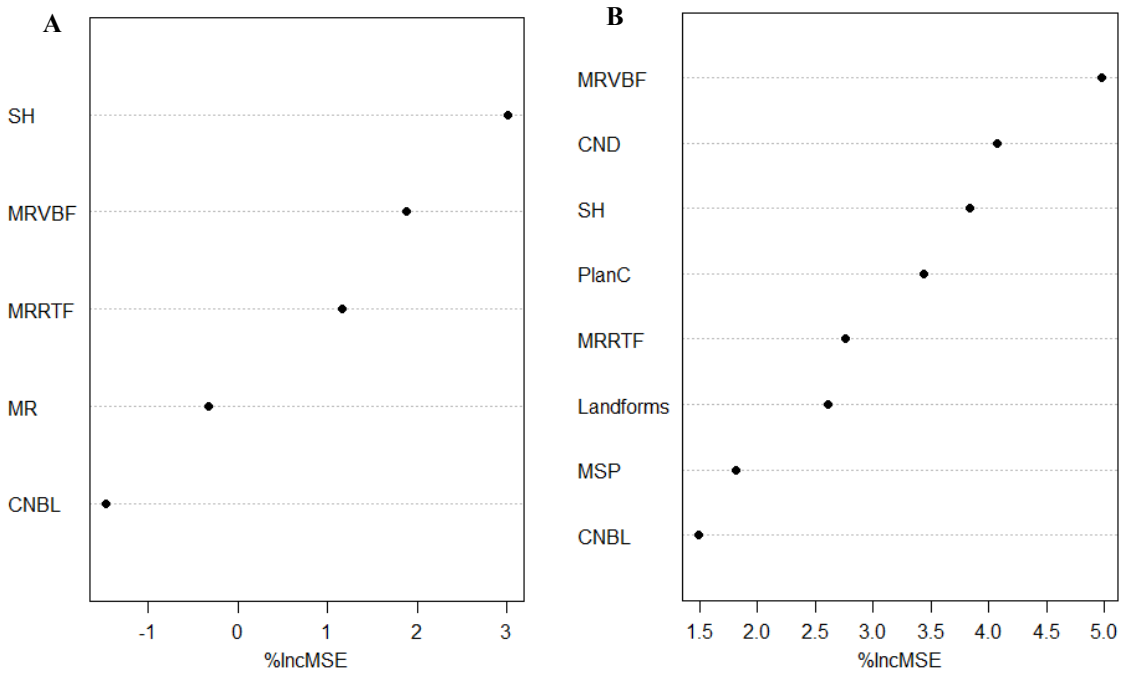


Figure 15. Importance of the covariates in RF model for (a) SOCS30 and (b) SOCS100. (image generated in RStudio program).

Analyzing the results for SOCS100, in addition to the MRVBF index, CND stood out as another important covariate (second place for both RT and RF). The CND map which is the additional calculation of the vertical distance to the stream network that can serve as the simplest indicator for areas subject to flooding. According to Ceddia et al. (2015) in the RA, the further away from the drainage channel we are, the areas will be better drained and will have higher clay content and more carbon stock at 100 cm. Also, with regard to the CND covariate, it is highlighted, through Figures 13 (A) and (B), its high positive correlation with the RSP attribute, considering both the dataset approaches, RA (r=0.88) and the TA (r=0.82), respectively. The RSP covariate was decisive in the greater differentiation of the Gower index of the blocks in relation to the RA (Figure 14D, 14E and 14F, respectively).

The differences found in the choice of covariates when comparing the RT and RF algorithm can be explained by the fact that when using RF, not only one tree is used, but a lot of regression trees. By creating several small regression trees that alone underperform, these together can outperform a single large tree (RT). Unfortunately, when using the SVM algorithm, no information is provided about the covariates used, nor their relative importance. This limitation makes the use of this algorithm more limiting because the soil and attributes mapper does not have the means to advance in explaining a possible connection between what is being used as a covariate and its meaning in the pedological knowledge existing in the study area.

The performance of ML models is presented in Tables 8 and 9. In both tables, it is possible to assess how the algorithms are affected by the type of data set used (RA-Table 8 and TA-Table 9), as well as the effect of whether or not to select the covariates to be made available for the algorithms (PCS - Previous Covariate Selection and WM - Wrapper Method). We observed that, in general, the further the validation samples distances from the Urucu block, the R^2 values decrease, but the result of VUJ is better than VUA (Table 8). Furthermore, analyzing the total validation dataset (V) the result is worse than analyzing VU and VUJ separately, mainly because the data from Araracanga are not well predicted. When the algorithms receive a lower group of covariates, which were previously selected (PCS), the best result for SOCS30 were obtained using RT (R^2=0.32) and RF (R^2=0.27). On the other hand, using all covariates available (WM), the performance of the three algorithms decreases for all validation datasets.

The performance of the algorithms was better when predicting SOCS100. Again, similar results were found when using a dataset with covariates that were previously selected (Table 8). The best result was obtained using RF algorithm (R^2=0.70 for VU and R^2=0.51 for VUJ).

When analyzing the TA approach (Table 9), it is possible to observe that despite the low performance of the models observed in the separate data for validation (V30), they are slightly superior to validation in the RA approach (V, Table 8). It was also possible to observe that the effect of the covariates selection was less effective in this approach. The models with the best performance were SVM for SOCS30 (R^2=0.04 and R^2=0.07, using PCS and WM respectively) and RF for SOCS100 (R^2=0.13 and R^2=0.22 using PCS and WM respectively). From the results presented in Tables 8 and 9, it can be said that the RA approach, using the previously selected covariates based on their association with the SOCS, provided the best results. In addition, the models are more parsimonious and pedological consistent, as suggested by Wadoux et al. (2020) and Wadoux & McBratney (2021).

Table 8. The metric errors of ML algorithms using Reference Area (RA) dataset.

SOCS	Data	RT			RF			SVM		
		R²	RMSE	MAE	R²	RMSE	MAE	R²	RMSE	MAE
SOCS30 PCS	T	0.37	0.66	0.52	0.83	0.43	0.35	0.31	0.69	0.58
	VU	0.32	1.66	1.33	0.27	1.53	1.17	0.20	1.55	1.19
	VUA	0.00	1.68	1.36	0.07	1.62	1.31	0.00	8.73	4.82
	V	0.00	1.59	1.28	0.04	1.54	1.25	0.00	9.77	5.78
	VUJ	0.08	1.50	1.20	0.08	1.43	1.13	0.01	1.84	1.42
SOCS30 *WM*	T	0.40	0.63	0.50	0.89	0.39	0.30	0.51	0.61	0.47
	VU	0.01	1.72	1.30	0.00	1.65	1.25	0.16	1.72	1.36
	VUA	0.04	1.78	1.42	0.00	1.67	1.33	0.00	3.87	8.46
	V	0.02	1.68	1.31	0.00	1.59	1.26	0.00	3.62	7.40
	VUJ	0.03	1.55	1.13	0.02	1.51	1.14	0.07	1.72	1.38
SOCS100 PCS	T	0.41	1.53	1.16	0.90	0.84	0.64	0.35	1.62	1.32
	VU	0.40	4.46	3.78	0.70	4.40	3.81	0.34	4.47	3.76
	VUA	0.00	4.26	3.64	0.05	4.14	3.54	0.00	1.45	8.01
	V	0.00	4.09	3.49	0.03	4.00	3.44	0.00	1.55	9.13
	VUJ	0.15	4.13	3.48	0.52	4.26	3.51	0.18	4.37	3.56
SOCS100 *WM*	T	0.42	1.51	1.13	0.91	0.82	0.62	0.22	1.77	1.39
	VU	0.33	4.45	3.71	0.45	4.71	4.03	0.21	5.13	4.41
	VUA	0.00	5.62	4.31	0.01	5.69	4.51	0.00	1.25	3.15
	V	0.02	3.35	3.35	0.00	4.16	3.55	0.00	1.17	2.93
	VUJ	0.14	4.12	3.44	0.19	4.52	3.76	0.06	5.01	4.17

PCS - Previous Covariate Selection; WM - Wrapper Method; T: Training dataset; V: validation data set; VU: Urucu block validation dataset; VUA: Urucu/Araracanga block validation dataset; V: Urucu/Araracanga/Jurua block validation dataset; VUJ: Urucu/Jurua block validation dataset.

Table 9. The metric errors of ML algorithms using Total Area (TA) dataset.

SOCS	Data	RT			RF			SVM		
		R²	RMSE	MAE	R²	RMSE	MAE	R²	RMSE	MAE
SOCS30	T90	0.36	0.76	0.59	0.90	0.43	0.33	0.69	0.52	0.40
PCS	V30	0.02	1.35	1.08	0.00	1.21	1.02	0.04	2.37	1.83
SOCS30	T90	0.36	0.76	0.59	0.93	0.41	0.31	0.98	0.09	0.01
WM	V30	0.01	1.35	1.08	0.00	1.13	0.93	0.07	5.24	2.49
SOCS100	T90	0.33	2.07	1.65	0.89	1.10	0.84	0.24	2.24	1.71
PCS	V30	0.10	2.50	2.02	0.13	2.40	1.89	0.02	2.91	2.25
SOCS100	T90	0.34	2.05	1.64	0.93	1.02	0.75	0.48	1.93	1.48
WM	V30	0.07	2.55	2.12	0.22	2.35	1.93	0.01	3.09	2.50

PCS - Previous Covariate Selection; WM - Wrapper Method; T: Training dataset; V: validation data set; RT: regression tree; RF: random forest; SVM: support vector machine.

The scatterplots of observed vs. estimated SOCS using the RF, RT and SVM algorithms are shown in Figures 16, 17 and 18, respectively. The prediction SOCS30 is less accurate and is better performed using RT algorithm and only for the Urucu block (VU), explaining 32% of the SOCS variability. On the other hand, when evaluating SOCS100, it appears that there is a better pattern relating pedological and landscape characteristics to soil carbon dynamics. At depth of 100 cm, the RF algorithm was superior to the others, allowing more accurate predictions not only for the Urucu block but also for the Juruá block (explaining 70 and 51% of the SOCS, respectively).

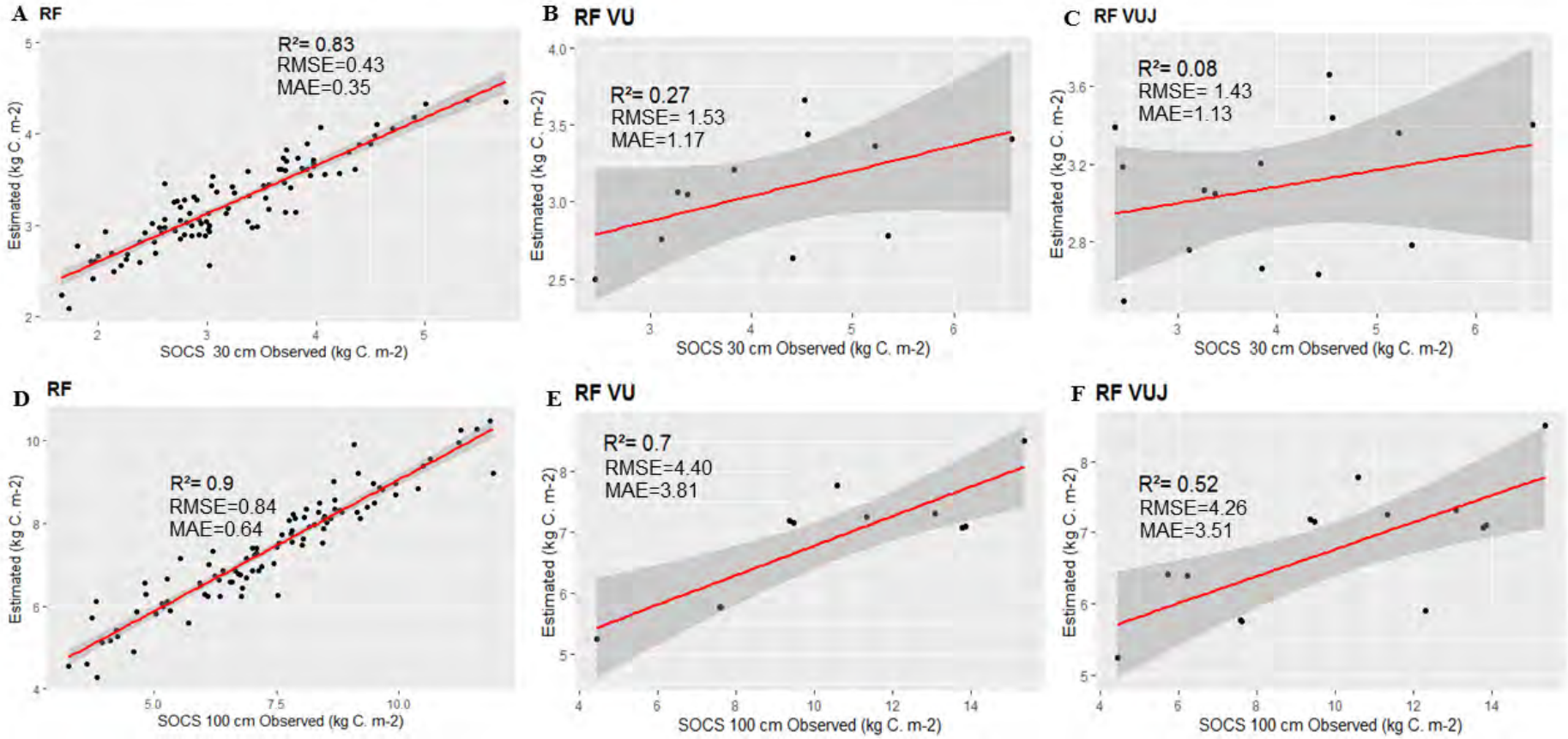


Figure 16. Error metrics of the RF model for training data and validations of carbon stock at 30 and 100 cm in RA. (A) Error metrics of the RF model training data carbon 30 cm; (B) Error metrics of the RF model for Urucu validation data carbon 30 cm; (C) Error metrics of the RF model for Urucu Juruá validation data carbon 30 cm; (D) Error metrics of the RF model training data carbon 100 cm; (E) Error metrics of the RF model for Urucu validation data carbon 100 cm; (F) Error metrics of the RF model for Urucu Juruá validation data carbon 100 cm. (images generated in RStudio program).

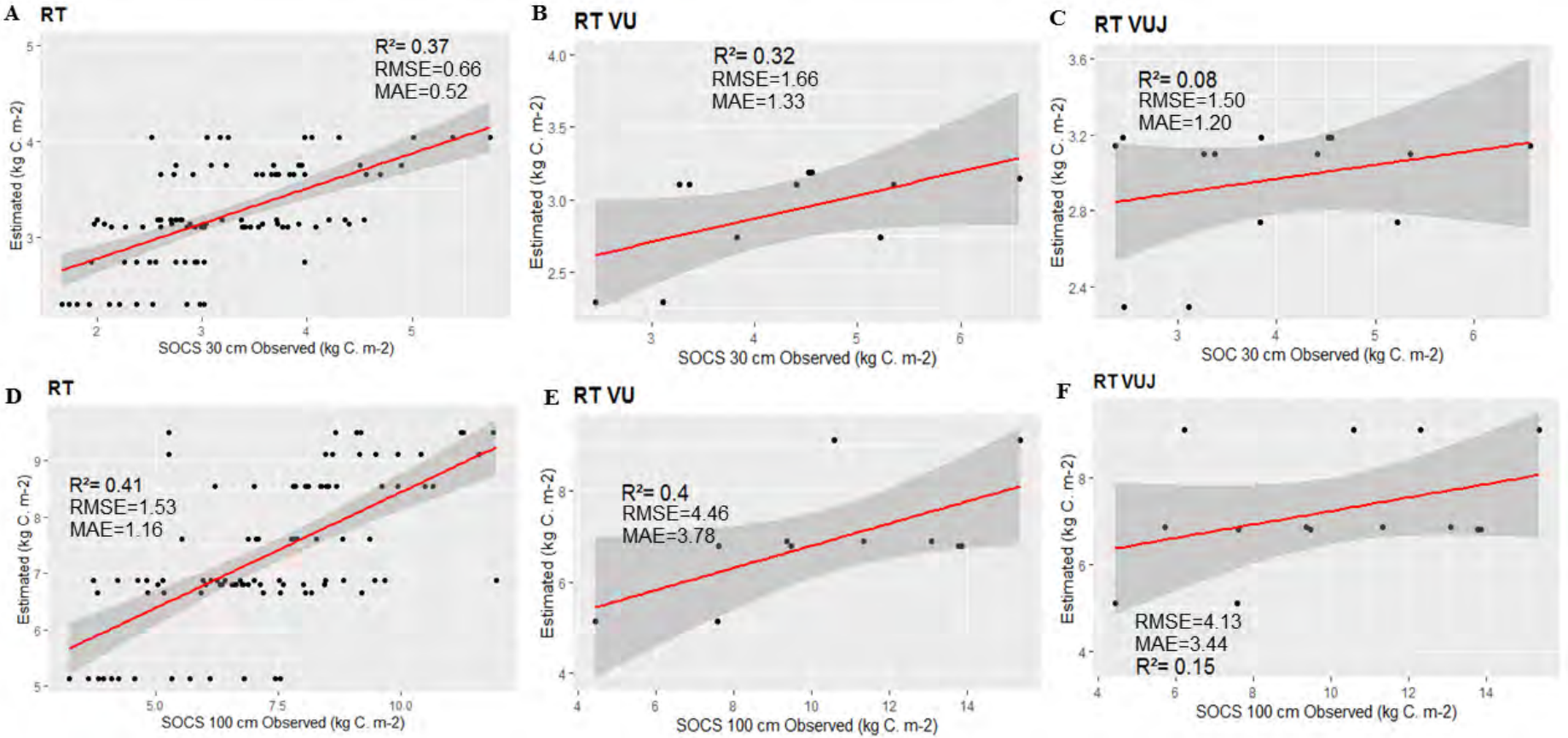


Figure 17. Error metrics of the RT model for training data and validations of carbon stock at 30 and 100 cm in RA. (A) Error metrics of the RT model training data carbon 30 cm; (B) Error metrics of the RT model for Urucu validation data carbon 30 cm; (C) Error metrics of the RT model for Urucu Juruá validation data carbon 30 cm; (D) Error metrics of the RT model training data carbon 100 cm; (E) Error metrics of the RT model for Urucu validation data carbon 100 cm; (F) Error metrics of the RT model for Urucu Juruá validation data carbon 100 cm. (images generated in RStudio program).

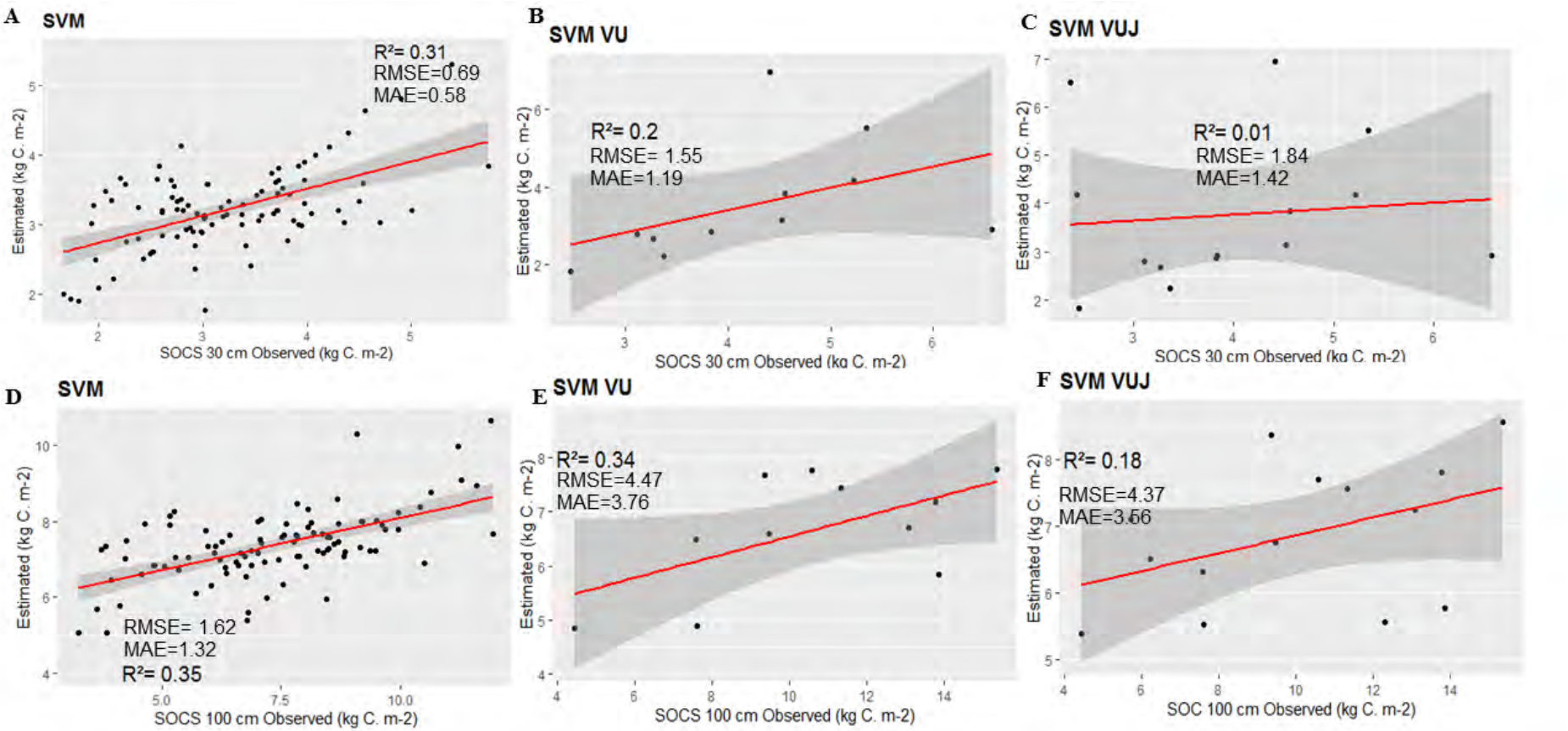


Figure 18. Error metrics of the SVM model for training data and validations of carbon stock at 30 and 100 cm in RA. (A) Error metrics of the SVM model training data carbon 30 cm; (B) Error metrics of the SVM model for Urucu validation data carbon 30 cm; (C) Error metrics of the SVM model for Urucu Juruá validation data carbon 30 cm; (D) Error metrics of the SVM model training data carbon 100 cm; (E) Error metrics of the SVM model for Urucu validation data carbon 100 cm; (F) Error metrics of the SVM model for Urucu Juruá validation data carbon 100 cm. (images generated in RStudio program).

Mapping the SOCS is not an easy task, and according to Grimm et al. (2008) the difficulty of mapping the carbon stock may be due to the fact that the spatial distribution pattern of this attribute is highly variable due to small-scale differences in deposition processes, redistribution and intrinsic stabilization, combined with the high random variability of SOCS. For Somarathna et al. (2017) most of the models that they tested require a minimum of 15 samples/km2 to reach their maximum predictive capability for their particular study site. The authors conclude that the spatial prediction accuracy of soil carbon is less dependent on the model type than training sample size and the limiting factor in DSM is often the number of soil data used for model calibration. This reinforces our approach of making a denser sampling in a smaller and representative area (RA).

