MAPBIOMAS
[SOLO]

MapBiomas Solo “Handbook”

ATBD

Algorithm Theoretical Basis Document

Coleção beta

Versão 1

Junho, 2023

SUMÁRIO

Para citar este documento em suas publicações, utilize o seguinte formato:

> MapBiomas, 2023, "Annual mapping of soil organic carbon stock in Brazil 1985-2021 (beta collection). Algorithm theoretical basis document and results", https://doi.org/10.58053/MapBiomas/3KXXVV, MapBiomas Data, V1
>
> MapBiomas, 2023, "Mapeamento anual do estoque de carbono orgânico do solo no Brasil 1985-2021 (coleção beta). Documento de base teórica do algoritmo e resultados", https://doi.org/10.58053/MapBiomas/3KXXVV, MapBiomas Data, V1

RESUMO EXECUTIVO

O projeto MapBiomas Solo, criado em 2021, visa revelar a dinâmica espaço-temporal das propriedades do solo e suas relações com as mudanças na cobertura e uso da terra no Brasil. O produto desta coleção consiste de uma série de mapas anuais dos estoques de carbono orgânico do solo (COS), medidos em toneladas por hectare (t/ha), nos primeiros 30 cm do solo brasileiro para o período de 1985-2021. A coleção beta é a primeira versão pública da série de mapas do MapBiomas Solo. Ela foi obtida utilizando modelos de regressão que computam as relações entre os dados pontuais de estoque de COS e covariáveis ambientais espacialmente explícitas.

Os dados de solo utilizados são provenientes de amostragem no campo e que estão disponíveis no repositório SoilData (https://soildata.mapbiomas.org/). Já as covariáveis ambientais são variáveis preditoras que representam os fatores de formação do solo (clima, organismos, relevo, material parental e tempo) e foram obtidas de bancos de dados espaciais abertos, incluindo dados anuais de cobertura e uso da terra produzidos pelo MapBiomas (Coleção 7.1). O cômputo das relações entre o estoque de COS e covariáveis ambientais foi implementado utilizando o algoritmo de aprendizado de máquina Random Forest no Google Earth Engine.

A série de mapas pode ser acessada na plataforma do MapBiomas (http://solo.mapbiomas.org/), onde são apresentados os estoques totais (tonelada) e por unidade de área (tonelada por hectare) para diferentes recortes territoriais e por classe de cobertura e uso da terra. Os dados revelam que o estoque total de COS nos primeiros 30 cm do solo brasileiro no ano de 2021 somavam aproximadamente 37,5 gigatoneladas (Gt), valor maior do que a estimativa do estoque total potencial de COS do Brasil, estimada pela literatura para solo sob vegetação nativa, em 36,4 Gt (Bernoux et al., 2002) para esta mesma camada de solo. Mapas produzidos por iniciativas nacionais e internacionais também apontam estoques similares de 34,6 Gt (Vasques et al., 2017) e 37,5 Gt (Poggio et al., 2021).

Apesar de serem resultados preliminares, eles apresentam potencial de utilização para aprimorar o entendimento da dinâmica de remoção e emissão de carbono no solo. O número de amostras disponíveis ainda é baixo, considerando a dimensão territorial do Brasil. Cada amostra utilizada representa cerca de mil quilômetros quadrados (1 amostra/ mil km²), o que equivale a área do município de Rio de Janeiro, o que confere elevada incerteza às estimativas (ainda não computadas). A meta do MapBiomas Solo é produzir atualizações anuais dessa série histórica. Com isso, o feedback dos usuários contribui para melhorias e aprimoramentos na precisão e qualidade dos dados, a fim de apoiar às necessidades e expectativas dos usuários para que a coleção se torne uma ferramenta valiosa e útil para pesquisadores, cientistas e tomadores de decisão em todo o país.

Este documento tem como objetivo fornecer as etapas metodológicas para produzir a coleção beta do MapBiomas Solo e descrever os conjuntos de dados utilizados. Todos os

mapas e conjuntos de dados do MapBiomas Solo estão disponíveis gratuitamente na plataforma do projeto (http://solo.mapbiomas.org/), bem como todos dados e os códigos utilizados para o processamento dos dados que estão listados no material suplementar deste documento e GitHub do MapBiomas (https://github.com/mapbiomas-brazil).

1. Introdução

1.1 Escopo e Conteúdo do Documento

Este documento descreve a base teórica, justificativa e métodos utilizados para produzir a coleção beta do MapBiomas Solo. O produto desta coleção beta consiste de uma série de mapas anuais dos estoques de carbono orgânico do solo (COS), medidos em toneladas por hectares (t/ha), nos primeiros 30 cm do solo brasileiro para o período de 1985-2021. A coleção beta é a primeira versão pública da série de mapas do MapBiomas Solo. Ela é resultado do primeiro ciclo de desenvolvimento e avaliação internos e representa aquilo que de melhor os dados e informações disponíveis permitiram produzir. A versão beta permite que os usuários finais (*beta testers*) utilizem os mapas nas mais diversas aplicações, identifiquem inconsistências críticas (*bugs*) e forneçam os meios necessários (sugestões e dados) para os desenvolvedores realizarem na sua correção e atualização.