Comparing the results of this study with the literature, it is possible to observe that the accuracy values are reasonable because given the accessibility difficulties of the Central Amazon Rainforest and low sampling density available (0.0083 samples/km²) we delivered good results (especially for the Urucu block - 0.026 samples/km2). This was possible because we combined specialized pedological knowledge with algorithm parameters optimization, Previous Covariate Selection and model transferability through the RA approach. For example, Schillaci et al. (2017) modeling the topsoil carbon stock (30 cm depth) of agricultural land in a semi-arid Mediterranean region achieved an accuracy of R2=0.47 using a Stochastic Gradient Treeboost model and at a sample density of 0.11 samples/km². Mitran et al. (2018), studying the spatial distribution of soil carbon stocks (at 0-30 cm depth) in a semi-arid region of India achieved an accuracy of R^2=0.55 and RMSE=3.08 for Linear Regression Kriging and R^2=0.63 and RMSE=1.89 for Geographically Weighted Regression Kriging with a sample density of 0.0011 samples/km². Guo et al. (2015) working with a sampling density of 2.52 sample/km² obtained R^2=0.65 with RF modeling; Sreenivas et al., (2014) modeling carbon stock in India (at 30 cm depth) with RF found a result of R^2=0.85 and RMSE=2.36 kg m-2 using covariates from relief, climate, soils, remote sensing data and sample density of 0.0016 samples/km2.

Despite the R^2 being higher than that observed in our study, the error (RMSE) was comparatively higher (almost double).

Ramifehiarivo et al. (2017) mapping soil organic carbon stock in Madagascar (at 30 cm depth) with RF and using several covariates from relief, climate, soils and remote sensing data, achieved an accuracy of R^2=0.59 using a sample density of 0.003 profiles/km². Díaz et al. (2020) using RF and climate, vegetation, relief and soil covariates to predict organic carbon (at 30 cm depth) in Paramo ecosystem soils in Colombia obtained a result of R^2 from 0.48 to 0.52 by cross-validation and sample density of 0.026 profiles/km²; Ma et al. (2017) mapping soil organic carbon and other key soil properties (at 0-20 cm depth) to support agricultural production in Eastern China, using Cubist and Regression Kriging (RK) models, achieved an accuracy of R^2=0.25 and RMSE=1.12, R^2=0.31 and RMSE=1.08 for Cubist and RK model, respectively (using a density of 0,013 samples/km²). Our results also corroborate those of Hounkpatin et al. (2021) concluding that local models (similar to our case with the model fitted in RA) were generally more effective for predicting SOCS after testing on independent validation data (for example VU e VUJ in our case) than global models (TA, in our case). The authors found, using all covariates and a sampling density of 0,051 soil profiles/km2, values of R^2 ranging from 0.22 to 0.28 and RMSE from 27.7 to 44.9 tC ha-1 for the models fitted with data from North, Centre and South of Swedish, respectively.

The same pattern seems to be followed when analyzing the results for 100 cm soil depth. For example, Wiesmeier et al. (2011) modeling SOCS (at 100 cm depth), using RF algorithm and a sampling density of 0.03 profile/km² (120 soil samples in 3600 km²), achieved an accuracy of R^2=0.74 and RMSE= 5.46 kg C m-2, that is, errors (RMSE) greater than this study for the same depth (RMSE=4.00 and RMSE=3.55 with and without covariate selection respectively for RA approach (Table 7).

3.5.5 Spatial prediction of SOCS

The spatial prediction of SOCS30 and SOCS100, using the RT, RF and SVM algorithms are shown in Figures 19 and 20 (Urucu Block) and 21 (SOCS100, Juruá Block), respectively. Analyzing the Urucu block, there is a visual similarity between the maps of SOCS30 generated by RF and RT algorithms (Figure 19 e 21). Although the RT has a higher R^2 than the RF, the values predicted by the models are similar, ranging from 2.08 kg C. m-² to 4.36 kg C. m-² for RF and 2.29 kg C. m-² to 4.04 kg C. m-² for RT. The SVM algorithm predicted some very discrepant SOCS values, extrapolating unreal both the maximum and the minimum values (-43.70 kg C. m-² and 33.11 kg C. m-², respectively). For SOCS100 the predicted values of the RF algorithm ranged from 3.89 kg C. m-² to 10.64 kg C. m-², while the RT algorithm ranged from 5.12 kg C. m-² to 9.50 kg C. m-² Again, the results generated by the SVM algorithm is more discrepant, ranging from -5.10 kg C. m-² to 34.09 kg C. m-². The algorithm RF not only had a greater range of variation compared to RT, but also the predictions of SOCS were closer to what was observed in the RA.

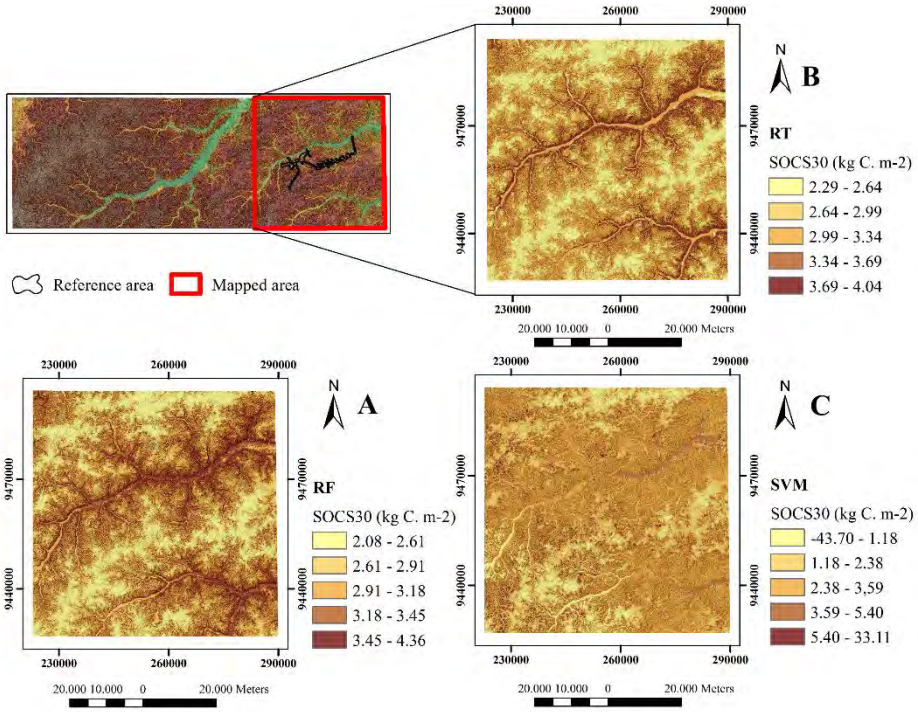


Figure 19. SOCS30 spatial prediction for the Urucu Block. (a) Random Forest model; (b) Regression Three model; (c) Support vector machine model.

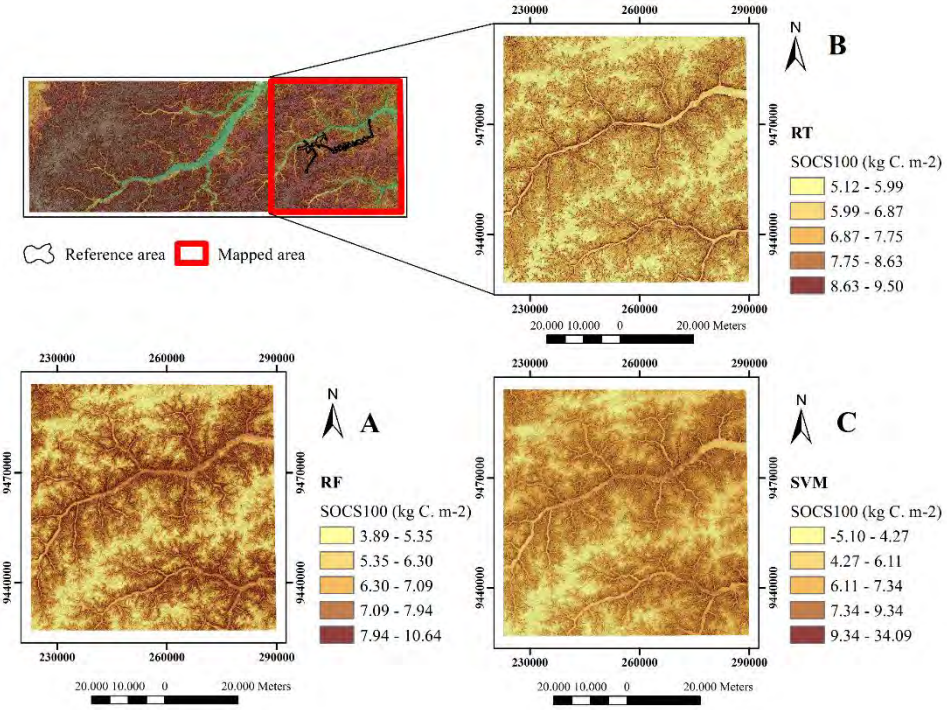


Figure 20. SOCS100 spatial prediction for the Urucu Block. (a) Random Forest model; (b) Regression Three model; (c) Support vector machine model.

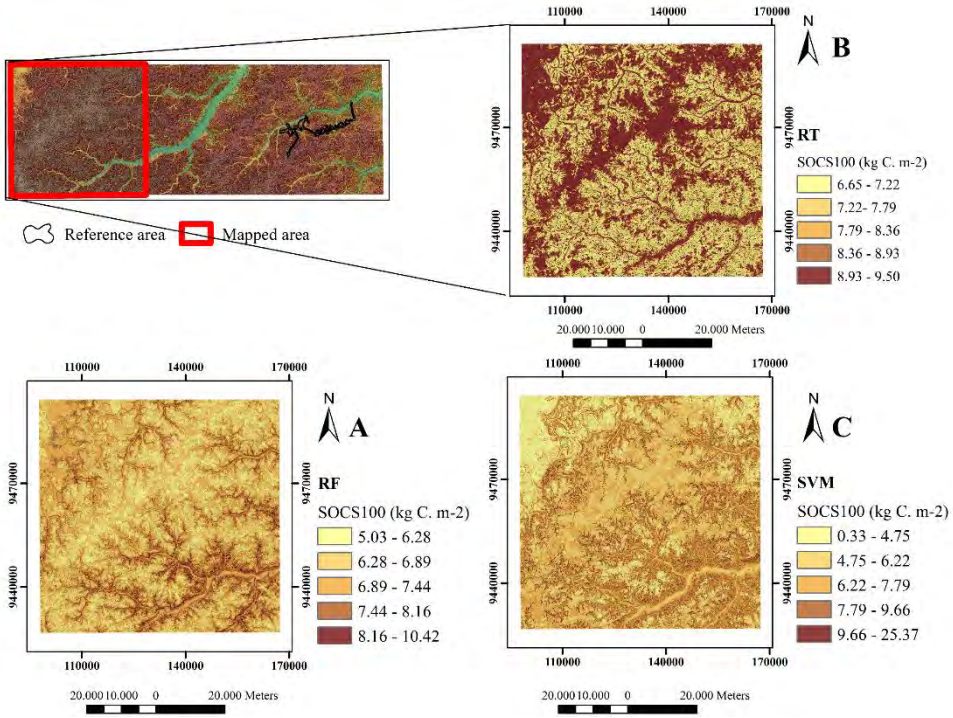


Figure 21. SOCS100 spatial prediction for the Juruá block. (a) Random Forest model; (b) Regression Three model; (c) Support vector machine model.

For SOCS100 in Juruá block, the values predicted by RF ranged from 4.97 kg C. m-² to 10.07 kg C. m-², while the RT algorithm predicted values ranging from 6.72 kg C. m-² to 9.02 kg C. m-². As observed for Urucu Block, the SVM algorithm presented the worst performance, ranging from a minimum of 0.33 kg C. m-² to a maximum value of 17.28 kg C. m-².

The hydromorphic flat tops generally presented low SOCS100 (between 5.06 and 6.8 kg C.m-2). The floodplain areas close to the rivers present intermediate values (between 6.8 and 7.4 kg C.m-2) and the slope areas with higher values (greater than 8.16 kg C.m-2).

A remarkable aspect, observed in Figure 21 (Juruá region), refers to the SOCS100 values predicted by the RT algorithm. The SOCS100 present a pattern inverse to those performed by the RF and SVM algorithms. As only at a depth of 100 cm the RT algorithm used the CND covariate as the second most important covariate (Table 7), it is believed that this is the cause of the difference in the results. This covariate is one of those that present minimum and maximum values that are more different from those observed in RA (Table 6). This demonstrates that although there is a potential transferability of the models developed in the RA to other blocks, the statistical differences of some covariates (Table 6), which did not appear so clearly in the Gower index (Figure 14), can impair the performance of some algorithms such as RT that uses only one tree compared with RF, which used more trees. Despite all the limitations imposed by the peculiar conditions of the study area, the results are promising and serve as a basis for other studies in Central Amazon.

3.6 CONCLUSIONS

A Despite the limitation of SOCS observation points available (0.0083 samples/km2) and its irregular spatial distribution along the remote forested area at Central Amazonian region in Brazil, it was possible to generate maps of SOCS at 30 and 100 cm soil depth using ML algorithms and relief covariates. This was possible because we combined the specialized pedological knowledge developed in a RA to previously select relief covariates that present a better correlation with SOCS. The Gower index was an important tool not only to show the transferability of the prediction models developed in the RA but also to highlight those relief covariates that most affect its transferability and which can be used to create or to remodel a RA.

The best map of SOCS30 was generated using RT algorithm (R^2=0.32) and the most important covariates used were SH, MRVBF and MRRTF. The SOCS30 map of the Urucu block region presented a range from 2.29 kg C. m-² to 4.04 kg C. m-².

The RF algorithm generated the most accurate maps to predict SOCS100 for the regions of the Urucu and Juruá Blocks (R^2=0.70 and 0.51, respectively). These maps presented higher accuracy and transferability than those developed to predict SOCS30. The covariates CND, MRVBF and SH were the most important used by the RF algorithm. The SOCS100 values of the maps generated to the Urucu Block region range from 3.89 kg C. m-² to 10.64 kg C. m-², while the Jurua block region (the farthest region in relation to the RA) range from 4.97 kg C. m-² to 10.07 kg C. m-².

4. CAPÍTULO II

PREDIÇÃO DA COMPOSIÇÃO GRANULOMÉTRICA DO SOLO EM ÁREAS REMOTAS DA AMAZÔNIA CENTRAL USANDO TÉCNICAS DE APRENDIZAGEM DE MÁQUINAS

4.1 RESUMO

A textura do solo é considerada um elemento importante na descrição, identificação e classificação do solo tendo grande influência no comportamento físico-hídrico e químico dos solos. Sob regiões de floresta amazônica, a densa cobertura vegetal e a limitada disponibilidade de vias de acesso, a execução de levantamentos e mapeamentos de solos torna-se bastante limitada. Quando existem dados, esses são relativamente pouco densos e a distribuição é bastante irregular. Nesse contexto, a utilização de algoritmos de aprendizagem de máquina (machine learning) associados a covariáveis de sensores remotos pode contribuir consideravelmente para a geração de mapas digitais da textura do solo. Os objetivos deste estudo foram: a) avaliar dois tipos diferentes de abordagem de amostragem (Área de Referência - AR e Área Total - AT) para desenvolver modelos de predição das frações granulométricas areia, silte e argila, em superfície e subsuperfície; b) avaliar a transferibilidade e o desempenho de três algoritmos de Aprendizado de Máquina (AM): "Regression Tree" (RT), "Random Forest" (RF) e "Support Vector Machine" (SVM). O local do estudo foi dividido em três blocos, denominados Urucu, Araracanga e Juruá. O conjunto de dados consistiu-se de 151 observações de areia, silte e argila em superfície e subsuperfície e 21 covariáveis (20 covariáveis de relevo e banda P do radar). O conjunto de dados de AR, utilizou 114 observações para treinar os algoritmos e 37 para validação e o conjunto de AT, 114 observações usadas para treinamento representando 75% dos dados e 37 para validação (25%). Com exceção da argila, o desempenho geral dos modelos foi melhor utilizando o conjunto de dados de AR. A maior acurácia das predições foi observada para a fração silte comparada as predições de areia e argila. O melhor desempenho foi obtido com o algoritmo de RF que gerou os mapas mais acurados de silte em superfície e subsuperfície para os Blocos de Urucu e Juruá (R^2 = 0,58 e 0,52, 0,51 e 0,56 respectivamente). Os valores de silte em superfície e subsurperficie dos mapas gerados para a região do Bloco de Urucu variam de 208.97 g kg^{-1} a 576.68 g kg^{-1} e 215,32 g kg^{-1}a 517,06 g kg^{-1}, enquanto para o bloco Juruá variam de 236.10g kg^{-1} a 555.70 g kg^{-1}e 229.83 g kg^{-1} a 460.56 g kg^{-1} respectivamente. Para areia e argila o RF também foi o melhor algoritmo, porém os valores de R^2 e os erros métricos foram menores comparados aos valores de silte. Apesar da baixa densidade de observação do conjunto de dados disponível, os resultados mostraram potencial dos algoritmos de AM para mapear as frações granulométricas do solo.

Palavras-chave: Mapeamento digital de atributos solo. Textura do solo. Propriedades do solo. Aprendizado de máquinas. SIG. Árvore de regressão. Floresta aleatória. Máquina de vetor de suporte.

4.2 ABSTRACT

Soil texture is considered an important element in the description, identification and classification of soil, having great influence on the physical-hydric and chemical behavior of soils. Under regions of the Amazon rainforest, the dense vegetation cover and the limited availability of access roads, the execution of surveys and soil mapping becomes quite impeded. When data do exist, they are relatively sparse and the distribution is quite uneven. In this context, the use of machine learning algorithms associated with remote sensor covariates can contribute considerably to the generation of digital maps of soil attributes. The objectives of this study were: a) to evaluate two different types of sampling approach (Reference Area - RA and Total Area - TA) to develop prediction models of clay, silt and sand granulometric fractions, in surface and subsurface; b) to evaluate the transferability and performance of three ML algorithms: regression tree (RT), random forest (RF) and support vector machine (SVM). The study site was divided into three blocks, called Urucu, Araracanga and Juruá blocks. The dataset consisted of 151 surface and subsurface sand, silt and clay observations and 21 covariates (20 relief and P-band radar covariates). The RA dataset used 114 observations to train the algorithms and 37 for validation and the TA set used 114 observations used for training representing 75% of the data and 37 for validation (25%). With the exception of clay, the overall performance of the models was better using the RA dataset. The highest accuracy of the predictions was observed for the silt fraction compared to the predictions of sand and clay. The prediction models developed to predict silt fractions showed greater accuracy and transferability than those developed to predict sand and clay. The best performance was obtained with the RF algorithm that generated the most accurate maps of surface and subsurface silt for the Urucu and Juruá Blocks (R^2 = 0.58 and 0.52, 0.51 and 0.56 respectively). The surface and subsurface silt values of the maps generated for the Urucu Block region range from 208.97 g kg-1 to 576.68 g kg-1 and 215.32 g kg-1 to 517.06 g kg-1, while for the Juruá block ranges from 236.10g kg-1 to 555.70 g kg-1 and 229.83 g kg-1 to 460.56 g kg-1 respectively. For sand and clay, RF was also the best algorithm, but R^2 values and metric errors were lower compared to silt values. Despite the low observation density of the available dataset, the results showed the potential of ML algorithms to map soil granulometric fractions.

Keywords: Digital mapping of soil attributes. Soil texture. Soil properties. Machine learning. GIS. Regression tree. Random forest. Support vector machine.

4.3 INTRODUÇÃO

A textura do solo é uma propriedade física fundamental que influencia fortemente muitas outras propriedades do solo. Essas frações granulométricas são relevantes para a produção agrícola, influenciando a fertilidade, capacidade de retenção de água, conteúdo de carbono, sendo responsável pela permeabilidade, porosidade e muitas outras propriedades do solo. Também desempenha um papel importante no sistema de classificação do solo, usado para identificar horizontes diagnósticos e classificar os solos em nível de família (AKPA et al., 2014; MEHRABI-GOHARI et al., 2019; SANTOS et al., 2018).

Os dados de frações granulométricas (areia, silte e argila), exercem funções importantes nos dados de entrada necessários para a maioria dos modelos hidrológicos, climáticos e ambientais. Também podem ser usados em algumas funções de pedotransferência para estimar propriedades como densidade, condutividade hidráulica e capacidade de retenção de água do solo (LIESS et al., 2012; MINASNY & HARTEMINK, 2011).

As geotecnologias aplicadas à análise da paisagem e modelagem ambiental possibilitam a extração de informações sobre as covariáveis relevantes na identificação das diferentes frações de textura do solo. O uso dessas covariáveis ambientais e derivadas de sensores remotos, em combinação com algoritmos de aprendizado de máquina (AM), mostra-se promissor na variabilidade espacial dos atributos dos solos, como por exemplo a textura do solo. Muitos pesquisadores vêm utilizando essas informações dentre outras diversas covariáveis na caracterização e predição da composição granulométrica do solo (AKPA et al., 2014; BHERING et al., 2016; LIESS et al., 2012; PINHEIRO et al., 2018; VAYSSE & LAGACHERIE, 2015).

Assim sendo, uma estratégia importante para gerar e expandir mapas detalhados da composição granulométrica do solo em áreas remotas, seria o uso dessas técnicas juntamente com a abordagem da área de referência (AR). Identificando todos os tipos de solo da região em uma área menor que seja representativa e, mapeando posteriormente áreas circunvizinhas, determinando suas relações espaciais.

A aplicação de tais técnicas pode ser uma importante estratégia para gerar e/ou expandir mapas detalhados da composição granulométrica do solo, especialmente em áreas remotas como é o caso da Amazônia Central. A combinação de métodos de AM, conhecimento especializado, análise exploratória dos dados de solo e covariáveis e a abordagem da área de referência (AR) podem ser extremamente úteis nesse tipo de mapeamento.

Considerando o potencial das técnicas AM, do uso de covariáveis derivadas de modelo digital de elevação e de radar, e da representatividade da AR analisada pela avaliação da semelhança de paisagens através do índice de similaridade de Gower, foi levantada a hipótese de que a textura do solo pode ser predita com boa qualidade (acurácia) combinando modernas técnicas de AM, seleção de covariáveis com auxílio de especialista a partir de uma AR representativa. Para validar ou recusar a hipótese o presente estudo teve os seguintes objetivos: 1- avaliar dois tipos diferentes de abordagem amostral (Área de Referência - AR e Área Total - AT) no treinamento dos algoritmos de AM; 2- avaliar duas categorias de seleção de covariáveis: "método wrapper", que se baseia na inferência feita por um modelo de AM calibrado, e "seleção de covariável" como etapa de pré-processamento, antes de calibrar o AM e 3- avaliar o desempenho de três algoritmos de AM: Regression Tree (RT), Random Forest (RF) e Support Vector Machine (SVM) na predição da composição granulométrica dos solos.

4.4 MATERIAL E MÉTODOS

4.4.1 Área de estudo

O estudo foi realizado em uma região da Floresta Amazônica entre os municípios de Carauari e Coari, a aproximadamente 640 km de Manaus, capital do Estado do Amazonas. A área de estudo é de aproximadamente 13.440 km², entre os paralelos 4°0'e 6 ° 0'S e 67 ° 0' e 64 °00'W (Figura 22). Segundo Köppen, o clima é classificado como Af (equatorial, com a temperatura do mês mais frio acima de 20 ° C, precipitação média anual de 2500 mm e período seco não acentuado). A região é remota, sendo possível o acesso somente por transporte aéreo e fluvial.

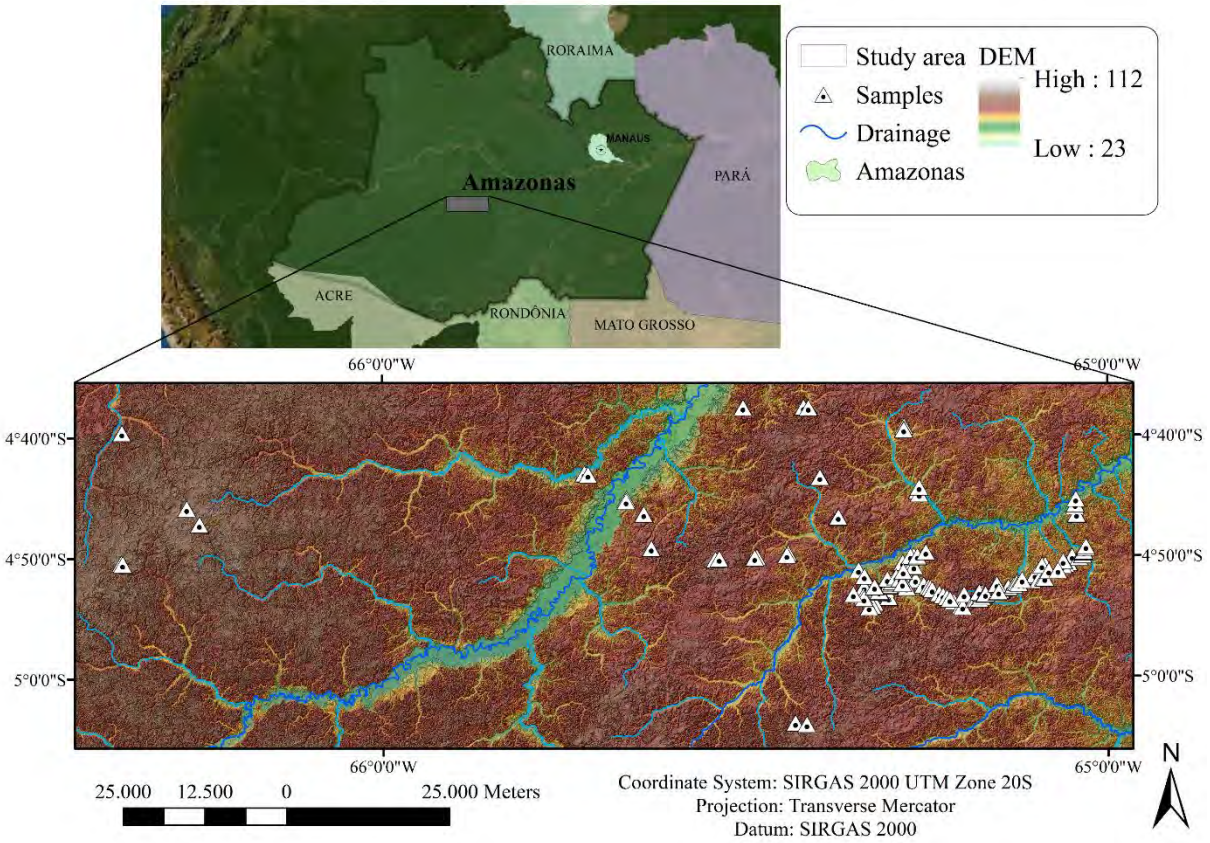


Figura 22. Localização da área de estudo. (Fonte: ArcGIS, elaborada pela Autora).

4.4.2 Banco de dados

Este trabalho utiliza os dados de solo coletados em duas etapas que foram descritas no capítulo 1 referente ao mapeamento de estoque de carbono. No entanto, foi utilizado um conjunto de dados maior inserindo 31 dados a mais compreendendo no total de 151 perfis de solos com a descrição da composição granulométrica (areia, silte e argila). A estratégia metodológica para prever areia, silte e argila em superfície e subsuperfície é apresentada no fluxograma da Figura 23.

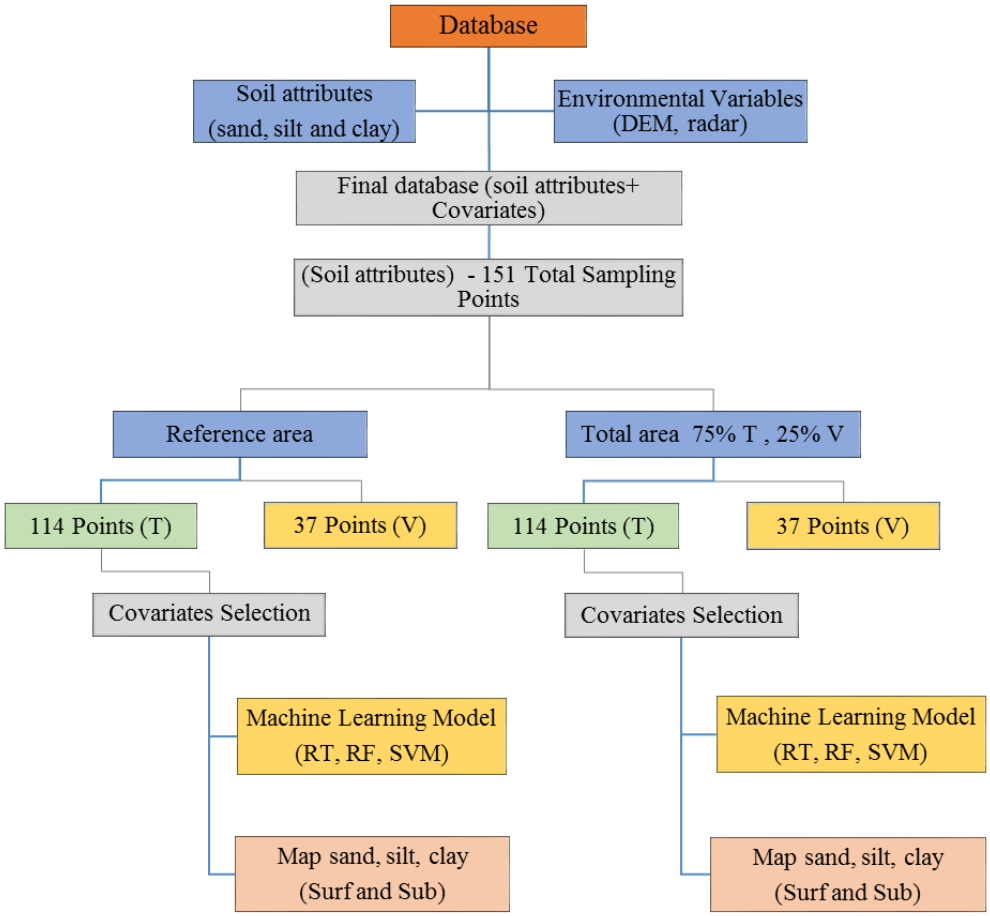


Figura 23. Fluxograma com a metodologia apresentada para mapeamento de areia, silte e argila em superfície e subsuperfície. T- Treino; V- Validação; RT- Regression Tree, RF- Random Forest; SVM-Support Vector Machine, sand- areia, silt- silte, clay- argila, surf- superfície, sub-subsuperficie.

A área de estudo de 13,440 km², coberta pela floresta Amazônica original, para fins de análise da transferibilidade dos modelos foi dividida em três blocos, denominados blocos de Urucu, Araracanga e Juruá. A divisão é mostrada na parte superior das Figuras 24 e 25. O critério de divisão dos blocos baseou-se na forma como os dados de relevo e imagens de radar foram gerados e disponibilizados pela Petrobras. A área de maior acessibilidade, definida como AR, compreende uma área de 80 km² localizada nas margens do Rio Urucu na Base de Operações Geólogo Pedro de Moura BOGPM - Petrobras-BR. Nessa área, de 2008 a 2013 foi realizado o levantamento e caracterização dos solos para fins de mapeamento de tipos e atributos do solo da AR.

Do total de perfis descritos 114 amostras com dados de areia, silte e argila, foram usados para treinar os modelos AM. No ano de 2018 foram realizadas campanhas de campo nas áreas remotas (16 clareiras), nos blocos de Urucu, Araracanga e Juruá. Nessas áreas, externa a AR, foram descritos e coletados um total de 40 perfis de solos para validação do mapeamento (extrapolação do conhecimento da AR), com total 37 pontos com dados de granulometria em superfície e subsuperfície (Figura 24). A essa abordagem em que os dados da AR foram usados para treinar modelos e os dados externos validar foi dado o nome de conjunto de dados 1 ou Área de Referência - AR)

A segunda abordagem testada foi a separação aleatória do total de pontos (n=151) em conjunto de treino (ajuste dos modelos de AM), n=114 (75% dos dados) e teste (validação dos

resultados), n=37 (25 % dos dados). A essa abordagem foi dado o nome de conjunto de dados 2 ou Área Total – AT (Figura 25). Os solos foram classificados de acordo com o Sistema Brasileiro de Classificação de Solos (SANTOS et al., 2018).

As principais classes de solos bem como o número de perfis, frequência e quantidade médias de areia, silte e argila em superfície e subsuperfície são mostrados na Tabela 10.

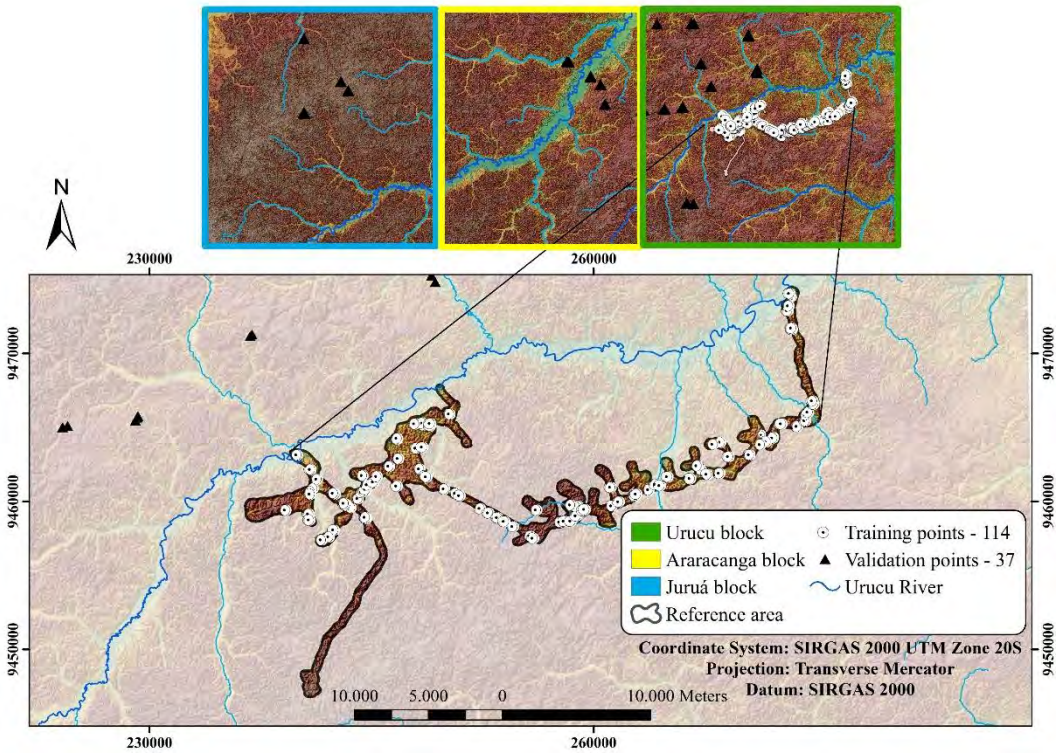


Figura 24. Amostragem baseada em área de referência. (Fonte: ArcGIS, elaborada pela Autora).

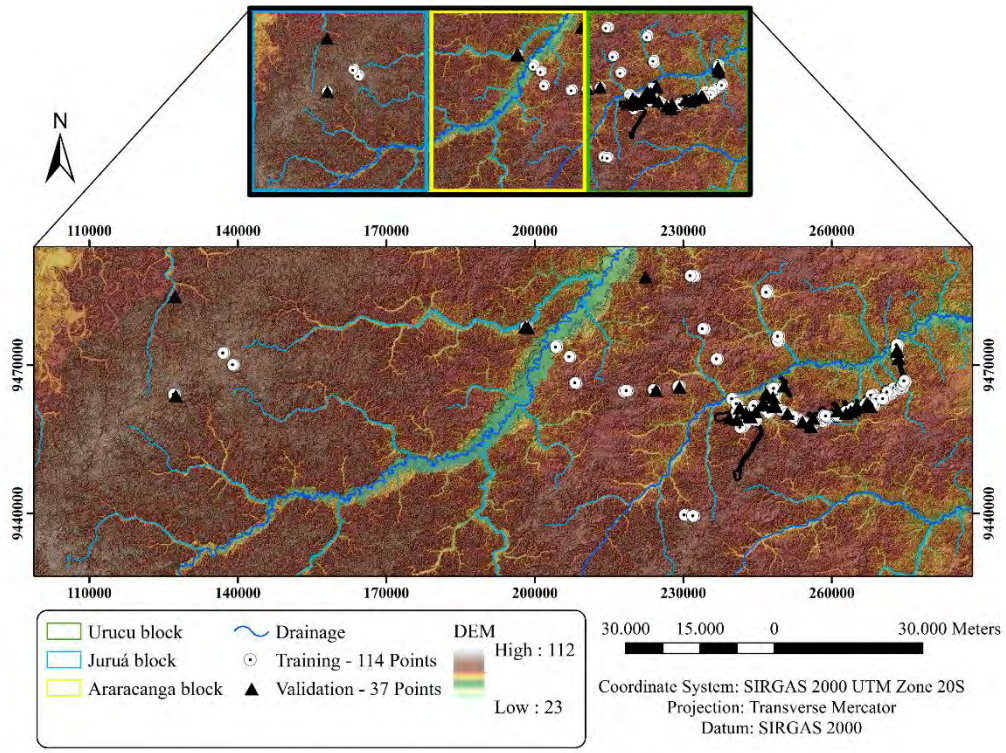


Figura 25. Amostragem baseada na área total, 75% treinamento 25% do conjunto de dados de validação. (Fonte: ArcGIS, elaborada pela Autora).

Tabela 10. Número de perfis de solo (n) e frequência de solo nos locais visitados.

*SiBCS[a]	AREIAA	AREIAB	SILTEA	SILTEB	ARGILAA	ARGILAB	N	Frequência (%)
Argissolo Amarelo	465,00 ± 154,45	335,76 ± 141,75	423,30 ± 145,69	357,00± 105,39	106,29± 50,66	312,70± 94,30	41	27,15
Argissolo vermelho	392,67 & 259,24	113,82 & 81,21	405,33 & 507,00	456,11 & 437,91	233,80 & 202,00	430,10 & 480,90	2	1.32
Argissolo Vermelho Amarelo	440,60 ± 152,64	303,08 ± 118,05	400,00 ± 146,62	326,04 ± 93,10	159,60 ± 64,00	368,03 ± 67,77	29	19,20
Argissolo Acizentado	560,70 ± 54,88	477,20 ± 71,87	341,30± 49,17	325,00± 62,77	97,98± 22,50	197,80± 40,93	3	1.98
Cambissolo Háplico	454,20 ± 122,50	357,60 ± 141,38	333,60± 99,95	306,20 ± 84,53	219,57± 80,03	333,70± 97,48	49	32.45
Cambissolo Flúvico	557,03 & 495,50	586,75 & 394,12	338,55 & 425,00	295,36 & 428,54	104,41 & 79,50	117,88 & 177,34	2	1.32
Espodossolos Humilúvicos	545,00	532,49	379,00	412,30	76,00	55,21	1	0.66
Espodossolos Ferri-Humilúvicos	702,10 ± 159,84	689,40 ± 119,04	258,57 ± 137,96	248,20 ± 99,81	39,36 ± 34,84	62,35 ± 48,58	4	2.65
Neossolo Quartzarênico	585,55	605.86	338,99	344,48	75,46	55,38	1	0.66
Neossolos Flúvicos	918,00 & 598,22	802,15 & 512,71	26,00 & 339,09	143,81 & 367,40	56,00 & 62,70	54,05 & 119,89	2	1.32
Planossolo Háplico	231,80 & 298,33	253,22 & 241,55	707,40 & 694,53	651,26 & 587,63	60,80 & 7,13	95,53 & 170,83	2	1.32
Gleissolos Háplicos	393,90 ± 198,78	341,35 ± 208,00	481,10± 174,79	422,90± 118,21	123,83 ± 99,02	235,72 ± 120,82	14	9,27
Gleissolos Melânicos	80,00	43,98	497,00	424,00	423,00	532,02	1	0.66
Total							151	100

*Sistema Brasileiro de Classificação de Solos. **Nota: Esta é a equivalência parcial entre classes de solos em alto nível categórico em SiBCS. A- superfície; B- subsuperfície; Unidade da textura g kg^{-1}.

4.4.3 Composição granulométrica do solo

No levantamento de solo, os perfis foram divididos em horizontes (A, B, C-horizontes). Os conteúdos de areia, silte e argila dos horizontes superficiais (Surf) e subsuperficiais (Sub) foram calculados como a média ponderada de profundidade ao longo dos horizontes. Para superfície foram considerados os horizontes A, AB, AC e AE e para subsuperfície os horizontes BA e B até 100cm. Os horizontes BC e C não foram incluídos, CA, C foram inclusos quando não havia horizonte B, ou seja, para solos como por exemplo: Neossolo Quartzarênico e Neossolos Flúvicos. Além disso, os transicionais AC, AE foram considerados pois, são bem semelhantes ao A, e também BA com características mais próximas de B.

$$PSF_{surf/sub} = \sum_{i=1}^{n} PSF_i * T_i \Big/ \sum_{i=1}^{n} T_i$$

Onde:

$PSF_{Surf/Sub}$ – é a fração do tamanho de partícula na camada desejada (superfície ou subsuperfície);
PSF_i - é o conteúdo de PSF no horizonte i;
Ti - é a espessura (m) da porção do horizonte i que se encontra dentro da camada desejada;
n - é o número de horizontes que possuem uma porção dentro da camada desejada.

4.4.4 Covariáveis ambientais

Como covariáveis ambientais, os atributos do terreno foram derivados do modelo digital de elevação (DEM) hidrologicamente consistente com resolução espacial de 20m, sendo também utilizado o coeficiente de retroespalhamento da banda P do radar da polarização HH. Maior detalhamento das covariáveis ambientais podem ser consultadas nos itens 3.4.4 e 3.4.5 do capítulo 1. As descrições, siglas e unidades das covariáveis podem ser consultadas na Tabela 3 do capítulo 1.

4.4.5 Modelos preditivos

Os algoritmos de AM, RT, RF e SVM foram utilizados na modelagem da composição granulométrica do solo em superfície e subsuperfície. Cada algoritmo pode encontrar relacionamentos complexos entre areia, silte, argila e covariáveis ambientais de maneiras diferentes. Os hiperparametros utilizados na modelagem podem ser consultados na Tabela 4 do capítulo 1.

4.4.6 Similaridade da paisagem entre as áreas de modelagem

A similaridade das condições ambientais das áreas e a semelhança entre AR e os blocos Urucu, Araracanga e Juruá foi abordada como no mapeamento de estoque de carbono e o método e avaliação encontram-se no item 3.4.6 do Capítulo 1.

4.4.7 Análise exploratória e seleção de covariáveis

Foram realizadas as análises de multicolinearidade e colinearidade e utilizado o valor de VIF>10 de acordo Gujarati (2000). Segundo Wadoux et al (2020) apenas um terço dos trabalhos que utilizam AM adotam um critério de seleção de covariáveis como etapa de pré-processamento dos dados antes de disponibilizá-los para os algoritmos treinarem um modelo. Neste estudo, foram testadas duas formas de desenvolvimento dos modelos, denominadas "método wrapper" e "seleção prévia de covariáveis". No primeiro caso ("método wrapper"), todas as covariáveis foram disponibilizadas para os algoritmos desenvolverem o treinamento, ou seja, um processo mais automático. No segundo caso (seleção prévia de covariável), foi considerado o conhecimento do especialista juntamente com a matriz de correlação como ferramenta de suporte para seleção das covariáveis que explicariam melhor a relação solo-relevo-vegetação (SRV), proposta por Ceddia et al. (2015). A seleção das covariáveis de entrada seguiu três etapas de pré-processamento, antes da calibração dos modelos de AM: a) análise exploratória para avaliação de dados anômalos que poderiam influenciar erroneamente os resultados; b) avaliação da correlação de Pearson entre a composição granulométrica e as covariáveis de relevo para melhor compreensão dos dados e da relação pedológica e ambiental e; c) avaliação da multicolinearidade. Mas detalhes podem ser vistos no item 3.4.8 do capítulo 1.

4.4.8 Avaliação da acurácia dos modelos

A acurácia dos modelos foi avaliada por meio dos seguintes parâmetros: coeficiente de determinação R^2, erro absoluto médio (MAE) e erro quadrático médio (RMSE), as equações estão descritas no item 3.4.9 do capítulo 1.

4.5 RESULTADOS E DISCUSSÃO

4.5.1 Estatísticas descritivas

A Tabela 11 apresenta as estatísticas dos dados de areia, silte e argila em superfície e subsuperfície para as duas abordagens utilizadas (Conjunto de dados 1- AR e conjunto de dados 2 - AT). Para fins de organização e comparação, a codificação dos dados apresenta algumas diferenças de acordo com a abordagem utilizada. Os códigos representam os termos associados à abordagem AR (dataset-1). Neste caso, os dados são apresentados da seguinte forma: 151 amostras totais (W), 114 amostras de treino AR (T-treinamento) e 37 amostras utilizadas como validação externa, os quais foram coletados nas clareiras remotas (V). Ainda nesta primeira parte da Tabela 11, as estatísticas dos 37 dados de validação são apresentadas separadamente, seguindo a divisão da área de estudo em blocos conforme a parte superior das Figuras 24 e 25, a saber: Bloco Urucu (VU - 21 observações), Bloco Araracanga (VA - 11 observações) e Bloco Juruá (VJ - 5 observações). As estatísticas das 32 observações dos blocos Urucu e Araracanga (VUA) e das 26 observações dos blocos Urucu e Juruá (VUJ) também são apresentadas. Com essas divisões, buscamos entender melhor a variação dos dados de areia, silte e argila em regiões localizadas a diferentes distâncias da AR e com diferentes padrões de covariáveis do relevo. Essa divisão também ajuda a entender o desempenho dos modelos de AM usados na predição dos atributos estudados. Na segunda parte da Tabela 11 apresenta as estatísticas de dados quando a abordagem de área total (AT) é adotada (dataset-2). Neste caso, além do conjunto de dados com 151 observações (W), são apresentadas as estatísticas de 114 dados de treinamento (T - 75%) e 37 dados de validação (V-25%), que foram selecionados aleatoriamente (Figura 25).

Analisando os dados do dataset 1, na areia em superfície as médias de VUJ e VJ são bem abaixo da média total e média do treinamento. Já a média da VA e VUA são mais próximas da média total, mas os valores mínimos são bem menores. O maior teor de areia em superfície está no conjunto de dataset 1 que representa a AR (918g.kg-1).

A areia em subsuperfície teve a menor amplitude e menor média nos pontos do bloco de Juruá (VJ). Comparando os dados de T e V o valor mínimo é menor em V. Isso pode ser um problema já que o modelo irá tentar predizer dentro do intervalo em que foi treinado.

Para silte em superfície o que mais se diferenciou foram os dados de VJ, isso porque nesse bloco de Juruá só tem 5 pontos, ou seja, é mais difícil cinco pontos representar toda a variação da W e T. Nesse caso a validação (V) tem valores máximos maiores que o treino (T). Isso talvez seja um problema porque os modelos de AM não são tão bons em extrapolar, tendendo a predizer no intervalo de valores que foram treinados.

Para silte em subsuperfície a amplitude de variação do treinamento (T) é maior que da validação (V), e relativamente próximos aos valores de máximo, mínimo e da média. Isso mostra que para essa fração, nessa profundidade o V é mais representativo do T. Novamente podemos observar que os pontos do bloco de Juruá, que tem menos pontos, é o que mais de difere dos dados de W e T.

Para argila em superfície os valores mínimos de treino são de 33 $g.kg^{-1}$ e para V bem menores 4,67 $g.kg^{-1}$ para V e VA e 6 $g.kg^{-1}$ para VU e VUJ. Novamente, os valores de treinamento (T) não cobrem toda a amplitude de variação (W). Para as áreas de Urucu (VU) e de Juruá (VJ) os valores máximos são bem menores que o observado na AR (T).