1.2 Visão Geral

Os mapas anuais da coleção beta foram obtidos por modelos de regressão implementados na forma de algoritmos de aprendizado de máquina. Esses algoritmos computam a relação numérica entre estoque de COS e covariáveis ambientais espacialmente explícitas. Os dados pontuais de estoque de COS utilizados para treinamento dos algoritmos são provenientes de amostragem de solo no campo desde a primeira metade do século XXI e estão disponíveis no Repositório Brasileiro de Dados do Solo, o SoilData (https://soildata.mapbiomas.org/). As covariáveis ambientais são variáveis preditoras que representam fatores de formação do solo (clima, organismos, relevo, material parental e tempo). Todas elas foram obtidas de bancos de dados espaciais abertos. O processamento das covariáveis, o treinamento dos modelos preditivos e as predições espaço-temporais dos estoques de COS foram realizados no Google Earth Engine. Todos os produtos estão armazenados no Google Cloud Storage Platform e disponíveis sob uma licença Creative Commons Attribution-ShareAlike (CC BY-SA).

1.3 Região de interesse

O MapBiomas Solo produziu mapas anuais do estoque de COS para todo o território brasileiro de modo a cobrir as áreas emersas dos seis biomas oficiais do país: Amazônia, Mata Atlântica, Caatinga, Cerrado, Pampa e Pantanal (Figura 1). Os mapas cobrem cinco das seis classes de cobertura e uso da terra no nível 1 da legenda na Coleção 7.1 do MapBiomas. São elas as áreas de terras emersas em: 1 - Floresta, 10 - Formação não-florestal, 14 - Agropecuária, 22 - Área não-vegetada e 27 - Não observado. Não foram feitas predições dos estoques de COS apenas para áreas consideradas como permanentemente submersas, isto é, áreas classificadas como Água (classe 26 e suas subclasses 31 - Aquicultura e 33 - Rio, Lago e Oceano). Considerando o regime sazonal e plurianual que algumas áreas alagáveis podem apresentar, como a planície pantaneira e áreas de várzea, foi definida como área de água permanente aquelas que permaneceram submersas por pelo menos 35 anos ao longo da

série de 37 anos (1985-2021). As demais áreas, que ficaram submersas por período menor que 35 anos, foram definidas como “áreas alagáveis” e receberam predições do estoque de COS.

Figura 1. Biomas brasileiros mapeados no projeto MapBiomas Solo para gerar os produtos da coleção beta de carbono orgânico do solo (fonte: MapBiomas, adaptado de IBGE, 2019).

1.4 Ciência e Aplicações-Chave

Conhecer esses estoques de carbono orgânico do solo e sua variação espaço-temporal é fundamental para implementar políticas públicas e programas eficientes de conservação, uso e manejo sustentável dos recursos naturais como estratégia de mitigação do impacto da

mudança climática. Uma série de mapas anuais dos estoques de COS tem potencial de aplicação em diversas áreas nos setores público e privados, dentre elas:

- Subsidiar a prevenção de perda de COS a partir da identificação de áreas e fatores naturais e antrópicos responsáveis pela redução dos estoques de carbono no país.
- Identificar solos com os maiores estoques de COS do país para direcionamento de políticas nacionais e regionais de conservação do solo para evitar a perda do carbono, do solo para atmosfera, nestas áreas. Esses solos incluem solos orgânicos, que são particularmente vulneráveis às mudanças de uso da terra e estão sob constante pressão antrópica devido ao seu alto potencial de produção.
- Fornecer informações espaço-temporais de qualidade para que os usuários da terra implementem e mantenham práticas apropriadas de manejo do solo e da terra para proteger e aumentar o COS sob as condições locais para benefício de longo prazo.
- Fornecer informações espaço-temporais de qualidade para auxiliar na avaliação de implementação de políticas públicas de recuperação de áreas degradadas (uma das metas até 2030).
- Monitorar e avaliar terras e propriedades rurais.
- Avaliar políticas e programas de mitigação e adaptação à mudança climática.
- Identificar áreas com maior viabilidade para entrada no mercado de comercialização de créditos de carbono.
- Monitorar os resultados da aplicação de linhas de crédito.
- Avaliar pedidos de financiamento voltados à mitigação e adaptação à mudança climática.
- Desenvolver novas e mais abrangentes pesquisas sobre a dinâmica do COS.

Realizar este potencial é a meta do MapBiomas Solo. A contribuição dos usuários com sugestões e dados para as correções das inconsistências críticas que identificarem nesta versão beta são fundamentais para que os produtos venham a atingir a qualidade requerida para viabilizar essas aplicações. Alcançar estes objetivos contribuirá para a mitigação das mudanças climáticas, reduzindo as emissões de carbono para a atmosfera e para a segurança alimentar, preservando a saúde do solo e os benefícios associados, incluindo a redução do risco de maior degradação do solo.