Para argila em subsuperfície, a amplitude de variação dos dados de treino é maior que validação. E relativamente próximos entre si. Nesse caso específico tanto os valores máximos como mínimo que são observados em V, VU, VUA, VUJ são referentes ao bloco de Urucu (VU).

Analisando os dados do dataset 2, a areia tanto em superfície quanto em subsuperfície apresenta valores máximos de V maiores que o treino, podendo talvez influenciar negativamente a avaliação do desempenho dos modelos.

Para o silte superficial ocorre variação semelhante, mas nesse caso o valor mínimo da V, é menor que o valor mínimo do T, ou seja, os modelos foram treinados com valores que não contemplam toda a variação do dataset de validação. Isso não ocorre no silte em subsuperfície.

Para a argila em superfície a amplitude de variação dos valores usados para treinar os modelos (T) contempla toda a possibilidade de valores da validação (V). Já em subsuperfície os dados de treino têm valor mínimo muito superior ao valor mínimo dos dados de validação.

De acordo com Ceddia et al. (2017) a classificação textural do solo na região é predominantemente como franco na superfície e franco- argiloso na subsuperfície, o que reflete na constituição do material de origem do local de estudo composto principalmente de arenito muito fino, argila ferruginosa e siltito. O autor também destaca o fato de encontrar valores relativamente elevados de silte na região, tanto na superfície quanto na subsuperfície já que esta quantidade de silte não é frequentemente encontrada nos solos brasileiros.

Os valores do coeficiente de variação (CV) elevados (>28%), em todos os casos, caracterizam a heterogeneidade dos conjuntos de amostras tanto nos dados de treinamento quanto nos dados de validação.

Tabela 11. Estatística descritiva da textura do solo.

Variáveis	Dataset	n	Mínimo	Máximo	Média	Mediana	SD	Sk	k	CV (%)
Área de Referência (dataset 1)										
Areia Surf (g kg^{-1})	**W**	**151**	80,00	918,00	458,73	437,20	156,55	0,36	-0,11	34.12
	T	**114**	182,13	918,00	468,60	450,75	154,06	0,48	-0,07	32,87
	V	**37**	80,00	793,33	428,33	409,58	162,33	0,11	-0,63	37,89
	VU	**21**	225,47	721,00	425,13	401,52	144,06	0,46	-0,99	33,88
	VUA	**32**	80,00	793,33	453,48	432,41	158,25	-0,06	-0,45	34,89
	VUJ	**26**	151,00	721,00	394,78	361,94	146,72	0,58	-0,66	37,16
	VA	**11**	80,00	793,33	507,62	548,81	176,66	-0,89	0,77	-
	VJ	**5**	151,00	359,73	267,34	273,00	75,05	-0,36	-1,38	-
Areia Sub (g kg^{-1})	**W**	**151**	43,98	855,00	353,19	314,42	160,52	0,50	-0,16	45.44
	T	**114**	81,21	855,00	351,93	307,02	155,12	0,65	0,24	44,07
	V	**37**	43,98	695,31	357,05	338,08	178,34	0,16	-1,09	49,94
	VU	**21**	86,46	674,50	342,27	314,42	169,47	0,41	-1,00	49,51
	VUA	**32**	43,98	695,31	382,72	377,13	177,10	-0,05	-1,03	46,27
	VUJ	**26**	86,46	674,50	313,52	278,77	165,08	0,66	-0,66	52,65
	VA	**11**	43,98	695,31	459,93	493,52	172,63	-0,97	0,51	-
	VJ	**5**	99,06	279,44	192,76	201,13	64,48	-0,12	-1,45	-
Silte Surf (g kg^{-1})	**W**	**151**	26,00	792,00	389,59	375,20	145,11	0,16	-0,27	37.24
	T	**114**	26,00	687,00	364,78	351,78	131,32	0,03	-0,12	36,00
	V	**37**	155,00	792,00	466,03	481,00	160,18	-0,11	-0,94	34,37
	VU	**21**	155,00	688,59	476,37	481,00	142,24	-0,42	-0,59	29,86
	VUA	**32**	155,00	688,59	434,45	431,00	146,20	-0,11	-1,02	33,65

Continua...

Continuação da **Tabela 11.**

Variáveis	Dataset	n	Mínimo	Máximo	Média	Mediana	SD	Sk	k	CV (%)
Área de Referência (dataset 1)										
Silte Surf (g kg^{-1})	**VUJ**	**26**	155,00	792,00	513,25	536,34	152,06	-0,43	-0,45	29,63
	VA	**11**	202,00	534,25	354,41	321,43	122,73	0,19	-1,70	-
	VJ	**5**	597,00	792,00	668,14	643,00	78,86	0,56	-1,59	-
	W	**151**	84,60	600,24	339,10	340,31	105,61	0,05	-0,21	31.14
	T	**114**	84,60	600,24	332,00	327,89	101,62	-0,04	0,01	30,61
	V	**37**	168,32	570,11	360,96	349,40	115,77	0,14	-1,05	32,07
Silte Sub (g kg^{-1})	**VU**	**21**	191,77	570,11	359,58	343,95	113,50	0,39	-0,87	31,56
	VUA	**32**	168,32	570,11	342,35	341,62	111,51	0,37	-0,74	32,57
	VUJ	**26**	191,77	570,11	382,75	374,6	115,12	0,06	-1,16	30,08
	VA	**11**	168,32	485,88	309,46	303,98	104,73	0,19	-1,42	-
	VJ	**5**	388,66	551,38	480,08	479,53	61,30	-0,30	-1,61	-
	W	**151**	4,67	500,06	152,54	140,09	86,08	0,87	1,12	56.43
	T	**114**	33,83	500,06	169,68	155,53	81,93	0,79	1,08	48,29
	V	**37**	4,67	423,00	99,71	78,14	77,48	1,99	5,82	77,71
Argila Surf (g kg^{-1})	**VU**	**21**	6,00	203,61	98,49	86,00	51,29	0,23	-0,65	52,08
	VUA	**32**	4,67	423,00	105,21	82,75	80,83	1,87	5,04	76,83
	VUJ	**26**	6,00	203,61	91,95	78,82	50,46	0,38	-0,62	54,88
	VA	**11**	4,67	423,00	118,03	73,29	121,32	1,34	0,83	-
	VJ	**5**	27,55	130,00	64,51	57,00	39.94	0,66	-1,37	-
	W	**151**	13,00	573,23	308,10	326,76	111,45	-0,28	-0,27	36.17
	T	**114**	13,00	530,58	314,60	330,53	108,02	-0,60	0,00	34,34
	V	**37**	70,48	573,23	288,07	267,14	120,77	0,52	-0,43	41,92
Argila Sub (g kg^{-1})	**VU**	**21**	70,48	573,23	298,13	288,95	131,68	0,36	-0,76	44,17
	VUA	**32**	70,48	573,23	281,86	261,35	127,20	0,62	-0,50	45,13
	VUJ	**26**	70,48	573,23	303,84	294,37	120,78	0,27	-0,48	39,75
	VA	**11**	150,68	532,02	250,79	199,86	117,73	1,12	0,20	-
	VJ	**5**	259.74	410.21	327.82	339.96	59.92	0.13	-1.86	-
Área total (dataset 2)										
Variáveis	**Dataset**	**n**	**Mínimo**	**Máximo**	**Média**	**Mediana**	**SD**	**Sk**	**k**	**CV (%)**
Areia Surf (g kg^{-1})	**W**	**151**	80,00	918,00	458,73	437,20	156,55	0,36	-0,11	34.12
	T	**114**	80,00	883,50	451,26	435,00	150,85	0,21	-0,36	33,43
	V	**37**	208,00	918,00	481,77	460,50	173,12	0,59	-0,19	35,93
Areia Sub (g kg^{-1})	**W**	**151**	43,98	855,00	353,48	314,42	160,62	0,50	-0,16	45.43
	T	**114**	43,98	695,31	337,60	307,92	145,78	0,24	-0,75	43,18
	V	**37**	102,24	855,00	402,42	381,03	193,72	0,54	-0,65	48,14
Silte Surf (g kg^{-1})	**W**	**151**	26,00	792,00	389,59	375,20	145,11	0,16	-0,27	37.24
	T	**114**	58,50	792,00	397,80	378,46	139,42	0,21	-0,40	35,05
	V	**37**	26,00	696,00	364,30	350,57	160,82	0,17	-0,32	44,14
Silte Sub (g kg^{-1})	**W**	**151**	84,60	600,24	339,10	340,31	105,61	0,05	-0,21	31.14
	T	**114**	84,60	600,24	349,35	349,44	100,72	0,07	-0,21	28,83
	V	**37**	112,23	582,00	309,00	306,03	116,32	0,23	-0,43	37,64

Continua...

Continuação da **Tabela 11.**

Variáveis	Dataset	n	Mínimo	Máximo	Média	Mediana	SD	Sk	k	CV (%)
Área total (dataset 2)										
Argila Surf (g kg^{-1})	**W**	**151**	4,67	500,06	152,54	140,09	86,08	0,87	1,12	56.43
	T	**114**	4,67	500,06	152,06	139,68	89,92	0,87	1,10	59,13
	V	**37**	39,00	351,19	154,02	142,75	74,12	0,81	0,33	48,12
Argila Sub (g kg^{-1})	**W**	**151**	13,00	573,23	308,52	326,76	111,63	-0,28	-0,28	36.18
	T	**114**	70,48	573,23	314,62	327,94	105,80	-0,09	-0,57	33,63
	V	**37**	13,00	530,58	289,72	317,19	127,69	-0,49	-0,46	44,07

W: Dataset total; T: Treino dataset; V: Validação dataset; VU: Validação do bloco de Urucu; VA: bloco de Araracanga; VJ: bloco de Juruá; VUA: Validação bloco de Urucu/Araracanga; VUJ: Validação bloco de Urucu/Juruá n: número de observações; SD: desvio padrão; Sk: assimetria; K: Curtose; surf: superfície sub; subsuperfície.

4.5.2 Correlação e importância das covariáveis com a textura do solo

As matrizes de correlação de Pearson entre as covariáveis e os valores de areia, silte e argila superficial e subsuperficial, usando os conjuntos de dados AR e AT, são mostradas na Figura 26. Os dois conjuntos de dados apresentaram semelhanças quanto aos valores de correlação. Tanto no conjunto de AR quanto no de AT, a maior parte das covariáveis tiveram correlações menores que 0,50. Os maiores valores de correlação foram dos atributos de argila em AR Figura 26A e silte em AT Figura 26B. Ambos os conjuntos proporcionam o mesmo número de dados usados para treinamento e validação (114T/37V), a principal mudança observada, está na disposição dessas amostras e nos valores de amplitude dos valores mínimo e máximo de alguns atributos dos conjuntos de dados (Tabela 11).

Na seleção prévia de covariável, foi considerado o conhecimento do especialista juntamente com a matriz de correlação como ferramenta de suporte para seleção das covariáveis que explicariam melhor a relação solo-relevo-vegetação (SRV), proposta por Ceddia et al. (2015).

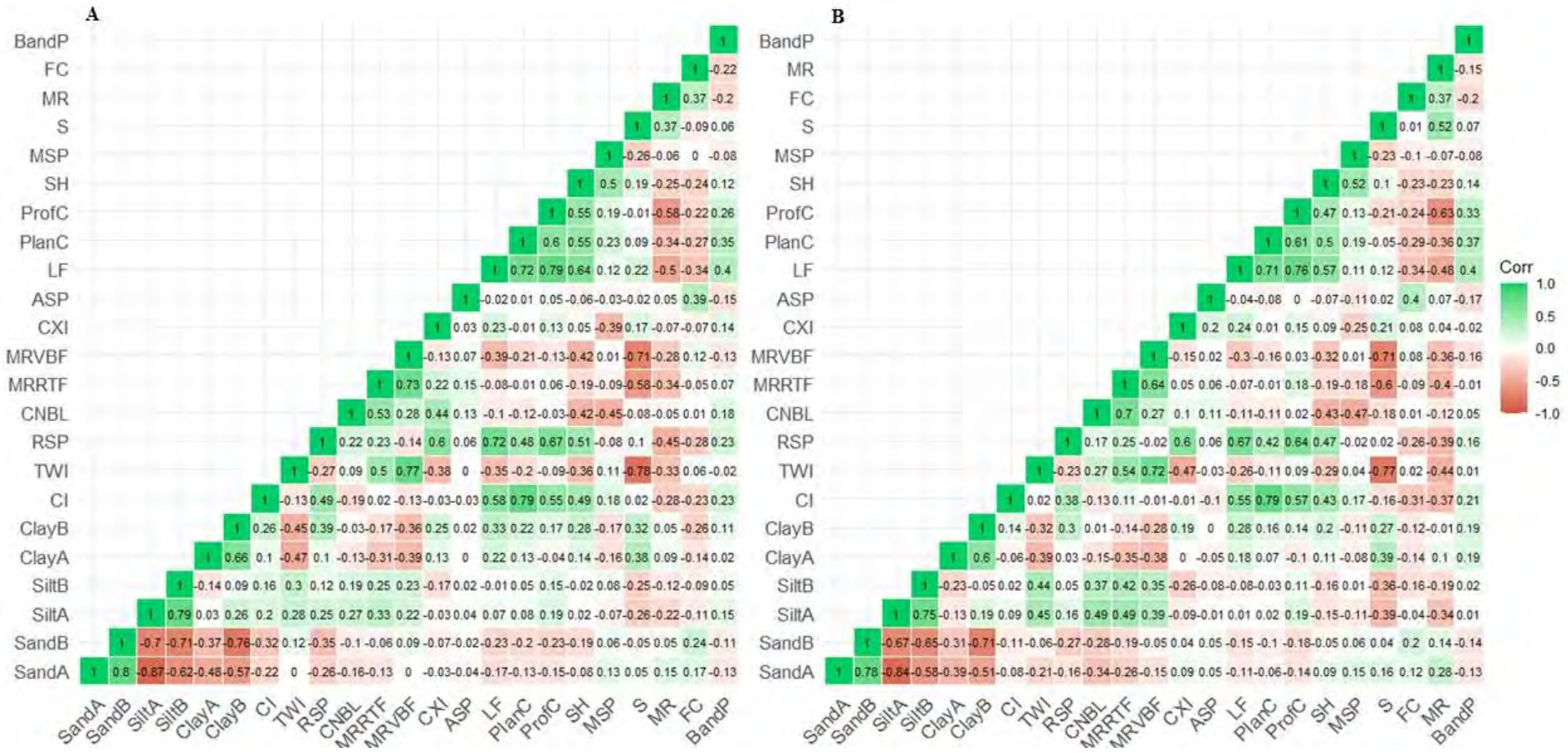


Figura 26. Matriz de correlação de covariáveis ambientais. (A) Matriz de covariáveis correlacionadas com os dados de composição granulométrica em área de referência (AR). (B) Matriz de covariáveis correlacionadas com os dados de composição granulométrica em área total (AT). (imagem generada no programa RStudio).

Os resultados do índice geral de Gower (Figura 14 do capitulo 1) mostram que há pouca dissimilaridade entre AR e os blocos de Urucu, Araracanga e Juruá (valores de 0,155, 0,164 e 0,171, respectivamente). No entanto, mesmo que esses valores de dissimilaridade sejam baixos, a maior parte das covariáveis que tiveram maiores correlações (Figura 26) também tiveram maiores contribuições de valores de índices de dissimilaridade em relação ao índice geral de Gower (RSP, CI, MRVBF, MRRTF, LF) (Figura 14 do capitulo 1).

Como apenas o RF obteve valores na modelagem significativos em relação a RT, somente foi mostrada a importância das covariáveis para efetuar predições dos atributos em superfície e subsuperfície fornecida pelo modelo RF apresentada na Figura 27. A importância das correlações para areia em superfície e subsuperfície e silte em superfície e subsuperfície refere-se ao conjunto de dados de AR, já para argila ao conjunto de dados de AT. Suas correlações com os atributos (Figura 26), também variaram quanto ao ranking de importância referido pelo modelo.

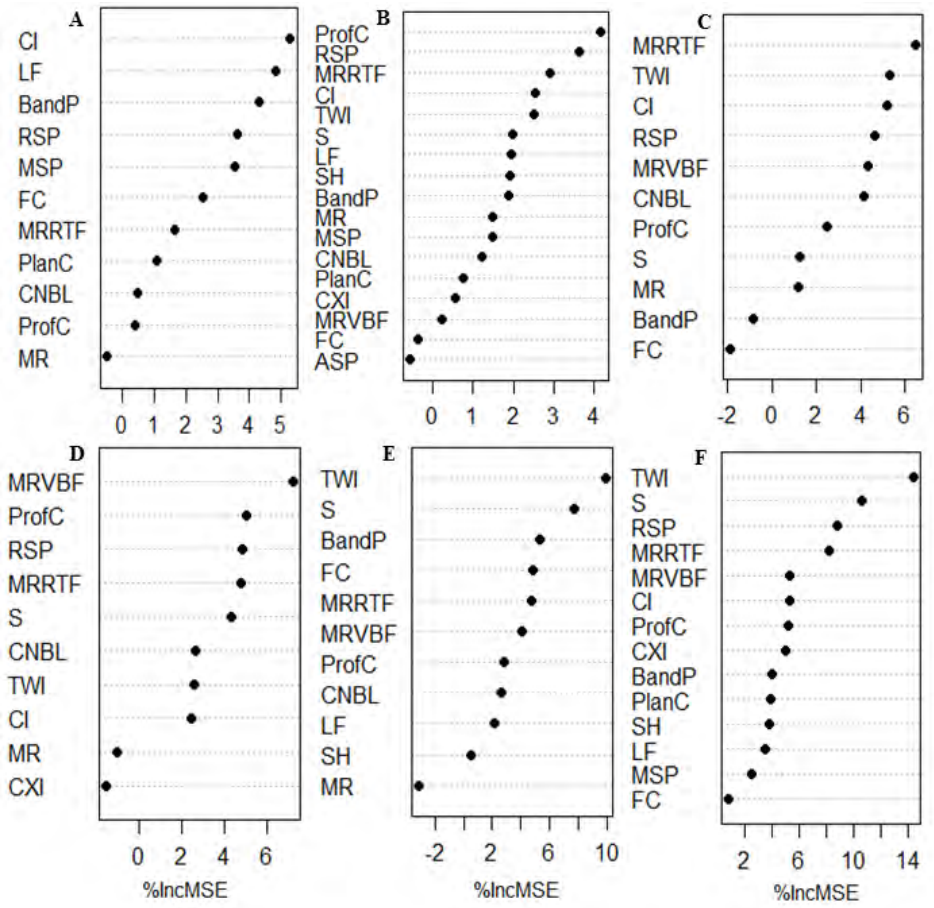


Figura 27. Importância das covariáveis preditoras para os atributos avaliados no modelo RF. (A) Areia Surf (B); Areia Sub; (C) Silte Surf; (D) Silte Sub; (E) Argila Surf; (F) Argila Sub. *Surf- superfície; Sub- subsuperfície. (imagem generada no programa RStudio).

Os atributos morfométricos do relevo podem auxiliar na discriminação e delimitação das unidades de mapeamento de solos, bem como das frações texturais do solo. Em suma, a questão é se as correlações entre as covariáveis de relevo e as frações texturais seguem um conhecimento existente ao longo da região do estudo. Em geral, tanto os valores dos coeficientes de correlação encontrados, quanto as covariáveis mais importantes associadas a composição granulométrica nesse estudo, coincidem com as informações e hipóteses levantadas no projeto RADAMBRASIL (1978), bem como do estudo de Villela (2013), Ceddia et al.

(2015) e Ceddia et al. (2017), onde verificou-se que o material de origem, relevo, vegetação e clima, atuaram de forma interdependente, explicando a relação dos tipos de solo e seus respectivos atributos. Estas mesmas covariáveis foram contextualizadas no modelo Soil-Relief-Vegetation (SRV) da Figura 28, que representa uma adaptação ao apresentado por Ceddia et al. (2015).

Em geral, as covariáveis CI, LF, BandP, RSP e ProfC, foram definidas como as mais importantes na predição de areia. O índice de convergência, representou o comportamento do escoamento superficial, que foi influenciado principalmente pelas formas dos relevos (landforms- LF). Os teores de areia, são mais elevados principalmente nas regiões das calhas dos rios, onde os valores negativos de CI, determinam principalmente as áreas convergentes referentes aos canais de baixada. Os valores positivos de CI indicam as áreas divergentes, onde predominam os topos bem drenados e os topos de encostas, a partir dos quais o escoamento superficial ocorre em todas as direções. Nessas regiões os teores de areia são menores. O RSP pode ser aplicado para identificar características topográficas. Seus valores variam de 0 a 1. Os valores mais próximos a 0 são caracterizados pelas regiões de baixadas, compondo os vales em V e U, que possuem altos teores de areia. Os valores mais próximos de 1 representam os declives superiores e topos de cume com baixos teores de areia. A curvatura do perfil (ProfC), expressou principalmente a diferença entre curvaturas convexas das côncavas, influenciando a velocidade do fluxo superficial das partes mais elevadas para as mais baixas (Figura 28 A). Permitiu, além disso, maior poder discriminatório para os solos de topo de elevação, na distinção entre os solos bem drenados (superfícies convexas) e os ambientes imperfeitamente drenados (superfícies planas). Os vales em formato de V também foram identificados com maior facilidade por meio das superfícies côncavas identificadas no modelo.

As covariáveis MRRTF, TWI, MRVBF e ProfC, possuem correlações positivas com o silte. Os topos planos de altitude são representados pelos valores altos de MRRTF e os fundos de vale de encostas, definidos pelos maiores valores de MRVBF. Essas covariáveis associadas a TWI, que caracteriza a distribuição espacial das zonas de saturação superficial, acrescentam importantes informações referentes principalmente as áreas com a presença de solos hidromórficos. Essas áreas são destacadas pelos teores de silte mais elevados (Figura 28 B), onde se encontram as regiões de baixadas (MU2 - Gleissolos) e as regiões de topos planos de altitude (MU4 - Argissolos acinzentados e Gleissolos). A curvatura do perfil auxiliou também no complemento nas regiões de solos bem drenados (superfícies convexas) e os ambientes imperfeitamente drenados (superfícies planas), principalmente em subsuperfície.

A combinação das covariáveis Slope e TWI, permitiu identificar as regiões com maiores teores de argila, onde se encontram as unidades MU1 e MU3. As regiões MU1 são representadas por encostas com declividade mais acentuadas e geralmente mais próximas das grandes redes de drenagem onde a declividade possui influência sobre a velocidade dos fluxos superficiais e subsuperficiais. A declividade tem grande potencial de auxiliar na predição dos solos Argissolos vermelho amarelo, onde predomina os maiores teores de argila. A unidade MU3 são as regiões de topos bem drenados onde encontramos áreas de encostas mais suaves e relevo relativamente plano a suave ondulado com boa drenagem, também com teores de argila elevados, compondo os Argissolos amarelos (Figura 28 C).

Algumas dessas covariáveis, também aparecem como importantes na predição da fração de textura do solo em outros trabalhos. Adhikari et al. (2013), destacaram os preditores slope e TWI em 80% de importância ao prever argila em superfície (0 a 30 cm) e as covariáveis TWI e MRVBF na predição de silte. No mapeamento de textura do solo em regiões áridas do Irã, Mehrabi-gohari et al. (2019) também destacaram o TWI como uma das covariáveis mais importantes na predição de argila e o TWI e MRVBF na predição de silte, considerando principalmente o TWI como uma variável preditiva significativa no modelo de predição da fração de silte.

Acredita-se que a bandaP foi selecionada como covariável importante no modelo RF de areia e argila em superfície, devido as características distintas dessas duas frações e pelo fato desse comprimento de onda ter a capacidade de penetrar no dossel da floresta, interagindo principalmente com a radiação eletromagnética da superfície do solo. O coeficiente de retroespalhamento da bandaP teve correlação positiva com a argila e negativa com a areia.

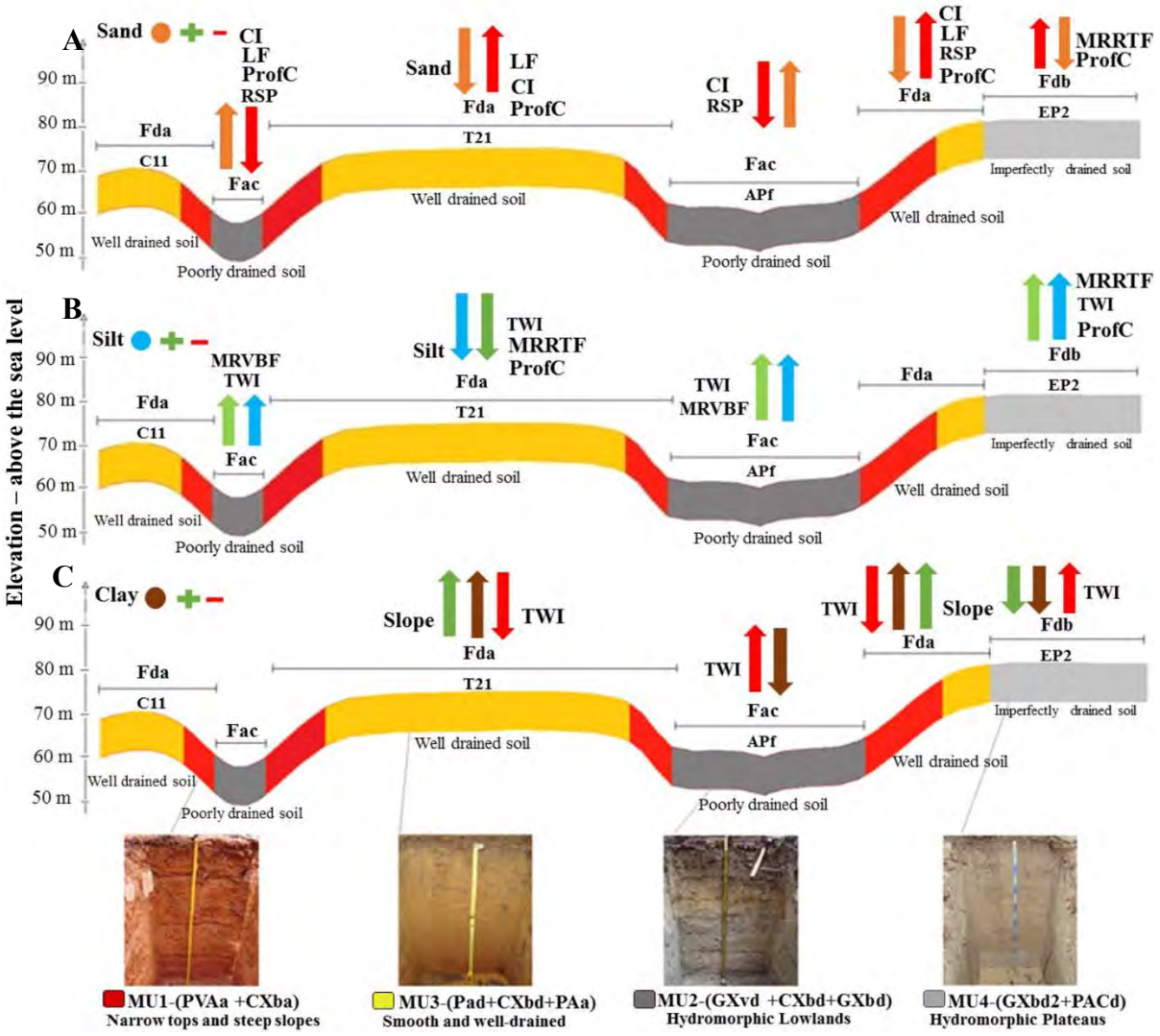


Figura 28. Relação solo- relevo- vegetação para areia (A), silte (B) e argila (C). Seta verde - correlação positiva com a covariável; seta vermelha- correlação negativa com a covariável. Fac- Floresta Tropical Aberta de Planície Inundada; Fda- Floresta Tropical Densa de Terras Altas; Fdb- Floresta Tropical Aberta de Planalto; APf- Planícies fluviais; C11- Áreas bem drenadas em topo plano; T21- Interflúvios Tabulares; EP2 - Superfícies biplanícies-planícies. H.S.—Sedimentos Holocênicos; P.S.—Sedimentos Pleistoceno. (Fonte: modificado CEDDIA et al. 2015).

4.5.3 Comparação de modelos preditivos

O desempenho dos modelos de AM é apresentado nas Tabelas 12, 13, 14 e 15. Nas tabelas 12, 13 e 14 os atributos são avaliados quanto ao conjunto de dados de AR e na Tabela 15 quanto ao conjunto de dados em AT. Em todas as tabelas, podemos avaliar como os algoritmos são afetados pelo tipo de conjuntos de dados utilizados (AR e AT), bem como a efeito dos métodos de seleção de covariáveis para etapa de modelagem dos algoritmos (PCS – Seleção Prévia de Covariáveis e WM - Wrapper Method). Para uma melhor visualização, também podemos observar o comportamento dos dados de treino e validação do algoritmo RF nos gráficos de dispersão dos atributos observado vs. estimado nas Figuras 29, 30, 31, 32 e 33.

Analisando areia em superfície, os melhores valores de predição se encontram no conjunto de dados de AR pelo algoritmo RF selecionando previamente as covariáveis (PCS). Os blocos Urucu e Juruá são as regiões com menores erros métricos e maiores R^2. Destaca-se que a região de Juruá é a área mais distante da área de referência e de mais difícil acesso, além disso, onde a estatística dos atributos é mais diferente. Logo, para estimar areia em subsuperfície a melhor predição foi observada no bloco de urucu (bloco que envolve a área de referência). Nesse caso o melhor resultado foi obtido usando conjunto de dados AR e deixando todas as covariáveis disponíveis para o algoritmo (WM). Os valores de R^2 variaram de 0,24 a 0,31 para os blocos de Urucu e Urucu + Juruá em superfície e 0,25 a 0,36 em subsuperfície. Os erros RMSE e MAE para areia em superfície variaram de 131,95g kg^{-1} a 143,59g kg^{-1} e 113,65g kg^{-1} a 120,15g kg^{-1}, respectivamente. Os erros RMSE e MAE para areia em subsuperfície variaram de 137,05g kg^{-1} a 158,11g kg^{-1} e 113,82g kg^{-1} a 134,07g kg^{-1}respectivamente (Tabela 12, Figuras 29 e 30).

O silte, foi o atributo que apresentou os melhores resultados utilizando o algortimo RF. Os valores de R^2 em superfície foram de 0,58, 0,37 e 0,52 e em subsuperfície 0,51, 0,41, 0,56 nos blocos de Urucu, Araracanga e Juruá respectivamente. Tanto em superfície quanto em subsuperfície, os melhores resultados foram encontrados utilizando a abordagem de AR - combinado com o método de seleção prévia de covariáveis (PCS). Os erros métricos em superfície variaram de 120,78 a 155,23 para o RMSE e 99,50 a 131,16 para o MAE, já em subsuperfície, 88,62g kg^{-1} a 90,04g kg^{-1} para RMSE e 71,49g kg^{-1} a 74,73g kg^{-1} para MAE (Tabela 13, Figuras 31 e 32).

A argila, foi a fração que apresentou o pior resultado, sendo melhor predita pela abordagem de AT com as covariáveis previamente selecionadas (PCS) tanto em superfície quanto em subsuperfície. Os valores de R^2, RMSE e MAE foram de 0,23, 64,79g kg^{-1}, 50,16g kg^{-1} e 0,31, 106,90g kg^{-1}, 81,17g kg^{-1}, em superfície e subsuperfície respectivamente (Tabela 15 e Figura 33).

Comparando os resultados de predição das frações areia, silte e argila deste estudo com a literatura, é possível observar que os valores de acurácia são razoáveis, diante das dificuldades impostas pela região e da baixa densidade amostral disponível (0,011 amostras/km^2).

Ließ et al. (2012) utilizando random forest, com densidade de amostragem de 1,87 perfis km^2 conseguiram explicar 30% da variação da areia e 43% da argila, na camada superficial dos solos, por meio do uso de atributos morfométricos. Bhering et al. (2016) utilizaram dois conjuntos de dados com resolução de modelo digital de 30 e 90m e explicaram 44 e 40% e 45 e 46%, na predição de areia e argila respectivamente. Os autores destacaram que a resolução espacial das covariáveis preditoras tiveram pouca influência sobre a predição dos atributos, e a abordagem por Random Forest apresentou potencial de utilização para estimar atributos do solo. Akpa et al. (2014), com densidade de amostragem de 0,001 perfis km^2 e utilizando random forest estimaram valores na camada superficial do solo de 0 a 15 cm de 16 a 53, 21 a 48 e 21 a 26% da variação nos teores de argila, areia e silte respectivamente. Vaysse & Lagacherie, (2015) com densidade de amostragem de 0,07 perfis km^2 e atributos morfométricos, dados

geológicos, climáticos e dados da imagem Landsat 7 explicaram 33 a 35% da variação de areia e 31 a 35% de argila. Chagas et al. (2016) na modelagem de areia, silte e argila com random forest e com uso de covariáveis ambientais (razões de bandas, NDVI, Clay minerals, dentre outros derivados de imagem Landsat 5) explicaram 63, 56 e 25% da variabilidade espacial de areia, argila e silte respectivamente.

De forma geral, os algoritmos RT e SVM não tiveram valores muito bons de R^2, diferentemente de Mehrabi-gohari et al. (2019), que usando RT obteve bons resultados no mapeamento de textura do solo em regiões áridas do Irã, sendo R^2 para a argila entre 0,68 e 0,51, para a areia 0,70 e 0,51 e silte 0,63-0,51. Pinheiro et al. (2018) também encontraram bom desempenho na aplicação do modelo de árvore de regressão em conjunto de dados harmonizados, com valores de 0,52 para argila na camada 0,00-0,05 m, e 0,69 para silte na camada de 0,05-0,15 camada m e R^2 maiores que 0,52 em areia. Kovacevic et al. (2010) utilizando algoritmo SVR obtiveram na predição de areia R^2 0,59 e argila R^2 0,76.

Dentre vários fatores, há um consenso na literatura que o fraco desempenho dos modelos geralmente é devido à baixa densidade de amostragem.

Tabela 12. Acurácia dos algoritmos de AM usando o conjunto de dados de área de referência (AR) Areia.

		RT			RF			SVM		
ATRIBUTOS	**DATA**	**R²**	**RMSE**	**MAE**	**R²**	**RMSE**	**MAE**	**R²**	**RMSE**	**MAE**
Areia Surf PCS	T	0,34	124,63	96,08	0,93	67,59	53,65	0,47	113,35	91,67
	VU	0,09	144,21	117,44	**0,24**	**131,95**	**113,65**	0,07	141,41	125,11
	VUA	0,03	165,44	129,18	0,19	140,60	116,81	0,01	162,06	140,10
	V	0,01	176,78	135,46	**0,24**	**143,59**	**120,15**	0,08	173,02	149,45
	VUJ	0,03	166,08	128,63	**0,31**	**138,09**	**119,00**	0,12	162,56	141,28
Areia Surf *WM*	T	0,36	122,20	96,81	0,94	67,14	53,93	0,57	106,18	86,47
	VU	0,21	132,60	103,71	0,20	132,26	114,97	0,04	144,79	129,51
	VUA	0,06	164,81	128,25	0,18	141,60	116,76	0,20	140,00	118,36
	V	0,03	175,96	132,01	0,19	148,08	123,59	0,19	145,37	123,52
	VUJ	0,09	157,57	113,77	0,22	143,72	125,04	0,08	151,27	134,70
Areia Sub PCS	T	0,45	114,43	90,63	0,92	63,85	50,63	0,47	113,51	95,48
	VU	0,09	161,85	137,87	0,24	147,39	126,58	0,20	148,15	126,38
	VUA	0,01	181,34	152,47	0,15	163,02	140,97	0,18	166,08	141,00
	V	0,00	190,00	155,52	0,11	165,91	143,28	0,24	168,50	139,67
	VUJ	0,02	179,18	145,01	0,17	154,95	132,64	0,21	155.50	127,30
Areia Sub *WM*	T	0,48	111,43	87,06	0,92	64,37	51,11	0,57	105,72	86,38
	VU	0,14	159,06	133,47	**0,36**	**137,05**	**113,82**	0,13	154,29	128,75
	VUA	0,05	181,84	155,73	**0,25**	**158,11**	**134,07**	0,15	173,11	146,64
	V	0,00	194,35	163,99	0,16	161,85	138,48	0,17	167,05	141,37
	VUJ	0,03	183,01	149,50	**0,25**	**147,25**	**123,99**	0,17	148,60	124,70

PCS - Seleção de Covariáveis Anteriores; WM - Método Wrapper; T: Conjunto de dados de treinamento; V: conjunto de dados de validação; VU: validação do bloco Urucu; VUA: validação do bloco Urucu/Araracanga; V: validação do bloco Urucu/Araracanga/Juruá; VUJ: validação do bloco Urucu/Juruá. RT: árvore de regressão; RF: floresta aleatória; SVM: máquina de vetores de suporte.

Tabela 13. Acurácia dos algoritmos de AM usando o conjunto de dados de área de referência (AR) para Silte.

		RT			RF			SVM		
ATRIBUTOS	**DATA**	**R²**	**RMSE**	**MAE**	**R²**	**RMSE**	**MAE**	**R²**	**RMSE**	**MAE**
Silte Surf PCS	T	0,49	93,12	72,31	0,91	56,79	43,59	0,50	93,10	71,32
	VU	0,19	163,88	138,90	**0,58**	**130,93**	**112,79**	0,33	130,12	107,02
	VUA	0,07	154,96	129,70	**0,37**	**120,78**	**99,50**	0,17	175,88	123,11
	V	0,07	175,06	144,43	0,36	140,78	114,20	0,28	185,24	133,30
	VUJ	0,18	189,07	158,10	**0,52**	**155,23**	**131,16**	0,38	156,24	124,62
Silte Surf *WM*	T	0,46	95,43	73,60	0,92	55,86	42,90	0,58	87,55	65,19
	VU	0,26	163,90	144,15	0,46	139,14	120,85	0,24	143,20	122,14
	VUA	0,06	157,48	136,16	0,26	128,44	106,70	0,13	149,53	119,28
	V	0,08	174,84	149,59	0,26	149,18	122,41	0,22	159,14	128,22
	VUJ	0,26	186,22	161,73	0,42	164,33	140,48	0,26	158,37	134,31
Silte Sub PCS	T	0,47	73,28	58,41	0,91	43,95	32,84	0,39	79,86	61,66
	VU	0,36	90,34	72,33	**0,51**	**89,04**	**71,49**	0,38	91,90	77,29
	VUA	0,38	86,52	72,32	**0,41**	**88,62**	**73,72**	0,33	111,76	91,59
	V	0,26	99,93	80,22	**0,46**	**89,39**	**74,73**	0,39	131,27	101,27
	VUJ	0,22	106,81	83,93	**0,56**	**90,04**	**73,35**	0,39	126,38	93,81
Silte Sub *WM*	T	0,49	72,17	57,35	0,92	43,83	32,80	0,53	72,01	56,70
	VU	0,35	89,71	72,40	0,42	93,02	74,63	0,42	84,22	67,87
	VUA	0,33	89,52	73,82	0,31	93,09	76,26	0,39	91,14	76,04
	V	0,22	101,84	81,57	0,37	94,71	78,93	0,39	115,79	89,47
	VUJ	0,21	106,75	83,70	0,50	95,33	78,74	0,37	120,78	88,56

PCS - Seleção de Covariáveis Anteriores; WM - Método Wrapper; T: Conjunto de dados de treinamento; V: conjunto de dados de validação; VU: validação do bloco Urucu; VUA: validação do bloco Urucu/Araracanga; V: validação do bloco Urucu/Araracanga/Juruá; VUJ: validação do bloco Urucu/Juruá. RT: árvore de regressão; RF: floresta aleatória; SVM: máquina de vetores de suporte.

Tabela 14. Acurácia dos algoritmos de AM usando o conjunto de dados de área de referência (AR) para Argila.

		RT			RF			SVM		
ATRIBUTOS	**DATA**	**R²**	**RMSE**	**MAE**	**R²**	**RMSE**	**MAE**	**R²**	**RMSE**	**MAE**
	T	0,53	55,62	41,81	0,91	31,00	23,39	0,47	61,33	45,55
	VU	0,09	73,66	59,19	0,24	71,52	59,91	0,21	67,02	53,50
Argila Surf PCS	VUA	0,04	90,00	70,97	0,02	92,19	72,47	0,08	115,35	73,04
	V	0,03	90,96	73,81	0,02	92,45	73,84	0,04	111,00	73,09
	VUJ	0,06	78,66	65,49	0,19	76,38	64,28	0,17	69,17	57,33
	T	0,54	54,99	40,35	0,92	31,15	23,51	0,56	56,22	41,10
	VU	0,08	74,09	59,95	0,18	71,21	59,87	0,27	65,19	50,25
Argila Surf *WM*	VUA	0,04	89,95	70,40	0,02	91,54	71,99	0,17	82,31	61,32
	V	0,03	90,92	73,31	0,02	91,24	72,95	0,10	96,26	72,56
	VUJ	0,05	78,99	66,10	0,15	75,02	63,57	0,15	91,00	68,36
	T	0,61	67,08	53,46	0,91	39,92	30,50	0,58	70,11	52,14
	VU	0,16	119,24	90,48	0,20	114,50	86,75	0,14	120,00	93,82
Argila Sub PCS	VUA	0,02	136,87	101,37	0,08	122,28	95,36	0,17	117,28	95,30
	V	0,02	130,98	93,54	0,07	116,48	89,68	0,13	113,43	92,55
	VUJ	0,15	113,33	81,42	0,18	107,22	80,31	0,12	114,06	90,20
	T	0,62	65,74	52,30	0,92	38,92	29,74	0,65	65,90	49,20
	VU	0,02	138,97	103,56	0,18	115,89	87,83	0,07	128,03	99,55
Argila Sub *WM*	VUA	0,00	146,94	111,05	0,08	120,79	93,80	0,14	118,69	93,12
	V	0,00	140,81	104,34	0,07	114,86	88,35	0,03	152,95	116,13
	VUJ	0,03	131,32	95,44	0,17	108,00	81,22	0,02	170,86	131,05

PCS - Seleção de Covariáveis Anteriores; WM - Método Wrapper; T: Conjunto de dados de treinamento; V: conjunto de dados de validação; VU: validação do bloco Urucu; VUA: validação do bloco Urucu/Araracanga; V: validação do bloco Urucu/Araracanga/Juruá; VUJ: validação do bloco Urucu/Juruá. RT: árvore de regressão; RF: floresta aleatória; SVM: máquina de vetores de suporte.

Tabela 15. Acurácia dos algoritmos de AM usando o conjunto de dados Área Total (AT) para areia, silte e argila.

		RT			RF			SVM		
ATTRIBUTES	**DATA**	**R²**	**RMSE**	**MAE**	**R²**	**RMSE**	**MAE**	**R²**	**RMSE**	**MAE**
Areia Surf PCS	T114	0,51	104,54	79,07	0,93	62,17	49,06	0,52	105,63	84,05
	V37	0,00	198,21	152,76	0.11	161,29	124,46	0,15	163,40	127,98
Areia Surf *WM*	T114	0,51	104,54	79,07	0,94	64,03	50,25	0,77	73,35	44,26
	V37	0,00	198,21	152,76	0,13	159,76	124,81	0,03	209,14	158,38
Areia Sub PCS	T114	0,54	97,51	80,50	0,93	58,48	47,98	0,40	113,73	94,53
	V37	0,03	201,98	148,69	0,23	174,62	137,93	0,21	180,93	138,08
Areia Sub *WM*	T114	0,55	97,35	80,07	0,95	59,01	48,27	0,81	64,91	41,40
	V37	0,03	202,34	148,76	0,22	177,14	145,21	0,19	207,12	162,41
Silte Surf PCS	T114	0,58	89,40	69,27	0,91	53,32	41,17	0,50	98,00	78,58
	V37	0,04	181,86	140,19	0,14	147,62	113,70	0,20	142,71	108,66
Silte Surf *WM*	T114	0,92	54,47	42,50	0,92	54,47	42,50	0,60	89,56	72,23
	V37	0,17	144,92	111,63	0,17	144,92	111,63	0,14	147,47	112,04
Silte Sub PCS	T114	0,49	71,55	57,04	0,91	39,51	31,47	0,42	76,60	61,36
	V37	0,06	123,68	97,63	0,03	120,87	98,83	0,29	102,91	79,98
Silte Sub *WM*	T114	0,51	69,49	55,04	0,92	38,62	30,06	0,55	69,35	54,23
	V37	0,04	126,21	99,85	0,06	116,38	94,03	0,21	107,44	84,62
Argila Surf PCS	T114	0,56	58,72	44,86	0,91	34,38	25,29	0,59	60,07	46,72
	V37	0,23	71,54	58,13	**0,23**	**64,79**	**50,16**	0,15	70,18	52,70
Argila Surf *WM*	T114	0,58	57,46	43,43	0,92	33,47	25,53	0,65	56,09	42,48
	V37	0,20	74,62	62,48	0,21	65,24	48,36	0,12	80,70	62,14
Argila Sub PCS	T114	0,54	70,82	55,10	0,93	38,97	30,26	0,57	70,58	56,11
	V37	0,19	117,26	94,70	**0,31**	**106,90**	**81,17**	0,29	114,94	92,21
Argila Sub *WM*	T114	0,51	73,71	58,79	0,93	39,40	30,31	0,61	68,41	53,36
	V37	0,21	116,46	93,56	0,30	107,88	82,67	0,26	122,35	94,83

PCS - Seleção de Covariáveis Anteriores; WM - Método Wrapper; T: Conjunto de dados de treinamento; V: conjunto de dados de validação; RT: árvore de regressão; RF: floresta aleatória; SVM: máquina de vetores de suporte.