2. Visão Geral e Informações Básicas

2.1 Contexto e Informações-Chave

O solo ocupa posição de destaque no sistema climático global, havendo mais carbono armazenado em seus primeiros 100 cm do que na atmosfera (Lal, 2013). Por isso, o solo pode ser tanto ameaça quanto solução à mudança climática. As mudanças no uso da terra e as práticas inadequadas de manejo do solo adotadas pela humanidade o tornaram fonte líquida de gases do efeito estufa.

Práticas conservacionistas podem reverter esse cenário, removendo gás carbônico atmosférico para armazená-lo na forma de COS. Tais práticas aumentam a qualidade do solo pela ação positiva do COS em suas propriedades (IPCC, 2014), recuperando áreas degradadas e enriquecendo solos com estoques de COS abaixo do seu potencial. Conhecer a localização desses estoques nos biomas brasileiros e sua dinâmica temporal é crucial no cômputo de estimativas realistas de emissões e remoções de gases de efeito estufa pelo país (SEEG, 2022). São essas estimativas que norteiam as políticas públicas e programas criados para atingir a meta de redução de emissões definida pelo Brasil no Acordo de Paris (Brasil, 2015).

O Brasil é um dos países cuja economia possui maior dependência do uso do solo. Ainda assim, o país não possui informações detalhadas do solo e suas propriedades no espaço e no tempo. Todos os esforços de mapear os estoques de COS até então ignoraram a dinâmica temporal provocada pelas mudanças na terra e no clima. Os estoques estimados foram reportados como valores absolutos ou mapas atemporais (Poggio et al., 2021). A principal razão para isso foi a indisponibilidade de dados de campo para treinamento de modelos preditivos. Apesar do grande volume de dados de solo produzidos no país (Camargo et al., 2010), a maioria deles ainda é difícil de encontrar, acessar e reutilizar (Hanson et al., 2011).

Os primeiros esforços para mudar esse cenário tiveram início em 2016 com o repositório brasileiro de dados do solo, o SoilData (Samuel-Rosa et al., 2020). A missão do repositório SoilData (https://soildata.mapbiomas.org/) é salvaguardar todos os tipos de dados do solo para reuso na geração de dados abertos secundários em escala nacional. Desde então, o SoilData já resgatou e curou dados de solo de mais de 20 mil pontos de observação e amostragem vindos de organizações públicas e privadas. Foi esse esforço que possibilitou vislumbrar, pela primeira vez, a produção de uma série histórica de mapas anuais dos estoques de COS em todos os biomas brasileiros pela iniciativa MapBiomas.

O MapBiomas Solo tem como objetivo entender os efeitos das mudanças na cobertura e uso da terra na variação espaço-temporal dos estoques de COS em todos os biomas brasileiros. A produção da coleção beta foi viabilizada pela colaboração com organizações públicas e privadas, que ampliaram a disponibilidade de dados de COS via SoilData. A meta do

MapBiomas Solo é produzir atualizações anuais da série histórica de mapas de estoque de COS e suas medidas de incerteza local e global.

2.2 Perspectiva Histórica

2.2.1 Dados de Solo de Campo

Os cientistas brasileiros são responsáveis pela produção do maior volume de dados e informações do solo em regiões tropicais do planeta (Camargo et al., 2010). Contudo, as práticas adotadas para armazenamento e compartilhamento de dados do solo costumam ser ineficientes, pois limitam a 'descobertabilidade' (do inglês, *discoverability*) dos dados do solo, isto é, a habilidade que um determinado conjunto de dados do solo tem de ser descoberto por outras pessoas. No Brasil, muitos conjuntos de dados são difíceis de serem descobertos porque estão disponíveis somente em documentos impressos ou (mal) escanerizados (Kämpf, 1971). Quando fáceis de serem descobertos, talvez eles precisem ser digitalizados ou podem estar contidos em mídias protegidas por senha ou em bancos de dados que são difíceis de usar e/ou dependem de programas de computador proprietários (Chagas et al., 2004; Ottoni et al., 2014).

Quando não existem barreiras tecnológicas, ainda assim os conjuntos de dados do solo podem conter dados de apenas algumas variáveis do solo ou profundidades do solo selecionadas do conjunto original de dados do solo, ou apenas uma versão agregada dos dados do solo está disponível (Cooper et al., 2005; Ottoni et al., 2014; Samuel-Rosa et al., 2013). E quando um conjunto de dados do solo está totalmente disponível, os metadados podem estar incompletos ou ausentes. Esses aspectos, junto de uma pobre organização dos dados do solo, limitam a 'reusabilidade' (do inglês, *reusability*) dos dados do solo, isto é, a habilidade que um determinado conjunto de dados do solo tem de ser usado novamente por seu produtor ou outra pessoa. Por fim, uma limitada descobertabilidade e reusabilidade dos dados do solo dificulta a observância de um dos princípios básicos do método científico, ou seja, a reprodutibilidade da pesquisa.