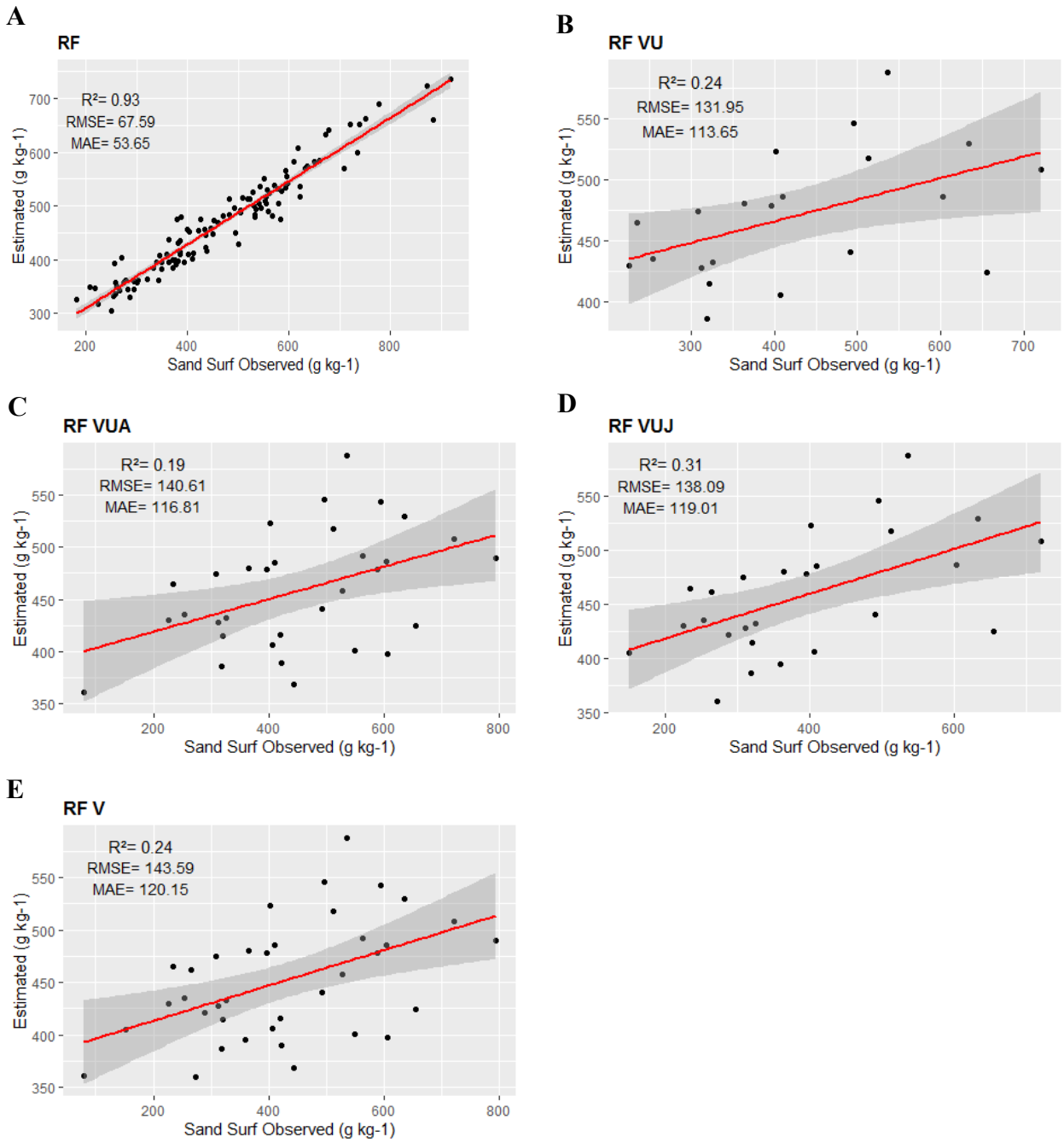


Figura 29. Ajuste linear do modelo de RF para dados de treinamento e validações de Areia A (em superfície na RA). (a) Ajuste linear dos dados de treinamento do modelo de RF; (b) Ajuste linear do modelo de RF para dados de validação de Urucu; (c) Ajuste linear do modelo de RF para dados de validação de Urucu/ Araracanga (d) Ajuste linear do modelo de RF para dados de validação de Urucu/Juruá; (e) Ajuste linear do modelo de RF para dados de validação de Urucu/Araracanga/Juruá. (imagens geradas no programa RStudio).

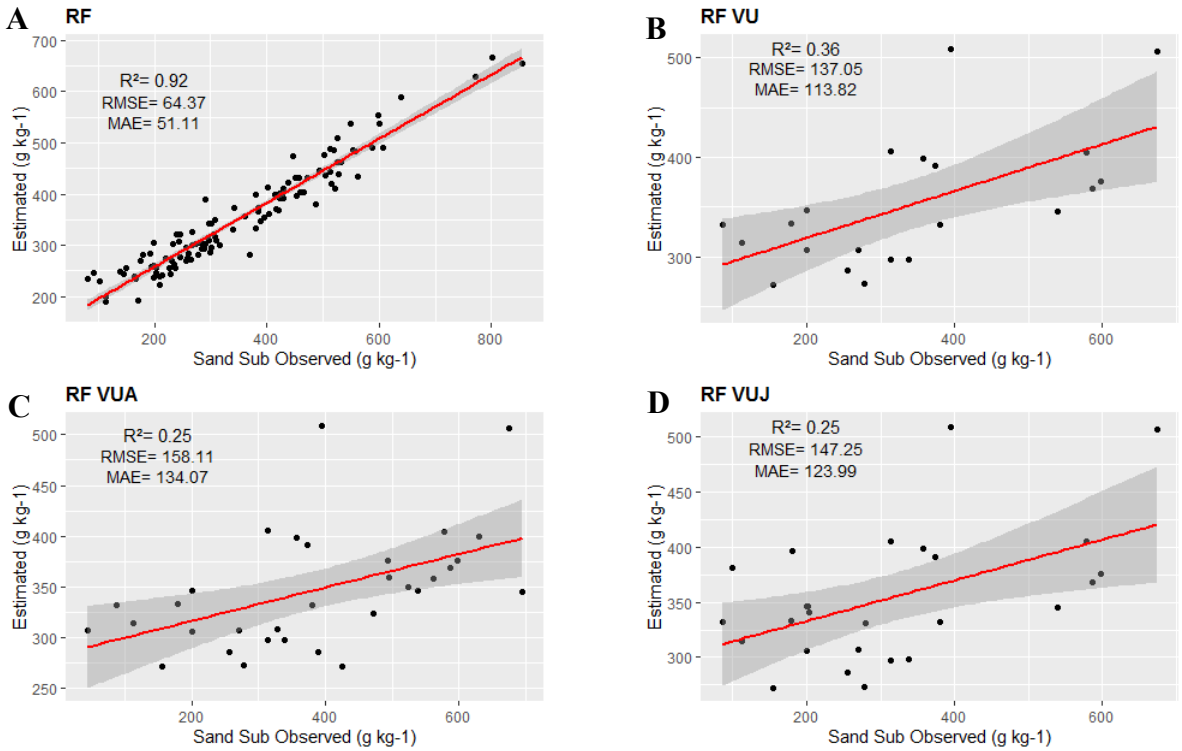


Figura 30. Ajuste linear do modelo de RF para dados de treinamento e validações de Areia B (em subsuperfície) na RA. (a) Ajuste linear dos dados de treinamento do modelo de RF; (b) Ajuste linear do modelo de RF para dados de validação de Urucu; (c) Ajuste linear do modelo de RF para dados de validação de Urucu/Araracanga (d) Ajuste linear do modelo de RF para dados de validação de Urucu/Juruá. (imagens geradas no programa RStudio).

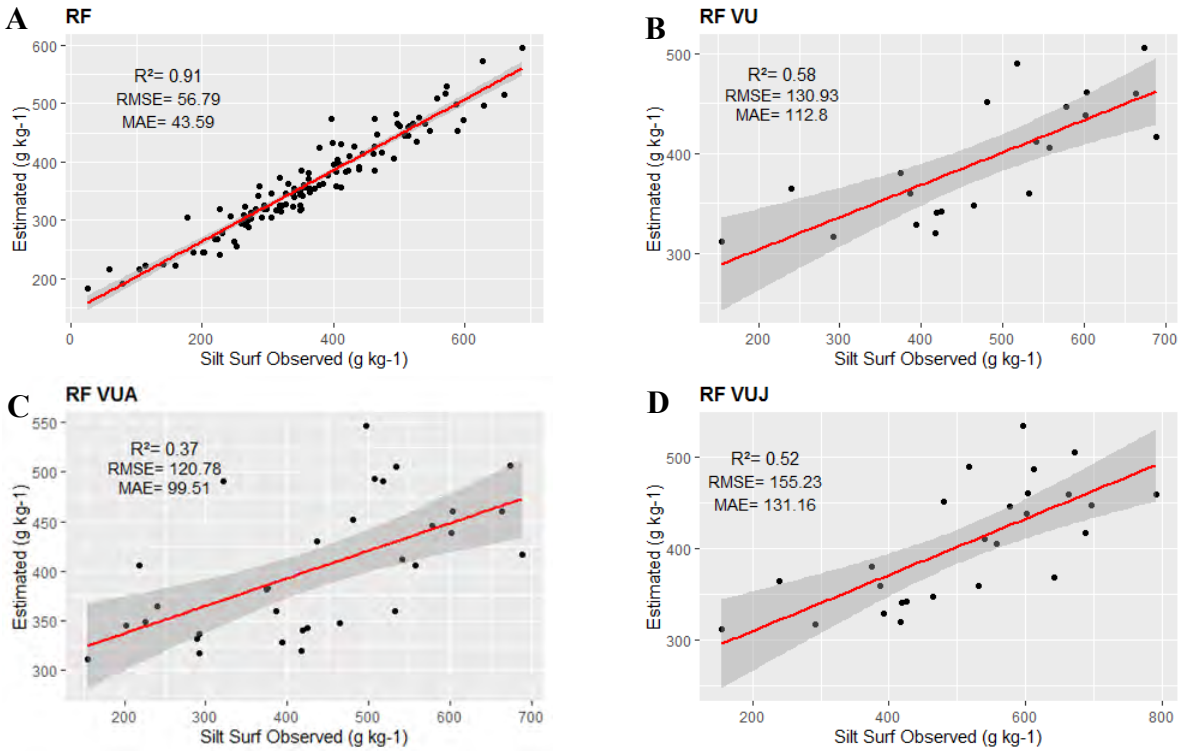


Figura 31. Ajuste linear do modelo de RF para dados de treinamento e validações de Silte A (em superfície) na RA. (a) Ajuste linear dos dados de treinamento do modelo de RF; (b) Ajuste linear do modelo de RF para dados de validação de Urucu; (c) Ajuste linear do modelo de RF para dados de validação de Urucu/Araracanga (d) Ajuste linear do modelo de RF para dados de validação de Urucu/Juruá. (imagens geradas no programa RStudio).

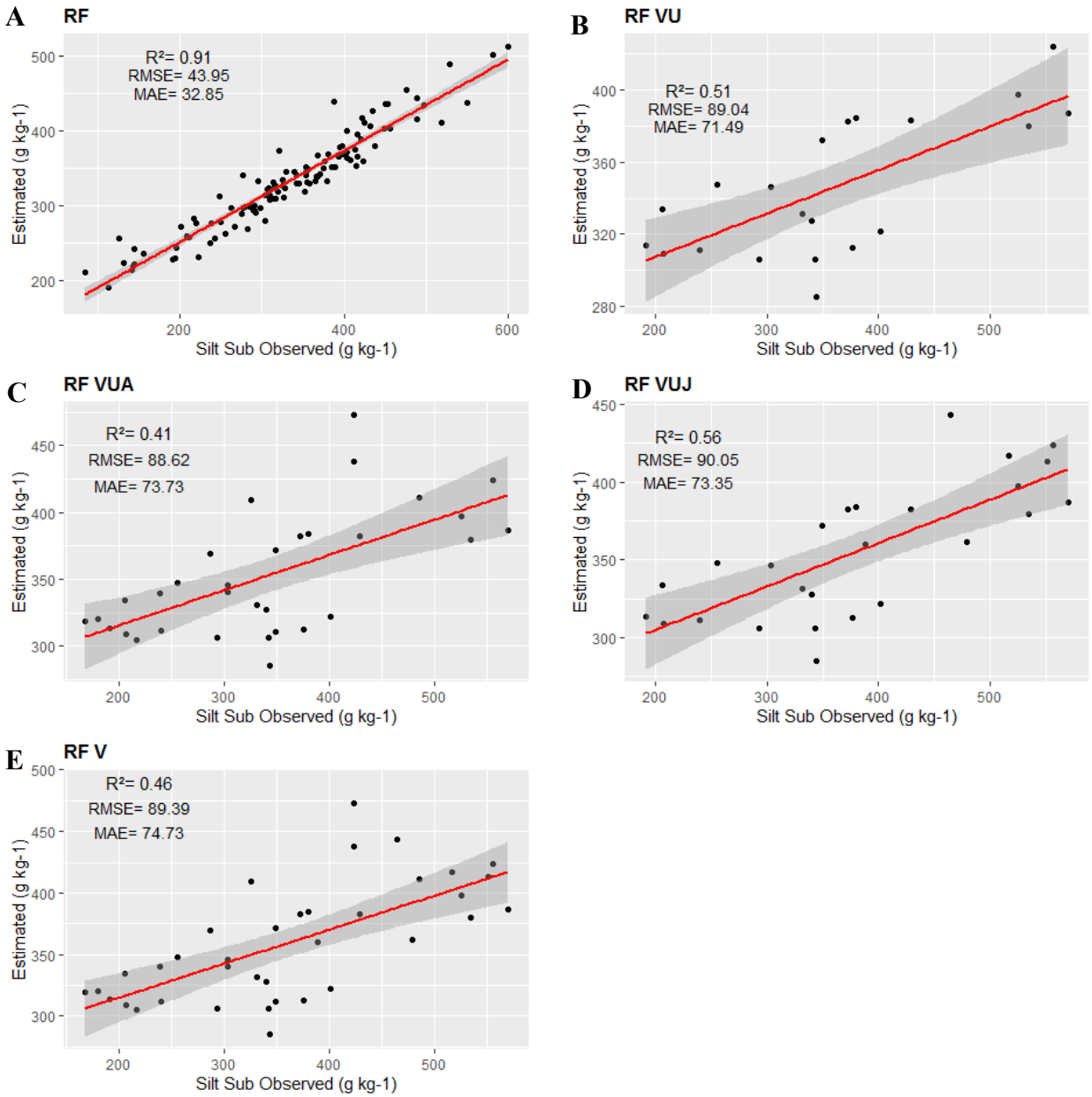


Figura 32. Ajuste linear do modelo de RF para dados de treinamento e validações de Silte B (em subsuperfície) na RA. (a) Ajuste linear dos dados de treinamento do modelo de RF; (b) Ajuste linear do modelo de RF para dados de validação de Urucu; (c) Ajuste linear do modelo de RF para dados de validação de Urucu/ Araracanga (d) Ajuste linear do modelo de RF para dados de validação de Urucu/Juruá; (e) Ajuste linear do modelo de RF para dados de validação de Urucu/Araracanga/Juruá. (imagens geradas no programa RStudio).

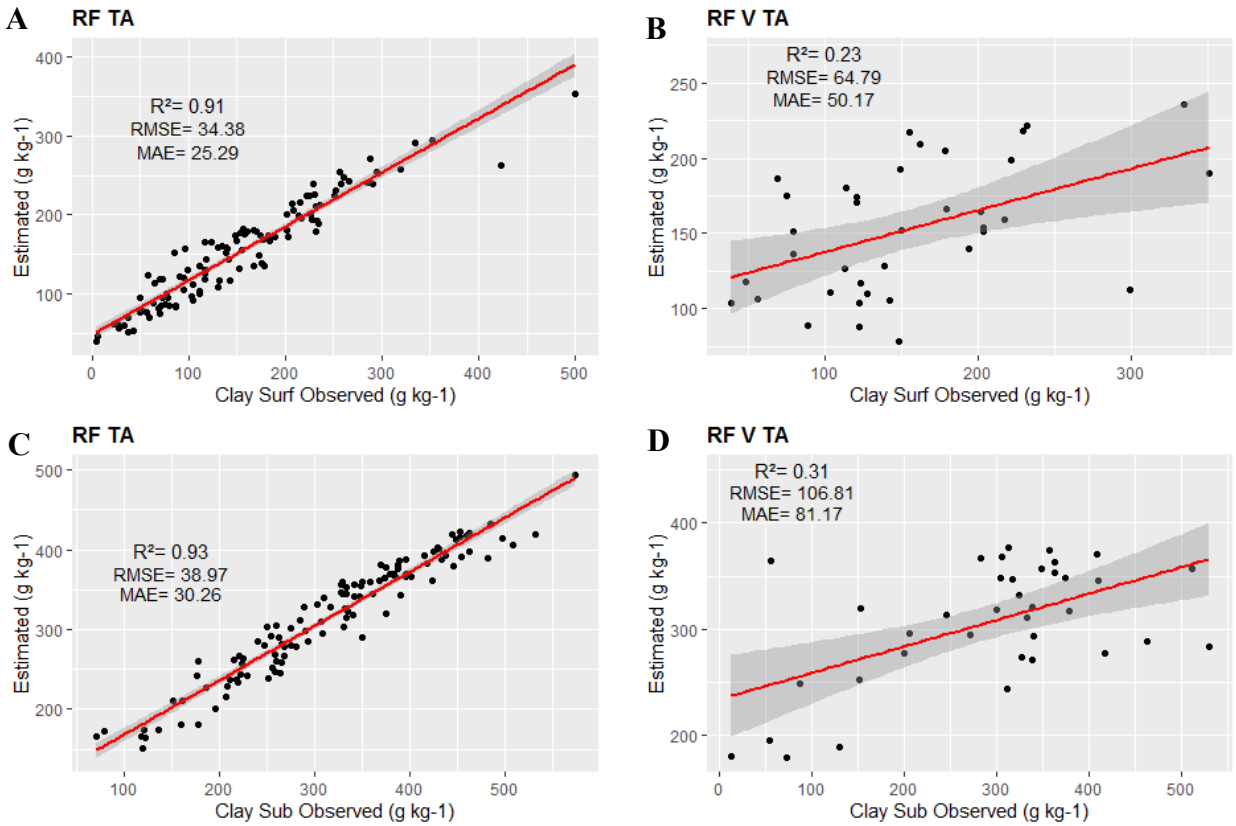


Figura 33. Ajuste linear do modelo de RF para dados de treinamento e validações de Argila A (em superfície) e Argila B (em subsuperfície) na TA. (a) Ajuste linear dos dados de treinamento do modelo de RF ArgilaA; (b) Ajuste linear do modelo de RF para dados de validação área total ArgilaA: (c) Ajuste linear dos dados de treinamento do modelo de RF ArgilaB; (d) Ajuste linear do modelo de RF para dados de validação área total ArgilaB. (imagens geradas no programa RStudio).

4.5.4 Predição espacial de areia, silte e argila

Os mapas finais preditos dos atributos do solo foi utilizando o algoritmo RF e são apresentados nas Figuras 34, 35, 36, 37, 38 e 39.

Os menores valores de areia, são preditos nos topos planos hidromórficos e as áreas com declividades mais acentuadas. Os maiores teores de areia estão presentes nas regiões de várzea, próximas as calhas dos grandes rios e igarapés e terraços no entorno do curso d'água principal, (vales em formato U), bem como nos vales mais encaixados (vales em formato V) das regiões de encostas. Esses ambientes são caracterizados pelo acúmulo de sedimentos arenosos proveniente dos processos erosivos naturais, tornando as baixadas entulhadas. Nessas regiões, os solos predominantes são classificados como Gleissolos Háplicos e Cambissolos Flúvicos e Háplicos (MU2). Em toda a área mapeada os teores de areia variaram de 302,79 a 720,60 g kg^{-1} para areia em superfície (Figura 34) e 211,56 a 634,59 g kg^{-1} para areia em subsuperfície (Figura 35).

A fração de silte (Figura 36 e 37) é o atributo que mais expressa o padrão na paisagem, os maiores teores são encontrados nas regiões de topos planos hidromórficos. Ocorrem, geralmente, nas maiores elevações da área de estudo, associados aos divisores das bacias hidrográficas, onde o relevo plano e a drenagem insuficiente, caracterizam essas regiões representadas pela unidade MU4 onde há o predomínio dos Argissolos acinzentados e

Gleissolos Háplicos. Nas regiões de baixada também são encontrados valores relevantes, onde encontram-se Gleissolos Háplicos (MU2).

Nas Figuras 38 e 39, observa-se que os maiores teores de argila ocorrem nas áreas de encostas declivosas e topos bem drenados. Esses teores aumentam em subsuperfície, variando de 154,22 a 458,17 g kg^{-1} (Figura 39). O sensível aumento de argila em profundidade é coerente com a observação de ocorrência de solos da ordem Argissolos, os quais apresentam horizonte diagnóstico B textural (Bt). Essas regiões são representadas pelas unidades de mapeamento MU1 e MU3 onde há predomínio dos solos Argissolos vermelho amarelo e Argissolos amarelo.

Os resultados espaciais das frações de textura encontrados nesse estudo corroboram com os resultados de Ceddia et al., (2017) realizados na mesma região de estudo.

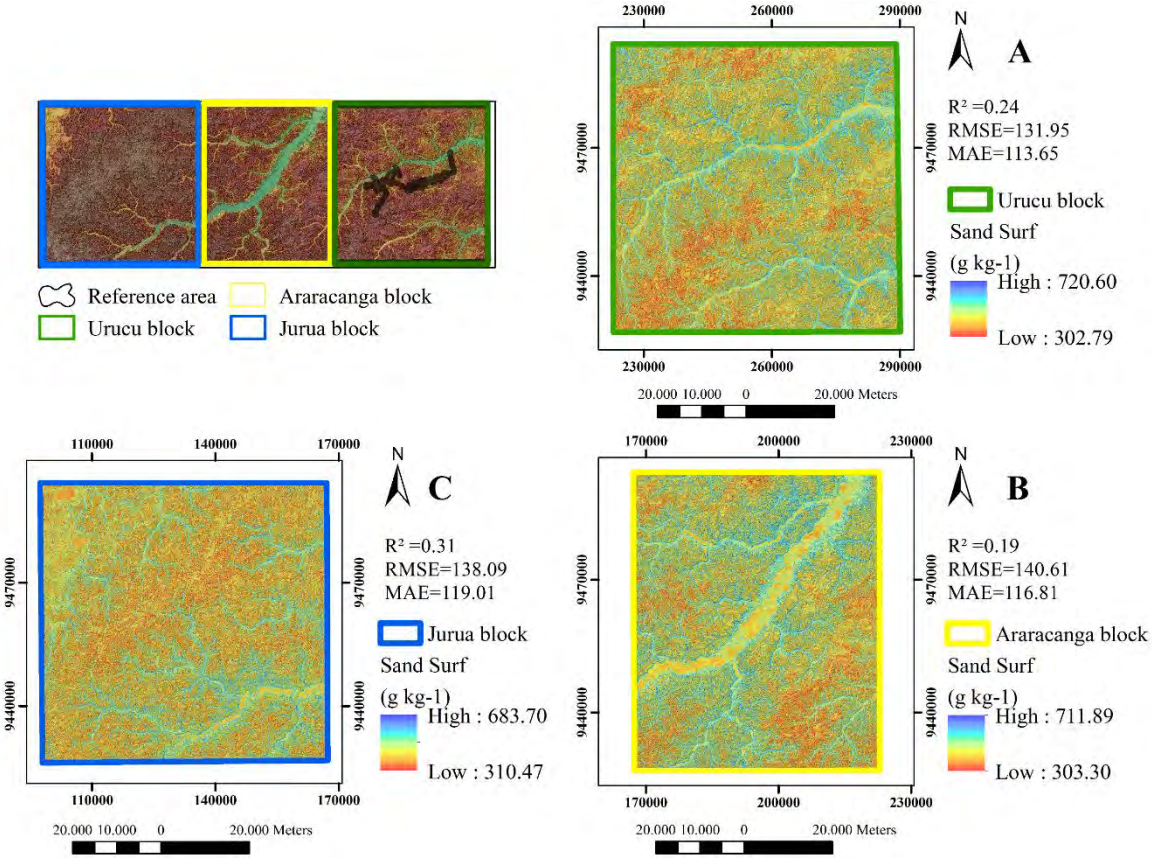


Figura 34. Predição espacial de areia em superfície. (A) bloco Urucu, (B) bloco Araracanga e (C) bloco Juruá.