Muitos esforços já foram empreendidos por diversas instituições do mundo todo para compilar, organizar e distribuir gratuitamente dados do solo, principalmente aqueles coletados nas muitas décadas passadas e deixados para as próximas gerações, os chamados dados legados (Arrouays et al., 2017; Botelho and Brilha, 2022). Por exemplo, a Africa Soil Profiles Database, o Archivo Digital de Perfiles de Suelo de México, o Sistema de Información de Suelos del INTA, o Soil and Landscape Grid of Australia - CSIRO, e a USA National Cooperative Soil Survey. Em nível global, alguns esforços colaborativos surgiram nos últimos anos com o objetivo de compilar e disponibilizar dados de propriedades específicas do solo, por exemplo, dados de infiltração e condutividade hidráulica saturada do solo (Rahmati et al., 2018). Atualmente, o maior mantenedor e distribuidor de dados legados do solo é o World Soil Information Service (WoSIS), criado e mantido pelo International Soil Reference and Information Centre (ISRIC) (Batjes et al., 2020, 2017). Os benefícios da construção de um serviço como esse já foram demonstrados inúmeras vezes.

Por exemplo, via uso de técnicas modernas de aprendizado de máquina e mineração de dados para análise e modelagem dos dados do WoSIS, o ISRIC produziu mapas globais de diversas características do solo, os quais são servidos gratuitamente via plataforma SoilGrids (Hengl et al., 2014, 2017; Poggio et al., 2021), e são usados pelo Painel Intergovernamental sobre Mudanças Climáticas (IPCC).

No Brasil, dois órgãos têm se dedicado de maneira mais consistente à compilação, organização e distribuição de dados do solo. São eles a Empresa Brasileira de Pesquisa Agropecuária (Embrapa) e o Instituto Brasileiro de Geografia e Estatística (IBGE), fruto da sua ativa participação em grandes projetos de levantamento de dados e mapeamento do solo, especialmente o antigo Projeto Radambrasil. O IBGE, por exemplo, atualiza periodicamente a classificação taxonômica dos perfil do solo coletados via Projeto Radambrasil, compilando dados de novos perfis para atualizar e refinar a escala dos mapas do solo originalmente produzidos naquele projeto. Já a Embrapa dedica-se tanto aos dados de perfis do solo coletados via Projeto Radambrasil, como aos dados produzidos via seus outros projetos, inclusive os mais recentes, tendo organizado o Sistema de Informação de Solos Brasileiros (SISB). Juntos, estima-se que os dados sob salvaguarda do IBGE e da Embrapa somam cerca de 10 mil perfis do solo (Samuel-Rosa and Vasques, 2017).

Também houve no Brasil diversos esforços particulares para compilar, organizar e distribuir gratuitamente dados legados e recentes do solo obtidos em projetos de universidades e instituições de pesquisa, ou mesmo para aprimorar os dados sob salvaguarda do IBGE e Embrapa. Uma das primeiras iniciativas particulares consistiu na compilação parcial de dados do Projeto Radambrasil por pesquisadores da Escola Superior de Agricultura “Luiz de Queiroz” (ESALQ), que recuperaram as coordenadas espaciais de cerca de 6000 observações do solo a partir de mapas impressos e descrições contidas em relatórios (Cooper et al., 2005). Enquanto isso, o Serviço Geológico do Brasil (CPRM) criou um banco de dados físico-hídricos de cerca de 1000 amostras do solo, dados esses compilados a partir de centenas de trabalhos encontrados na literatura científica (Ottoni et al., 2018). Mais recentemente, também na ESALQ, foi lançado o projeto de construção de uma biblioteca espectral (BESB), que reuniu mais de 20 mil espectros, tanto de dados publicados, quanto de amostras do solo coletadas e enviadas por pesquisadores de todas as regiões do Brasil (Sato, 2015).

As iniciativas de compilação e organização de dados do solo no Brasil não produziram, até agora, uma solução permanente para o problema de salvaguardar todos os tipos de dados de solos. Elas também não desenvolveram uma estratégia capaz de efetivamente maximizar a descobertabilidade e reusabilidade dos conjuntos de dados do solo. Isso se deve, principalmente, à falta de padrões para compilação e organização de dados e à especificidade das metas de cada iniciativa. Assim, pode-se dizer que, em geral, as estratégias que usamos para gerenciar dados do solo no Brasil estão em desacordo com as demandas da sociedade pós-moderna e dificultam a reprodutibilidade da pesquisa do solo. No curto prazo, isso representa um uso ineficiente dos já escassos investimentos feitos na

ciência do solo. No longo prazo, isso detém o avanço do conhecimento do solo, as consequências refletindo no modo como os recursos naturais serão geridos nas próximas décadas e pelas gerações futuras.