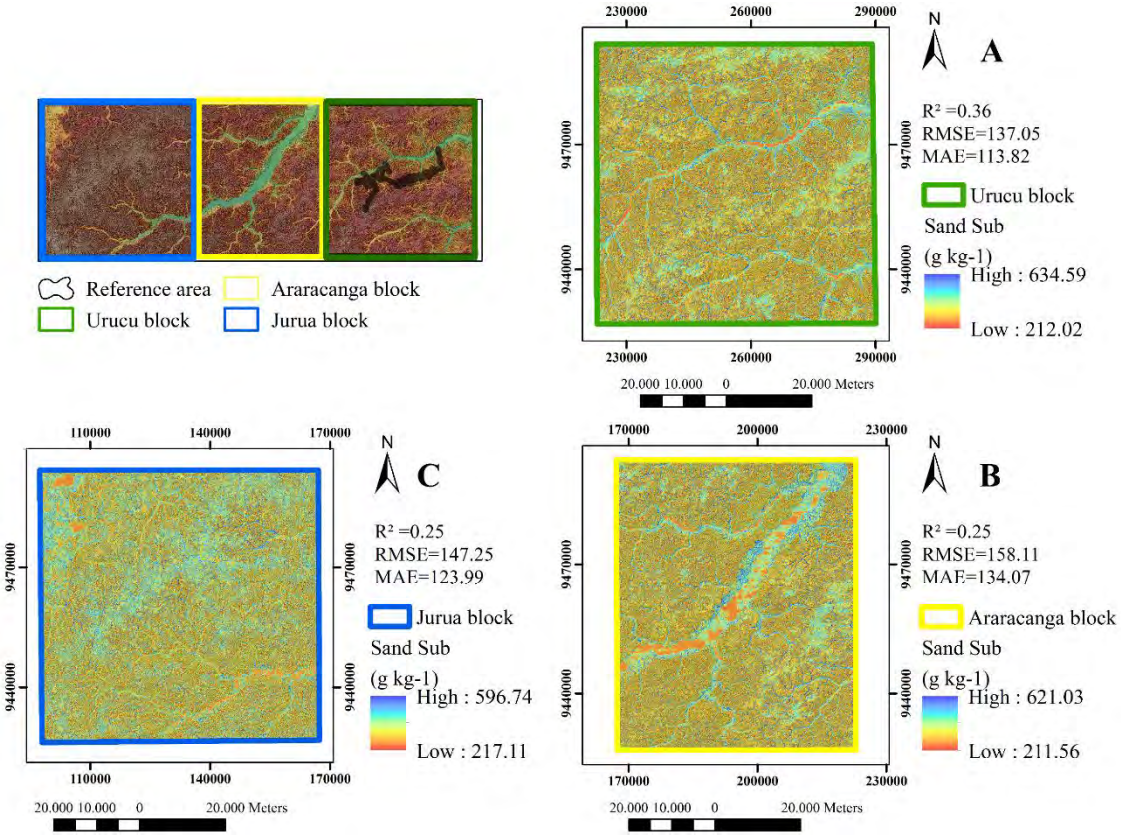


Figura 35. Predição espacial de areia em subsuperfície. A) bloco Urucu, (B) bloco Araracanga e (C) bloco Juruá.

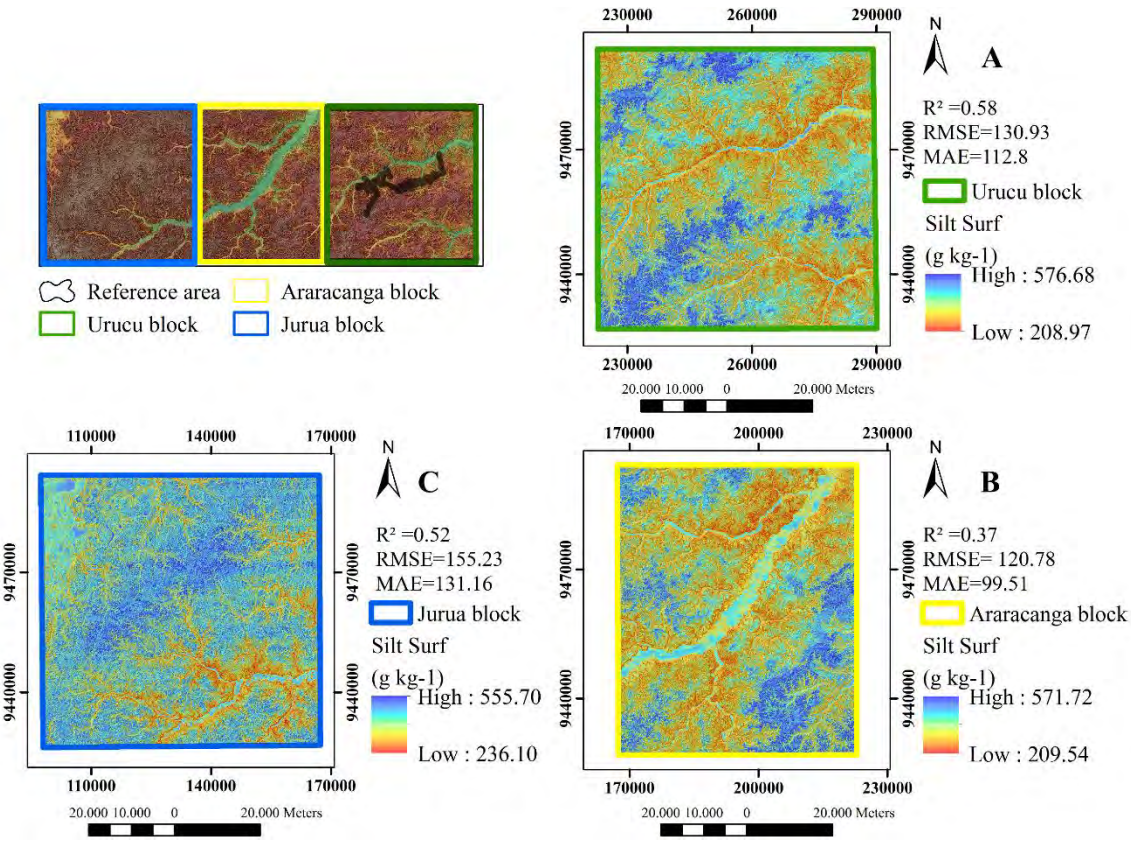


Figura 36. Predição espacial de silte em superfície. A) bloco Urucu, (B) bloco Araracanga e (C) bloco Juruá.

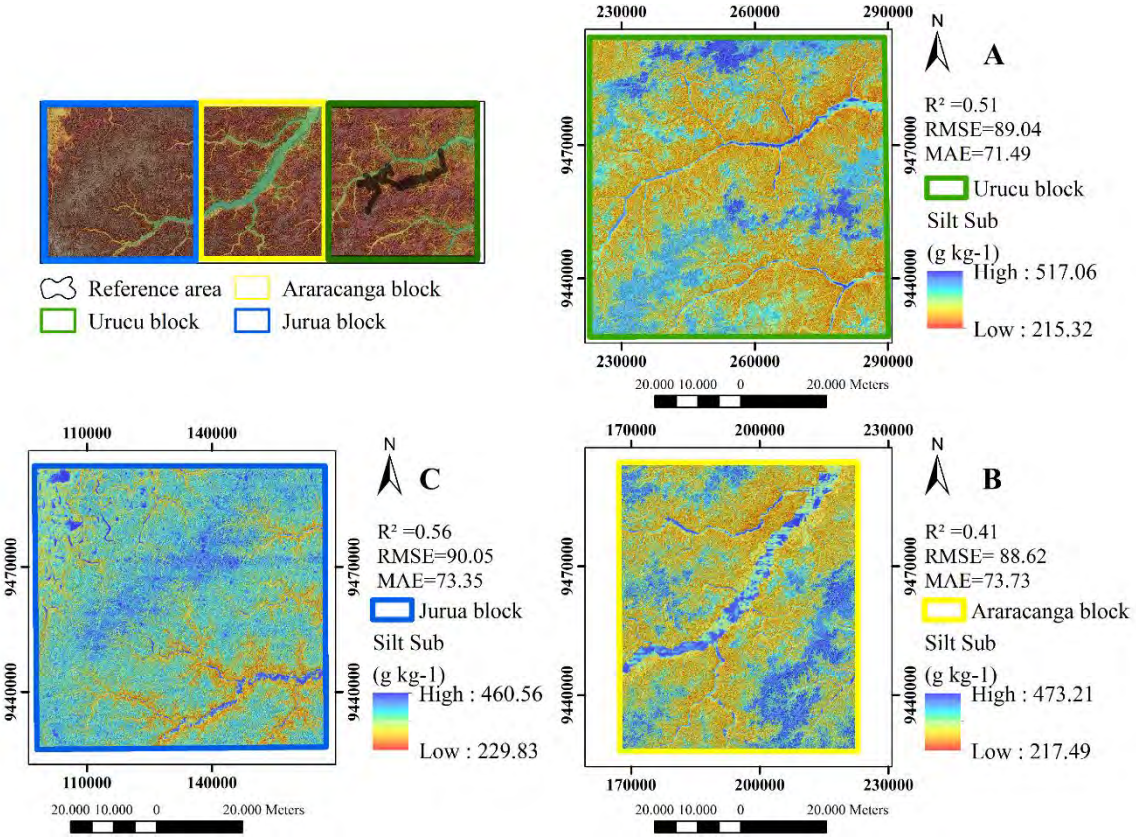


Figura 37. Predição espacial de silte em subsuperfície. A) bloco Urucu, (B) bloco Araracanga e (C) bloco Juruá.

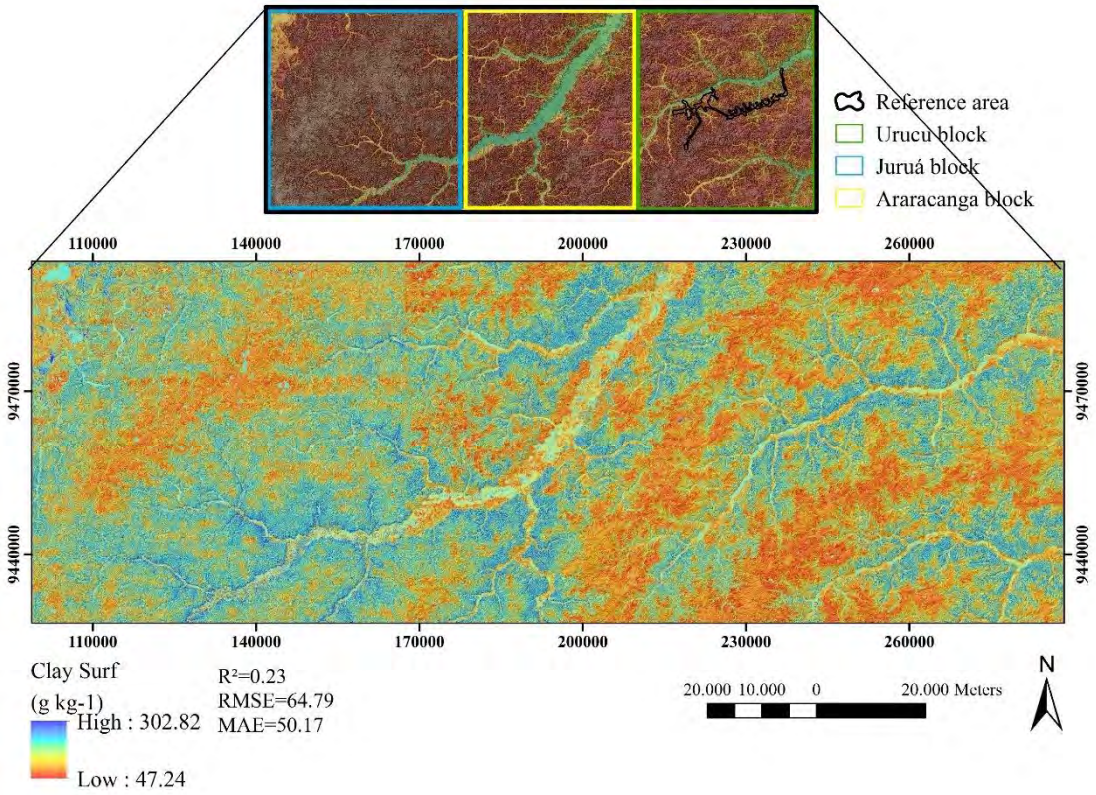


Figura 38. Predição espacial de argila em superfície para área total.

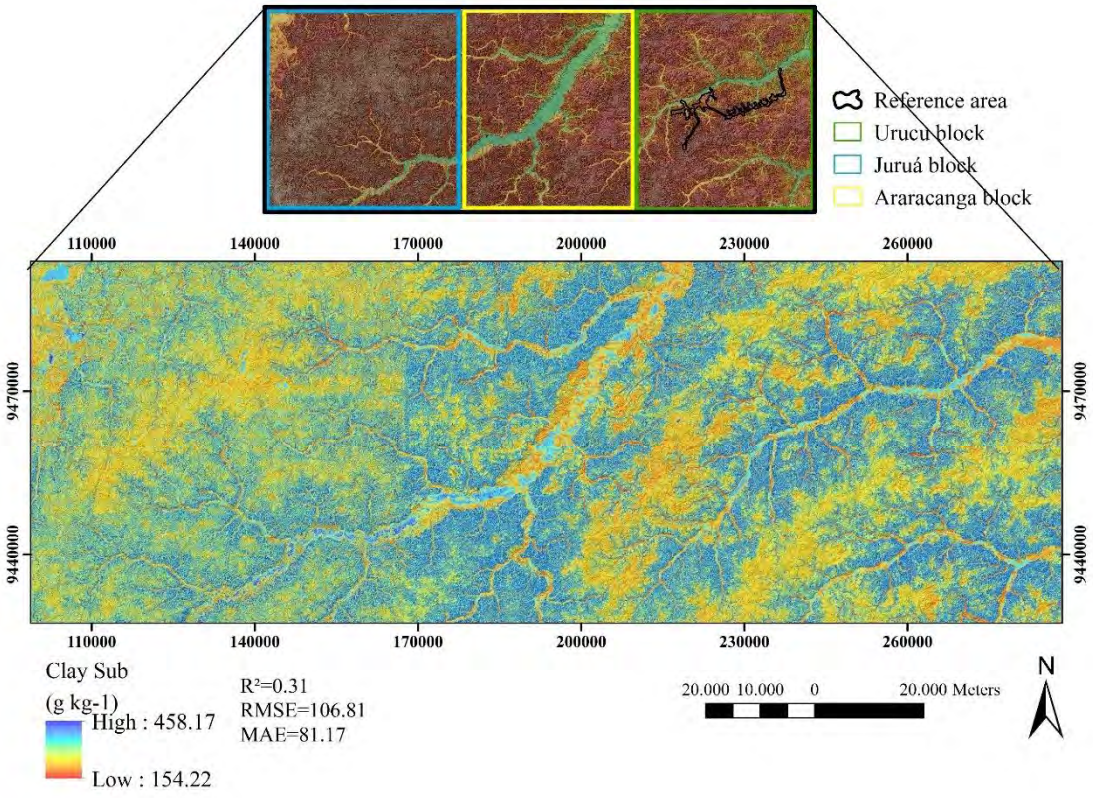


Figura 39. Predição espacial de argila em subsuperfície para área total.

4.6 CONCLUSÕES

Apesar da limitação de pontos de observação da composição granulométrica do solo disponíveis (0,01 amostras / km²), além da distribuição espacial irregular ao longo da área remota na região, foi possível gerar mapas com boas métricas de textura do solo em superfície e subsuperfície utilizando algoritmos de AM associados à covariáveis de relevo e sensoriamento remoto.

No geral, com exceção da argila, o conjunto de dados baseado no uso da área de referência (AR) obteve os melhores resultados de predição.

A partir dos atributos morfométricos derivados do MDE, foi possível estabelecer relações entre as frações granulométricas e a paisagem. O método de seleção de covariáveis (PCS), obteve no geral, melhores resultados no mapeamento da composição granulométrica do solo. As covariáveis que exerceram maior influência na distribuição da composição granulométrica do solo na região foram: CI, LF, MRRTF, MRVBF, TWI, SLOPE, ProfC, para areia, silte e argila, além da bandaP, particularmente no caso das frações de areia e argila.

O coeficiente de retroespalhamento da banda P foi considerado como covariável importante para predição de areia e argila em superfície pelo modelo de random forest, mostrando seu uso potencial na modelagem desses atributos.

Com base nas análises realizadas, o algoritmo Random Forest apresentou desempenho geral superior na predição da composição granulométrica do solo. A melhor predição foi estimada para silte em superfície e em subsuperfície para as regiões dos blocos de Urucu e Juruá.

5. CONCLUSÕES GERAIS

Este trabalho investigou o uso de covariáveis ambientais de fontes como modelos de elevação e imagens de radar, e o conceito de área de referência, para mapear os atributos do solo (estoque de carbono e composição granulométrica), e entender melhor a relação solo-paisagem em uma região remota de exploração de Petróleo e Gás na Amazônia Central.

Os resultados obtidos nesse estudo dão suporte à hipótese de que o mapeamento digital de atributos do solo, a partir de uma área de referência, associada as técnicas de AM, pode ser utilizado como alternativa no mapeamento para dimensões de áreas mais extensas da paisagem.

Com base nas análises realizadas, o modelo Random Forest apresentou desempenho superior para todos os atributos preditos. No geral, os melhores ajustes e menores erros métricos foram feitos para estoque de carbono a 100cm de profundidade (SOC100) e para silte em superfície e subsuperfície.

Os resultados mostram, não só a importância dos algoritmos de AM para mapear os atributos do solo, como também do uso de conhecimento pedológico especializado gerado em uma AR, para apoiar a seleção de covariáveis ambientais, antes de calibrar os algoritmos de AM. Os algoritmos de AM testados e os métodos de seleção de covariáveis mostraram que mesmo com poucos dados e uma área muito grande e remota, bons resultados podem ser obtidos.

Sugere-se, como forma de comparação e potencialidade dos dados, a investigação de novas técnicas de aprendizado de máquina, como, por exemplo, árvores de regressão impulsionadas (BRT), gradient boosting, ou Deep learning, caso a quantidade de dados aumente significativamente. Assim, poderá verificar se a estimativa dos atributos do solo sofre diferenciação ao ser modelada por diferentes algoritmos.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ADHIKARI, K.; KHEIR, R. B.; GREVE, M. B.; BØCHER, P. K.; MALONE, B. P.; MINASNY, B.; MCBRATNEY, A. B.; GREVE, M. H. High-Resolution 3-D Mapping of Soil Texture in Denmark. **Soil Science Society of America Journal** v. 77, n. 3, p. 860–876, 2013.

ADHIKARI, K. Digital Mapping of Topsoil Carbon Content and Changes in the Driftless Area of Wisconsin, USA. **Soil Science Society of America Journal**, v. 79, 2 dez. 2014.

AKPA, S. I. C.; ODEH, I. O. A.; BISHOP, T. F. A.; HARTEMINK, A. Digital Mapping of Soil Particle-Size Fractions for Nigeria. **Soil Science Society of America Journal**, v. 78, n. 6, p. 1953–1966, 2014.

ARRUDA, G. P.; DEMATTÊ, J.A.M.; CHAGAS, C. S.; FIORIO, P.R.; SOUZA, A.B.; FONGARO, C.T. Digital soil mapping using reference area and artificial neural networks. **Scientia Agricola**, v. 73, n. 3, p. 266–273, jun. 2016.

ARRUDA, G. P. DE; DEMATTÊ, J. A. M.; CHAGAS, C. S. Mapeamento digital de solos por redes neurais artificiais com base na relação solo-paisagem. **Revista Brasileira de Ciência do Solo**, v. 37, n. 2, p. 327–338, abr. 2013.

ARUN, K.; LANGMEAD, C. J. Structure based chemical shift prediction using random forests non-linear regression. Em: **Proceedings of the 4th Asia-Pacific Bioinformatics Conference**. Series on Advances in Bioinformatics and Computational Biology. [s.l.] Published by imperial college press and distributed by world scientific publishing co., 2005. v. Volume 3p. 317–326.

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, set. 2016.

BATJES, N. H.; DIJKSHOORN, J. A. Carbon and nitrogen stocks in the soils of the Amazon Region. **Geoderma**, v. 89, n. 3–4, p. 273–286, maio 1999.

BERNOUX, M.; CARVALHO, M. C. S.; VOLKOFF, B.; CERRI, C. C. Brazil's Soil Carbon Stocks. **Soil Science Society of America Journal**, v. 66, n. 3, p. 888–896, 2002.

BHERING, S. B.; CHAGAS, C. S.; CARVALHO JUNIOR, W.; PEREIRA, N. R.; CALDERANO FILHO, B.; PINHEIRO, H. S. K. Mapeamento digital de areia, argila e carbono orgânico por modelos Random Forest sob diferentes resoluções espaciais. **Pesquisa Agropecuária Brasileira**, v. 51, n. 9, p. 1359–1370, set. 2016.

BLUMBERG, D. G.; FREILIKHER, V.; KAGANOVSKII, Y.; MARADUDIN, A. A. Subsurface microwave remote sensing of soil-water content: Field studies in the Negev Desert and optical modelling. **International Journal of Remote Sensing**, v. 23, n. 19, p. 4039–4054, jan. 2002.

BOEHNER, J., KOETHE, R. CONRAD, O., GROSS, J., RINGELER, A., SELIGE, T. Soil Regionalisation by Means of Terrain Analysis and Process Parameterisation. In: MICHELI, E., NACHTERGAELE, F., MONTANARELLA, L. [Ed.]: Soil Classification 2001. **European Soil Bureau, Research Report** No. 7, EUR 20398 EN, Luxembourg. 2002. 213-222p.

BOEHNER, J., SELIGE, T. Spatial prediction of soil attributes using terrain analysis and climate regionalization. In: BÖHNER, J., MCCLOY, K.R., STROBL, J. [Eds.] **SAGA - Analyses and Modelling Applications**. Göttinger Geogr. 2006. 115. 13-27p.

BOEHNER, J., CONRAD, O. Module relative heights and slope positions. **System for Automated Geoscientific Analyses. SAGA**. 2008.

BÖHNER, J.; ANTONIĆ, O. Chapter 8 Land-Surface Parameters Specific to Topo-Climatology. Em: HENGL, T.; REUTER, H. I. (Eds.). . **Developments in Soil Science**. Geomorphometry. [s.l.] Elsevier, 2009. v. 33p. 195–226.

BRASIL. **Departamento Nacional de Produção Mineral. Projeto RADAMBRASIL. 1973-1987. (Levantamento de Recursos Naturais, 38 volumes).**, 5 nov. 2019. (Nota técnica).

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, R. A. Classification and Regression Trees (CART)**. Belmont, CA: Wadsworth International,** 1984**.**

BREIMAN, L. Random forests**. Mach. Learn. 45, 5–32, 2001.**, 2001.

CARVALHO JÚNIOR, W. de. **Classificação supervisionada de pedopaisagens no domínio dos mares de morros utilizando redes neurais artificiais**. Tese de doutorado. UFV, MG. 2005. 160p

CEDDIA, M. B.; VILLELA, A. L. O.; PINHEIRO, É.F. M.; WENDROTH, O. Spatial variability of soil carbon stock in the Urucu river basin, Central Amazon-Brazil. **Science of The Total Environment**, v. 526, p. 58–69, 1 set. 2015.

CEDDIA, M. B.; GOMES, A. S.; VASQUES, G. M.; PINHEIRO, É. F. M. Soil Carbon Stock and Particle Size Fractions in the Central Amazon Predicted from Remotely Sensed Relief, Multispectral and Radar Data. **Remote Sensing**, v. 9, n. 2, p. 124, fev. 2017.

CHAGAS, C. S.; CARVALHO JUNIOR, W.; BHERING, S. B.; CALDERANO FILHO, B. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. **Catena**, 2016a.

CHAGAS, C. S.; BHERING, S. B.; CARVALHO JUNIOR, W.; PEREIRA, N. R.. Mapeamento digital de atributos físicos e físico-hídricos do solo por técnicas de mineração de dados. 2016b.

COMLEY, J. W.; DOWE, D. L. Minimum message length and generalized Bayesian nets with asymmetric languages. **Advances in Minimum Description Length Theory and Applications**, p. 265–294, 2005.

CONRAD, O. Module Valley Depth. System for Automated Geoscientific Analyses. SAGA. 2012.

CRIVELENTI, R. C.; COELHO, R. M.; ADAMI, S. F.; OLIVEIRA, S. R. M. O. Mineração de dados para inferência de relações solo-paisagem em mapeamentos digitais de solo. **Pesquisa Agropecuária Brasileira**, v. 44, n. 12, p. 1707–1715, dez. 2009.

DÍAZ, J.S.G.; DELGADO, N.O.; GAMBOA, A.B.; BUNNING, S.; GUEVARA, M.; MEDINA, E.; OLIVERA, C.; OLMEDO, G.; RODRÍGUEZ, L.M.; SEVILLA, V.; VARGAS, R. Estimación del carbono orgánico en los suelos de ecosistema de páramo en Colombia. **Revista Ecosistemas**, v. 29, n. 1, 2 abr. 2020.

DU, J.; KIMBALL, J. S.; MOGHADDAM, M. Theoretical Modeling and Analysis of L- and P-band Radar Backscatter Sensitivity to Soil Active Layer Dielectric Variations. **Remote Sensing**, v. 7, n. 7, p. 9450–9472, jul. 2015.

EMADI, M.; TAGHIZADEH-MEHRJARDI, R.; CHERATI, A.; DANESH, M.; MOSAVI, A.; SCHOLTEN, T. Predicting and Mapping of Soil Organic Carbon Using Machine Learning Algorithms in Northern Iran. **Remote Sensing**, v. 12, n. 14, p. 2234, jan. 2020.

FARIA, M. M.; FERNANDES FILHO, E. I. **Avaliação dos algoritmos SVM e Maxver para a classificação de sistemas florestais monodominantes de candeia (Eremanthus sp.)**. 2013.

FAVROT, J.C. 1989. A strategy for large scale soil mapping: the reference areas method. Science du Sol 27: 351-368 (in French, with abstract in English).

FIGUEIREDO, S. R. Mapeamento supervisionado de solos através do uso de regressões logísticas múltiplas e sistema de informações geográficas. 2006.

GALLANT, J. C.; DOWLING, T. I. A multiresolution index of valley bottom flatness for mapping depositional areas. **Water Resources Research**, v. 39, n. 12, 2003.

GAMA, F.; SANTOS, J.; MURA, J.; RENNÓ, C. Estimativa de Parâmetros Biofísicos de Povoamentos de Eucalyptus Através de Dados SAR Estimation of Biophysical Parameters in the Eucalyptus Stands by SAR Data. **Ambiência**, v. 2, 21 out. 2009.

GAMA, J. Functional Trees. **Machine Learning**, v. 55, p. 219–250, 1 jun. 2004.

GARSON, G. D. **Quantitative Research in Public Administration. NC State University**, 2005.

GESSLER, P. E.; MOORE, I. D.; MCKENZIE, N. J.; RYAN, P. J.Soil-landscape modelling and spatial prediction of soil attributes. **International journal of geographical information systems**, v. 9, n. 4, p. 421–432, jul. 1995.

GIASSON, E.; HARTEMINK, A. E.; TORNQUIST, C. G.; TESKE, R.; BAGATINI, T. Avaliação de cinco algoritmos de árvores de decisão e três tipos de modelos digitais de elevação para mapeamento digital de solos a nível semidetalhado na Bacia do Lageado Grande, RS, Brasil. **Ciência Rural**, v. 43, n. 11, p. 1967–1973, nov. 2013.

GISLASON, P. O.; BENEDIKTSSON, J. A.; SVEINSSON, J. R. Random Forests for land cover classification. **Pattern Recognition Letters**, Pattern Recognition in Remote Sensing (PRRS 2004). v. 27, n. 4, p. 294–300, 1 mar. 2006.

GÖKCEOGLU, C.; AKSOY, H. Landslide susceptibility mapping of the slopes in the residual soils of the Mengen region (Turkey) by deterministic stability analyses and image processing techniques. **Engineering Geology**, v. 44, n. 1, p. 147–161, 1 out. 1996.

GOWER, J. C. A General Coefficient of Similarity and Some of Its Properties. **Biometrics**, v. 27, n. 4, p. 857–871, 1971.

GRIMALDI, S.; NARDI, F.; DI BENEDETTO, F.; ISTANBULLUOGLU, E.; BRAS, R. L. A physically-based method for removing pits in digital elevation models. **Advances in Water Resource**s. 2007. 30(10), 2151-2158.

GRIMM, R.; BEHRENS, T.; MARKER, M.; ELSENBEER, H. Soil organic carbon concentrations and stocks on Barro Colorado Island -- Digital soil mapping using Random Forests analysis. 2008.

GRINAND, C.; ARROUAYS, D.; LAROCHE, B.; MARTIN, M. P. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. **Geoderma**, v. 143, n. 1–2, p. 180–190, jan. 2008.

GUJARATI, D. N. **Econometria Básica. 3. ed. São Paulo: Makron Books**, 2000.

GUO, L.; ZHANG, H.; SHI, T.; CHEN, Y.; JIANG, Q.; LINDERMAN, M. Prediction of soil organic carbon stock by laboratory spectral data and airborne hyperspectral images. **Geoderma**, v. 337, p. 32–41, 1 mar. 2019.

GUO, P.-T.; LI, M.-F.; LUO, W.; TANG, Q.-F.; LIU, Z.-W.; LIN, Z.-M.. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. **Geoderma**, v. 237–238, p. 49–59, 1 jan. 2015.

HALL, G.F. & OLSON, C. G. Predicting variability of soils from landscape models. In: MAUSBACH, M. J. and WILDING, L. P. [Eds.] Spatial variabilities of soils and landforms. SSSA Special Publication 28. SSSA. Madison. WI. p.9-24,1991.

HÄRING, T.; DIETZ, E.; OSENSTETTER, S.; KOSCHITZKI, T.; SCHRÖDER, B. Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. **Geoderma**, v. 185–186, p. 37–47, set. 2012.

HENDERSON, F. M.; LEWIS, A. J. (EDS.). Principles and Applications of Imaging Radar. 3rd edition ed. New York: Wiley, 1998.

HENGL, T.; HEUVELINK, G. B. M.; KEMPEN B.; LEENAARS J.G. B.; WALSH M.G.; SHEPHERD K.D.; SILA A.; MACMILLAN R.A.; MENDES DE JESUS J. S.; TAMENE L.; TONDOH J.E. Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. **PLOS ONE**, v. 10, n. 6, p. e0125814, 25 jun. 2015.

HEUNG, B.; BULMER, C. E.; SCHMIDT, M. G. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. **Geoderma**, v. 214–215, p. 141–154, fev. 2014.

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, p. 958–966, 1 dez. 2014.

HOUNKPATIN, K. O. L.; STENDAHL, J.; LUNDBLAD, M.; KARLTUN, E. Predicting the spatial distribution of soil organic carbon stock in Swedish forests using a group of covariates and site-specific data. **Soil**, v. 7, n. 2, p. 377–398, 6 jul. 2021.

HUANG, C.; DAVIS, L. S.; TOWNSHEND, J. R. G. An assessment of support vector machines for land cover classification. **International Journal of Remote Sensing**, v. 23, n.4, p. 725-749, 2002.

HUTCHINSON, M. F.; GALLANT, J. C. **Digital elevation models and representation of terrain shape**. 2000.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE) - **Manuais:** tutorial de geoprocessamento SPRING. 2008.

JASIEWICZ, J.; STEPINSKI, T. F. Geomorphons — a pattern recognition approach to classification and mapping of landforms. **Geomorphology** 182, p.147–156, 2013.

KHEIR, R.B., GREVE, M. H., BØCHER, P. K., GREVE, M. B., LARSEN, R., MCCLOY, K. Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: The case study of Denmark. **Journal of Environmental Management** 91: 1150-1160. 2010.

KOVAČEVIĆ, M.; BAJAT, B.; GAJIĆ, B. Soil type classification and estimation of soil properties using support vector machines. **Geoderma**, v. 154, p. 340–347, 15 jan. 2010.

LAGACHERIE, P.; LEGROS, J. P.; BURFOUGH, P. A. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. **Geoderma**, v. 65, n. 3, p. 283–301, 1 mar. 1995.

LAGACHERIE, P.; ROBBEZ-MASSON, J.; NGUYEN-THE, N. Mapping of reference area representativity using a mathematical soilscape distance. **Geoderma**, v. 101, p. 105–118, 1 abr. 2001.

LAGACHERIE, P.; VOLTZ, M. Predicting soil properties over a region using sample information from a mapped reference area and digital elevation data: a conditional probability approach. **Geoderma**, v. 97, n. 3, p. 187–208, 1 set. 2000.

LAMICHHANE, S.; KUMAR, L.; WILSON, B. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. **Geoderma**, v. 352, p. 395–413, out. 2019.

LEHMANN, J.; KLEBER, M. The contentious nature of soil organic matter. **Nature**, v. 528, n. 7580, p. 60–68, dez. 2015.

LIESS, M.; GLASER, B.; HUWE, B. Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models. **Geoderma**, v. 170, p. 70–79, 15 jan. 2012.

LIEß, M., GLASER, B., & HUWE, B. Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models. **Geoderma**, 170, 70-79. 2012.

LIN, D. S.; WOOD, E. F.; BEVEN, K.; SAATCHI, S. Soil moisture estimation over grass-covered areas using AIRSAR. **International Journal of Remote Sensing**, v. 15, n. 11, p. 2323–2333, 20 jul. 1994.

MA, Y.; MINASNY, B.; WU, C. Modelling and mapping of key soil properties to support agricultural production in Eastern China. **Geoderma Regional**, v. 10, 1 jun. 2017.

MALLAVAN, B. P.; MINASNY, B.; MCBRATNEY, A. B. Homosoil, a Methodology for Quantitative Extrapolation of Soil Information Across the Globe. Em: BOETTINGER, J. L.; HOWELL, D.; MOORE, A.M.; HARTEMINK, A.E.; KIENAST-BROWN, S. (Eds.). **Digital Soil Mapping: Bridging Research, Environmental Application, and Operation**. Progress in Soil Science. Dordrecht: Springer Netherlands, 2010. p. 137–150.

MARCHI, L.; DALLA FONTANA, G. GIS morphometric indicators for the analysis of sediment dynamics in mountain basins. **Environmental Geology**, v. 48, n. 2, p. 218–228, jul. 2005.

MATHIAS, D. T.; LUPINACCI, C. M.; NUNES, J. O. R. The identification of runoff flows in an area of technogenic relief using hydrological models in GIS. **Sociedade & Natureza**, v. 32, p. 738–748, 2020.

MCBRATNEY, A. B.; MENDONÇA SANTOS, M. L.; MINASNY, B. On digital soil mapping. **Geoderma**, v. 117, n. 1, p. 3–52, 1 nov. 2003.

MCNICOL, G.; BULMER, C.; D'AMORE, D.; SANBORN, P.; SAUNDERS, S.; GIESBRECHT, I.; ARRIOLA, S.G.; ALLISON BIDLACK, A.; BUTMAN, D.; BUMA, B. Large, climate-sensitive soil carbon stocks mapped with pedology-informed machine learning in the North Pacific coastal temperate rainforest. **Environmental Research Letters**, v. 14, n. 1, p. 014004, jan. 2019.

MEHRABI-GOHARI, E.; MATINFAR, H. R.; JAFARI, A.; TAGHIZADEH-MEHRJARDI, R.; TRIANTAFILIS, J. The Spatial Prediction of Soil Texture Fractions in Arid Regions of Iran. **Soil Systems**, v. 3, n. 4, p. 65, dez. 2019.

MEIER, M.; SOUZA, E.; FRANCELINO, M. R.; FERNANDES, 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.

MELGANI, F.; BRUZZONE, L. Classification of Hyperspectral Remote Sensing Images with Support Vector Machines. **Geoscience and Remote Sensing, IEEE Transactions on**, v. 42, p. 1778–1790, 1 set. 2004.

MINASNY, B.; HARTEMINK, A. E. Predicting soil properties in the tropics. **Earth-Science Reviews**, v. 106, n. 1–2, p. 52–62, 2011.

MITRAN, T.; MISHRA, U.; LAL, R.; RAVISANKAR, T.; SREENIVAS, K. Spatial distribution of soil carbon stocks in a semi-arid region of India. **Geoderma Regional**, v. 15, p. e00192, 1 dez. 2018.

MOGHADDAM, M.; SAATCHI, S.; TREUHAFT, R. **Estimating soil moisture in a boreal old jack pine forest**. IGARSS'97. 1997 IEEE International Geoscience and Remote Sensing Symposium Proceedings. Remote Sensing - A Scientific Vision for Sustainable Development. **Anais**..IEEE, 1997.

MOORE, I. D.; GRAYSON, R. B.; LADSON, A. R. Digital terrain modelling: A review of hydrological, geomorphological, and biological applications. **Hydrological Processes**, v. 5, n. 1, p. 3–30, jan. 1991.

MOORE, I.D.; GESSLER, P.E.; NIELSEN, G.A.; PETERSON, G.A. Soil attribute prediction using terrain analysis. **Soil Science Society of America Journal**, v.57. p.443-452. 1993.

MORAES, J. L.; CERRI, C. C.; MELILLO, J. M.; KICKLIGHTER, D.; NEILL, C.; SKOLE, D. L.; STEUDLER, P. A. Soil Carbon Stocks of the Brazilian Amazon Basin. **Soil Science Society of America Journal**, v. 59, n. 1, p. 244–247, 1995.

NASCIMENTO, R. F. F.; ALCÂNTARA, E.H.; KAMPEL, M.; STECH, J.L.; MORAES, E.M.L.; FONSECA, L.M.G.; 2009. O algoritmo Support Vector Machines (SVM): avaliação da separação ótima de classes em imagens CCD-CBERS-2. **Simpósio Bras. Sensoriamento Remoto SBSR 14 2009 Natal Anais São José Campos INPE. 2009.**

NGUYEN, M. Q., ATKINSON, P. M., & LEWIS, H. G. Superresolution mapping using a Hopfield neural network with fused images. **Geoscience and Remote Sensing**, IEEE Transactions 2006, 44(3), 736-749.

PADARIAN, J.; MINASNY, B.; MCBRATNEY, A. B.; DALGLIESH, N. Predicting and mapping the soil available water capacity of Australian wheatbelt. **Geoderma Regional**, v. Complete, n. 2–3, p. 110–118, 2014.

PETROBRAS. Centro de Pesquisa da Petrobras (CENPES). **Mapa detalhado de Solos do entorno da malha viária da Província Petrolífera de Urucu/Coari-AM**. Convênio UFRRJ/PETROBRAS/EMBRAPA/INPA/UFAM, 2010.

PINHEIRO, H. S. K. **Métodos de mapeamento digital aplicados na predição de classes e atributos dos solos da bacia hidrográfica do rio Guapi Macacu, RJ**. 30 jul. 2015.

PINHEIRO, H. S. K.; CARVALHO JUNIOR, W.; CHAGAS, C. S.; ANJOS, L. H. C.; OWENS, P. R. Prediction of Topsoil Texture Through Regression Trees and Multiple Linear Regressions. **Revista Brasileira de Ciência do Solo**, v. 42, 16 abr. 2018.

PINTO, A. C. B.; AZEVEDO, C.; JR, O. DE C. Activity patterns and diet of the howler monkey Alouatta belzebul in areas of logged and unlogged forest in Eastern Amazonia. **Animal Biodiversity and Conservation**, p. 11, 2003.

RAMIFEHIARIVO, N.; BROSSARD, M.; GRINAND, C.; ANDRIAMANANJARA, A.; RAZAFIMBELO, T.; RASOLOHERY, A.; RAZAFIMAHATRATRA, H.; SEYLER, F., RANAIVOSON, N.; RABENARIVO, M.; ALBRECHT, A.; RAZAFINDRABE, F.;

RAZAKAMANARIVO, H. Mapping soil organic carbon on a national scale: Towards an improved and updated map of Madagascar. **Geoderma Regional**, Digital soil mapping across the globe. v. 9, p. 29–38, 1 jun. 2017.

RENNÓ, C. D. **Avaliação de medidas texturais na discriminação de classes de uso utilizando imagens SIR-C/X-SAR do perímetro irrigado de Bebedouro, Petrolina, PE.** São José dos Campos: INPE, 111p. 2003. (Dissertação de Mestrado)

RODRIGUEZ-GALIANO, V. F.; GHIMIRE, B.; ROGAN, J.; CHICA-OLMO, M.; RIGOL-SANCHEZ, J. P. An assessment of the effectiveness of a random forest classifier for land-cover classification. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 67, p. 93–104, 1 jan. 2012.

ROSA, R. A. S. **Correção Radiométrica de Imagens de Radar de Abertura Sintética Aerotransportado.** 86f. Tese de Mestrado em Engenharia Eletrônica e Computação – Instituto Tecnológico de Aeronáutica, São José dos Campos. 2009.

SAATCHI, S.; MARLIER, M.; CHAZDON, R. L.; CLARK, D. B.; RUSSELL, A. E. Impact of spatial variability of tropical forest structure on radar estimation of aboveground biomass. **Remote Sensing of Environment**, v. 115, n. 11, p. 2836–2849, nov. 2011.

SAMBATTI, M. B. J.; LEDUC, R.; LUBECK, D.; ROBERTO MOREIRA, J.; ROBERTO, S. J. Assessing Forest Biomass and Exploration in the Brazilian Amazon with Airborne InSAR: an Alternative for REDD. **The Open Remote Sensing Journal**, v. 5, n. 1, 30 maio 2012.

SANTOS, E. M. DOS. **Teoria e aplicação de support vector machines à aprendizagem e reconhecimento de objetos baseado na aparência.** 20 jun. 2002.

SANTOS, H. G.; JACOMINE, P.K.T.; ANJOS, L.H.C. dos; OLIVEIRA, V.A.; LUMBRERAS, J.F.; COELHO, M.R.; ALMEIDA, J.A.; ARAUJO FILHO, J.C.; OLIVEIRA, J.B.; CUNHA, T. J. F. **Sistema Brasileiro de Classificação de Solos.** Brasília, DF: Embrapa, 2018.

SCHILLACI, C.; LOMBARDO, L.; SAIA, S.; FANTAPPIÈ, M.; MÄRKER, M.; ACUTIS, M. Modelling the topsoil carbon stock of agricultural lands with the Stochastic Gradient Treeboost in a semi-arid Mediterranean region. **Geoderma**, v. 286, p. 35–45, 15 jan. 2017.

SCULL, P.; FRANKLIN, J.; CHADWICK, O. A. The application of classification tree analysis to soil type prediction in a desert landscape. **Ecological Modelling**, v. 181, n. 1, p. 1–15, 10 jan. 2005.

SILVA, S. H. G.; MENEZES, M. D.; OWENS, P. R.; CURI, N. 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, 1 abr. 2016.

SOMARATHNA, P. D. S. N.; MINASNY, B.; MALONE, B. P. More Data or a Better Model? Figuring Out What Matters Most for the Spatial Prediction of Soil Carbon. **Soil Science Society of America Journal**, v. 81, n. 6, p. 1413–1426, 2017.

SREENIVAS, K.; SUJATHA, G.; SUDHIR, K.; KIRAN, D. V.; FYZEE, M. A.; RAVISANKAR, T.; DADHWAL, V. K. **Spatial Assessment of Soil Organic Carbon Density Through Random Forests Based Imputation**. 2014.

SRIVASTAVA, H. S.; PATEL, P.; NAVALGUND, R. R. Incorporating soil texture in soil moisture estimation from extended low-1 beam mode RADARSAT-1 SAR data. **International Journal of Remote Sensing**, v. 27, n. 12, p. 2587–2598, jun. 2006.

STUM, A. K.; BOETTINGER, J. L.; WHITE, M. A.; RAMSEY, R. D. Random Forests Applied as a Soil Spatial Predictive Model in Arid Utah. *Em*: BOETTINGER, J.L.; HOWELL, D. W.; MOORE, A.C.; HARTEMINK, A. E.; KIENAST-BROWN, S. **Digital Soil Mapping**. Dordrecht: Springer Netherlands, p. 179–190. 2010

TAGHIZADEH-MEHRJARDI, R.; SARMADIAN, F.; MINASNY, B.; TRIANTAFILIS, J.; OMID, M.. Digital Mapping of Soil Classes Using Decision Tree and Auxiliary Data in the Ardakan Region, Iran. **Arid Land Research and Management**, v. 28, n. 2, p. 147–168, 3 abr. 2014.

TEN CATEN, A.; DALMOLIN, R. S. D.; PEDRON, F. A.; MENDONÇA-SANTOS, M. L. Regressões Logísticas Múltiplas: fatores que influenciam sua aplicação na predição de classes de solos. **Revista Brasileira de Ciência do Solo**, v. 35, n. 1, p. 53–62, fev. 2011.

TESFA, T.; TARBOTON, D.; CHANDLER, D.; MCNAMARA, J. Modeling Soil Depth from Topographic and Land Cover Attributes. **James P. McNamara**, v. 45, 1 out. 2009.

THOMPSON, J. A.; BELL, J. C.; BUTLER, C. A. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. **Geoderma**, v. 100, n. 1–2, p. 67–89, mar. 2001.

VASCONCELOS, G. T.; OLIVEIRA, S. R. DE M. **Avaliação da eficiência de algoritmos de aprendizado de máquina para classificação automática de solos.** [s.l.] In: Congresso interinstitucional de iniciação científica, 12., 2018, Campinas. Anais... [S.l: s.n], 2018., 2018.

VAYSSE, K.; LAGACHERIE, P. Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). **Geoderma Regional**, v. 4, p. 20–30, 1 abr. 2015.

VILLELA, A. L. O. **Mapeamento digital de solos da Formação Solimões sob floresta tropical amazônica.** 29 ago. 2013.

VOLTZ, M.; LAGACHERIE, P.; LOUCHART, X. Predicting soil properties over a region using sample information from a mapped reference area. **European Journal of Soil Science**, v. 48, n. 1, p. 19–30, 1997.

WADOUX, A. M. J. C.; MCBRATNEY, A. B. Hypotheses, machine learning and soil mapping. **Geoderma**, v. 383, p. 114725, 1 fev. 2021.

WADOUX, A. M. J.-C.; MINASNY, B.; MCBRATNEY, A. B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. **Earth-Science Reviews**, v. 210, p. 103359, 1 nov. 2020.

WALKLEY, A.; BLACK, I. An examination of the Degtjareff method for detemrining soil organic matter, and a proposed modification of the chromic acid titration method. **Soil Science** 37, 29–37.1934.

WANG, S.; ZHUANG, Q.; WANG, Q.; JIN, X.; HAN, C.. Mapping stocks of soil organic carbon and soil total nitrogen in Liaoning Province of China. **Geoderma**, v. 305, p. 250–263, 1 nov. 2017.

WIESMEIER, M.; BARTHOLD, F.; BLANK, F.; KÖGEL-KNABNER, I. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. **Plant and Soil**, v. 340, p. 7–24, 1 mar. 2011.

WILLMOTT, C. J. **Some comments on the evaluation of model performance. Bull Am Metereol Soc. 63: 1309-1313.**, 1982.

WILSON, J. P.; GALLANT, J. C. (EDS.). **Terrain analysis: principles and applications**. New York: Wiley, 2000.

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, ago. 2017.

WU, J. Effects of changing scale on landscape pattern analysis: scaling relations. **Landscape Ecology**, v. 19, n. 2, p. 125–138, 1 mar. 2004.

ZRIBI, M.; SAHNOUN, M.; BAGHDADI, N.; LE TOAN, T.; BEN HAMIDA, A. Analysis of the relationship between backscattered P-band radar signals and soil roughness. **Remote Sensing of Environment**, v. 186, p. 13–21, 2016a.

ZRIBI, M.; SAHNOUN, M.; DUSSEAUX, R.; AFIFI, S.; BAGHDADI, N.; HAMIDA, A. **Analysis of P band radar signal potential to retrieve soil moisture profile**. 1 mar. **Anais** 2016b.