O SoilData (https://soildata.mapbiomas.org/) foi criado com o propósito de servir de meio para a compilação, organização e publicação de todos os tipos de dados do solo no Brasil. Para isso são usados métodos baseados em experiências internacionais, principalmente uma política de dados abertos, primando pela facilidade de acesso, manutenção e uso. A meta do MapBiomas Solo é fazer do SoilData um repositório centralizado para armazenar e servir dados do solo em formato padronizado e harmonizado para várias aplicações. A principal dessas aplicações é a produção da dos mapas anuais de estoques de COS para todo o território nacional e viabilizar a sua atualização anualmente. Além disso, os dados do SoilData podem ser utilizados para o desenvolvimento de bases de dados especializadas, a melhoria do Sistema Brasileiro de Classificação do Solo, a criação do Sistema Universal de Classificação do Solo, a construção de sistemas inteligentes de recomendação de fertilizantes, e o recém-lançado Programa Nacional de Levantamento e Interpretação de Solos do Brasil (PronaSolos).

2.2.2 Mapeamento de Estoques de COS

Iniciativas globais e nacionais já produziram mapas do estoque de COS no Brasil, utilizando diferentes técnicas e abordagens. Esses mapas foram desenvolvidos com base em dados de campo nos quais o estoque de COS foi medido, juntamente com dados ambientais auxiliares que ajudam a explicar a distribuição espacial do COS em relação ao ambiente. Esses dados ambientais, as covariáveis, representam os fatores de formação do solo, desempenhando um papel importante na modelagem do estoque de COS. Entre essas covariáveis, destacam-se os dados geomorfométricos do terreno, como relevo, declividade e índice de umidade topográfica, que fornecem informações relevantes sobre a topografia e a hidrologia do local. Além disso, são considerados dados relacionados aos organismos presentes no ecossistema, características do material de origem do solo e outros fatores relevantes para compreender a distribuição do COS.

O primeiro mapa de estoque de COS do solo brasileiro foi gerado por Bernoux e colaboradores em 2002 (Bernoux et al., 2002). Os autores utilizaram diferentes bases de dados e uma metodologia que relaciona informações sobre a vegetação pretérita e tipo de solo em todo o território nacional, na profundidade de 0 a 30 cm. O resultado foi um mapa de estoque potencial de COS, em escala 1:1.000.000, resultante de 75 categorias diferentes de associação solo-vegetação. De acordo com o mapa produzido, as regiões brasileiras apresentam variações nos valores de COS devido aos diferentes fatores. De modo que, no Pantanal, a Região Central e noroeste da Bacia Amazônica apresentam solos úmidos com altos valores de estoque de COS. Na região Sul, os altos estoques de COS são resultado do clima mais frio, enquanto os baixos estoques da região nordeste são resultado do clima semiárido. Na Amazônia, a floresta densa tem um estoque de COS maior do que a floresta

aberta, enquanto no Cerrado, o tipo de solo exerce grande influência nos valores de COS. Os estoques médios representativos variam de 15,1 a 417,8 t/ha, totalizando cerca de 36,4 Gt de COS em todo país. É importante ressaltar que esses dados representam locais com vegetação nativa e não levam em consideração as alterações causadas pelo uso humano. O mapa de Bernoux é considerado referência para o cálculos das emissões de gases de efeito estufa do Brasil.

Recentemente, iniciativas nacionais se propuseram a mapear os estoques de COS do país (Gomes et al., 2019; Vasques et al., 2017). Essas iniciativas tiveram como objetivo identificar tendências em larga escala e priorizam a escala espacial, utilizando técnicas avançadas de mapeamento digital de solos. Devido a disponibilidade limitada de dados pontuais de solo estes mapas não levam em consideração a dimensão temporal. Neles, os dados coletados em diferentes épocas foram combinados para gerar um único produto. O mapa resultante representa um período de tempo, normalmente de décadas, do qual os dados de campo foram coletados. Como grande parte dos dados de solo disponíveis remontam às décadas de 1970 e 1980 (Samuel-Rosa et al., 2020), os produtos tendem a melhor representar estes períodos.

O mapa produzido por Gomes e colaboradores de 2019 (Gomes et al., 2019) utilizou 8227 perfis de solo e 37.693 amostras de solo oriundos do projeto RADAMBRASIL, coletadas entre as décadas de 70 e 80, que foram correlacionadas com 74 covariáveis ambientais, por quatro métodos de aprendizado de máquina (Random Forests, Cubist, Support Vector Machines e Generalized Linear Models) com diferentes características para avaliar a previsão mais precisa dos estoques de COS. O mapa final, na resolução de 1 km, revelou que os solos brasileiros armazenam aproximadamente 36 Gt de COS em seus primeiros 30 cm. O bioma Amazônia possui os maiores estoques de COS entre 0-100 cm, totalizando 36,1 Gt de COS, enquanto o Pantanal e a Caatinga menor valor total respectivamente (0,77 Gt de COS e 4,88 Gt de COS).

A Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), no âmbito do PronaSolos, já produziu duas versões de mapa dos estoques de COS brasileiro, uma lançada em 2017 para a profundidade de 0-30 cm e a versão mais recente lançada em 2021, que engloba também a profundidade de 0-100 cm, com uma resolução espacial de 1 km (Vasques et al., 2017). Foram mais de 8900 dados pontuais de solo de todo o país, disponíveis em bases de dados abertas e 58 covariáveis ambientais. Os mapas publicados em 2017 e 2021 apresentaram um estoque total de COS na profundidade de 0-30 cm estimado em 34,6 Gt de COS. Os mapas de COS, juntamente com os metadados correspondentes, estão disponíveis para consulta pública no Programa Nacional de Solos do Brasil (PronaSolos) e na Infraestrutura de Dados Espaciais da Embrapa (https://geoinfo.cnps.embrapa.br/).

Em nível internacional, o SoilGrids (https://soilgrids.org/), desenvolvido pela World Soil Information (ISRIC) em parceria com outras instituições, tem produzido um conjunto de

mapas globais de atributos do solo, incluindo o estoque de COS. Os mapas são gerados a partir de modelos de aprendizado de máquina (Random Forest) e combinam dados de campo, sensores remotos e outros dados auxiliares, para prever os atributos do solo em sete profundidades padrão (0, 5, 15, 30, 60, 100 e 200 cm) (Poggio et al., 2021). As previsões foram baseadas em cerca de 150 mil perfis de solo e contou com 158 covariáveis de solo baseadas em dados de sensoriamento remoto. No mapeamento foram utilizados dados de 5.086 perfis do Brasil que correspondem a 10.034 horizontes. Os mapas do SoilGrids gerados em 2016 têm resolução espacial de 250 m, pouco melhor que a primeira versão lançada em 2014, que teve uma resolução de 1 km. Uma das limitações nesta iniciativa é a distribuição heterogênea de dados de treinamento no mundo, por isso o mapa resultante melhor representar a variação em larga escala, favorecendo áreas que possuem maior volume de dados.

O Soils Revealed (https://soilsrevealed.org/) é uma plataforma para visualizar a influência do manejo do solo nos estoques de COS globalmente. Utilizando dados de solo disponíveis, informações sobre o ambiente e simulações computacionais ao longo do tempo. As projeções para as próximas décadas são baseadas em cenários elaborados pelo IPCC. A plataforma oferece uma abordagem dinâmica para exibir e comparar áreas com potencial de aumentar o COS, promovendo ações mitigadoras. No contexto brasileiro, o Soils Revelead destaca o Brasil como um dos detentores de maiores estoques de COS, graças à sua extensão territorial e à diversidade de ecossistemas presentes. Estima-se que os solos brasileiros possuem cerca de 31 Gt de COS. No entanto, dados revelam que houve uma perda média de 0,122 t/ha na profundidade de 0-30 cm entre os anos de 2000 e 2018. Estas projeções demonstram como os estoques de COS do país podem ser afetados pelas mudanças na cobertura e uso da terra e ressaltam a importância da adoção de práticas de manejo sustentável para preservar essa valiosa reserva nos solos brasileiros.

Recentemente a FAO, no âmbito da Aliança Global do Solo (Global Soil Partnership) produziu o mapa global de estoques de COS, o Global Soil Organic Carbon Map (GSOC). O GSOC (https://data.apps.fao.org/glosis/) é o primeiro mapa global produzido por meio de um processo consultivo e participativo pela FAO, abrangendo a camada de solo de 0-30 cm. Esse mapa foi elaborado com base em dados de inventários nacionais, estudos publicados, modelos de simulação e outras fontes e está disponível em formato raster com resolução de 1 km. O mapa produzido pela Embrapa foi incluído oficialmente como inventário de COS do Brasil.

A partir do mapa global, a FAO elaborou o potencial de sequestro de COS para o mundo. O Global Soil Organic Carbon Sequestration Potential Map (GSOCSeq), foi produzido com base no Global Soil Organic Carbon Map (GSOC), utilizando o modelo RothC. Esse mapa avalia a capacidade das diferentes regiões em armazenar carbono orgânico do solo ao longo do tempo. De acordo com os resultados desse mapeamento, o Brasil foi identificado como tendo o maior potencial de sequestro de COS em um cenários de adoção de práticas

conservacionistas. Isso destaca a relevância do país no contexto da mitigação das mudanças climáticas e ressalta a necessidade de desenvolver estratégias de conservação e manejo sustentável do solo para maximizar esse potencial. O relatório da FAO fornece mais detalhes sobre o uso do modelo RothC e os resultados específicos do mapeamento do potencial de sequestro de COS no Brasil e em outras regiões (FAO, 2020; Peralta et al., 2022).

Apesar de quase um terço do país estar sob uso antrópico, até agora, a dinâmica temporal de uso e cobertura da terra foram simplificados ou ignorados pelas iniciativas de mapeamento. A MapBiomas Solo, o Brasil está apresentando um avanço inédito no mapeamento temporal dos estoques COS, e pela primeira vez considerando os impactos das mudanças de uso e cobertura da terra ao longo dos últimos 37 anos. O projeto é elaborado com base em dados abertos e os produtos resultantes dessa iniciativa consistem em mapas anuais que retratam os estoques de COS (t/ha) nos primeiros 30 cm do solo a partir de sua superfície, abrangendo o período de 1985 a 2021. Os mapas fornecem uma resolução espacial detalhada de 30 m e serão atualizados anualmente, possibilitando uma visão dinâmica dos estoques de COS do Brasil.

3. Descrição do Algoritmo, Suposições e abordagens

A coleção de mapas anuais do estoque de COS do território brasileiro foi obtida utilizando modelos de regressão. Esses modelos de regressão computam a relação entre os dados pontuais de estoque de COS e covariáveis ambientais espacialmente exaustivas. Os dados de solo utilizados são provenientes de amostragem de solos em campo que possuem coordenadas espaciais e temporais conhecidas e estão disponíveis no repositório SoilData (https://soildata.mapbiomas.org/). Já as covariáveis ambientais são variáveis preditoras que representam os fatores de formação do solo (clima, organismos, relevo, material parental e tempo) e foram obtidas de bancos de dados espaciais abertos. O cômputo das relações entre o estoque de COS e covariáveis ambientais foi implementado utilizando o algoritmo de aprendizado de máquina chamado *Random Forest*. Para avaliar a qualidade das predições feitas pelo *Random Forest* ao longo da série de 37 anos (1985-2021), técnicas de validação cruzada foram utilizadas. As próximas seções descrevem cada um desses componentes em detalhe.

3.1 Dados pontuais de solo

O mapeamento espaço-temporal dos estoques de COS em todo o território brasileiro ao longo de uma série de 37 anos (1985-2021) requer grande volume de dados pontuais de propriedades do solo. As quatro essenciais são concentração de COS, volume de terra fina, densidade da terra fina e espessura da camada. Esses dados precisam ser coletados no campo e ser acompanhados das coordenadas do local e data do momento de amostragem. São as coordenadas espaciais que permitem relacionar o estoque de COS num ponto com dados ambientais que cobrem toda a área a ser mapeada. A coordenada temporal, por sua vez, é necessária para estabelecer tais relações ao longo do tempo.

O repositório SoilData, um dos produtos do MapBiomas Solo, contém a maior coleção de dados pontuais de propriedades do solo para o território brasileiro. São dados de mais de 20 mil pontos de observação e amostragem do solo, distribuídos em todo o território brasileiro, provenientes de mais de 250 conjuntos de dados coletados nos últimos 70 anos. A entrada desses dados no SoilData ocorre de duas maneiras. A primeira delas é pela atuação do MapBiomas Solo na compilação de dados de levantamentos de solos, dissertações e teses e artigos publicados em periódicos científicos. A segunda consiste no depósito de dados de pesquisa por seus produtores, incluindo estudantes de pós-graduação, docentes e pesquisadores de instituições públicas de ensino e pesquisa.

Após a entrada de um conjunto de dados no SoilData, a primeira ação da equipe de curadoria de dados é o seu mapeamento para um modelo padrão (padronização). Esse modelo, implementado na forma de planilha eletrônica, consiste numa coleção de cinco tabelas. Cada uma dessas tabelas armazena um grupo de informações específicas necessárias para a gestão e processamento dos dados. A tabela "citacao" recebe informações de identificação dos dados e seus autores. A tabela "evento" recebe os dados

relativos ao local e momento de observação e amostragem do solo. Já os dados das propriedades do solo determinadas nas amostras das camadas ou horizontes do solo são organizados na tabela "camada". Por fim, a tabela "metodo" armazena a descrição dos métodos analíticos empregados no campo e laboratório para produzir os dados armazenados nas tabelas "evento" e "camada". A quinta e última tabela do modelo de dados do SoilData, chamada "historico", é utilizada nas ações da equipe de curadoria de dados. Trata-se de uma tabela usada para registrar quaisquer alterações nos conjuntos de dados, servindo como um sistema de versionamento. Essas alterações são fruto da revisão dos dados, correção de inconsistências numéricas e erros de digitação e realização de melhorias. Dentre as melhorias realizadas nos conjuntos de dados estão a inferência de coordenadas espaciais e temporais faltantes e a melhoria da precisão das coordenadas existentes.

A reutilização dos dados depositados no SoilData para o mapeamento dos estoques de COS requer a sua harmonização. A harmonização é uma transformação dos dados de uma variável (propriedade do solo ou do ambiente) que tenham sido produzidos utilizando métodos díspares em cada conjunto de dados. Essa transformação resulta no mapeamento dos valores dessa variável para um mesmo espaço de atributos. Uma vez mapeados, os dados de uma variável provenientes de diferentes conjuntos de dados podem ser colados na mesma coluna da base de dados. No SoilData, a harmonização é feita usando soluções analíticas exatas (coordenadas espaciais) e regras heurísticas e modelos estatísticos (propriedades do solo e ambiente). No caso das regras heurísticas e modelos estatísticos, o resultado da harmonização são valores que mais provavelmente teriam sido obtidos caso um método de referência tivesse sido usado dadas as informações disponíveis.

As próximas seções descrevem o modo de obtenção e processamento dos dados do solo necessários para o cômputo do estoque de COS. Todos os procedimentos foram conduzidos usando o software R e pacotes disponíveis no CRAN (https://cran.r-project.org/). Os dados e códigos utilizados são de acesso aberto e estão disponíveis no repositório de dados do MapBiomas em https://doi.org/10.58053/MapBiomas/4FJWZC.

3.1.1 Concentração de carbono orgânico no solo (COS)

O COS é parte da matéria orgânica do solo, a fração do solo cuja origem são os subprodutos da atividade biológica e restos de plantas e animais. A concentração de COS é medida em gramas por quilograma (g/kg) da fração fina do solo (diâmetro esférico equivalente < 2 mm) seca em estufa (105°C). Ela pode ser quantificada utilizando diversos métodos analíticos. Esses métodos incluem procedimentos de química úmida (oxidação com cromo em meio ácido) e seca (combustão ou ignição por elevação da temperatura) e, mais recentemente, métodos espectrométricos baseados na reflectância difusa do solo. Fontes de carbono inorgânico como o carbonato de cálcio presentes numa amostra de solo são removidos no início da análise e, assim, não são incluídos na quantificação do COS.

Os métodos de quantificação da concentração do COS possuem eficiências distintas e podem produzir sub ou superestimativas sistemáticas. A harmonização de dados da COS requer medições pareadas usando metodologias analíticas distintas para construir modelos de regressão. Essas medições pareadas precisam ser acompanhadas por dados de outras variáveis pedológicas e ambientais que sejam capazes de explicar os erros observados. Os dados disponíveis no SoilData, entretanto, são carentes de medições pareadas da concentração do COS e, por isso, nenhuma estratégia de harmonização é implementada na presente versão.

3.1.2 Proporção das frações fina e grossa

A fração fina do solo inclui os constituintes minerais e orgânicos com diâmetro esférico equivalente menor que dois milímetros. É na fração fina que encontramos praticamente todo o COS. Já a fração grossa inclui os constituintes minerais do solo com diâmetro esférico equivalente igual ou maior que dois milímetros e conhecidos como cascalho (2 a 20 cm) ou calhau (20 a 200 cm). Juntos, cascalho e calhau compõem a fração grossa, também chamada de esqueleto. Em termos de estoques de COS, se assume que a participação da fração grossa é irrisória, sendo ignorada nesta versão beta.

As proporções das frações fina e grossa no solo quase sempre são quantificadas. Contudo, em muitos trabalhos, essas informações são omitidas ou registradas de maneira incompleta ou incerta. Em princípio, seria possível utilizar modelos de regressão para estimar essas proporções a partir de outras variáveis pedológicas e ambientais, as chamadas funções de pedotransferência. Contudo, testes preliminares mostraram que, para os dados disponíveis no SoilData, o desempenho desses modelos fica abaixo do aceitável, resultando em demasiada incerteza. Por isso, nesta versão, as camadas sem dados da proporção das frações fina ou grossa foram descartadas.

No cálculo dos estoques de COS, a proporção da fração grossa é usada para definir o volume do solo inteiro que é ocupado pela fração tida como capaz de estocar COS (fração fina). Em outros termos, a proporção volumétrica da fração grossa atua como um fator de correção. Contudo, na virtual totalidade dos casos, as frações fina e grossa são medidas como proporções gravimétricas, isto é, em gramas por quilograma (g/kg) de solo inteiro. Nesta versão, o conteúdo gravimétrico das frações fina e grossa são convertidos para volume (cm^3/cm^3) assumindo uma densidade de partículas finas e fragmentos grossos constante e uniforme de 2,65 g/cm^3.

3.1.3 Densidade do solo (DS) inteiro

A densidade do solo (DS) inteiro, medida em megagramas por metro cúbico (Mg/m^3), é a razão entre a massa dos constituintes minerais e orgânicos do solo estruturados, como encontrado no campo, e o volume por eles ocupado. No cálculo do estoque de COS, a DS serve para transformar a medida de concentração gravimétrica de COS para a base

volumétrica. Essa transformação desconta o volume do solo que não é ocupado pela fração fina e sim por poros preenchidos com ar ou água.

Assim, como a concentração de COS, a DS pode ser quantificada usando métodos diretos ou indiretos distintos. A escolha do método costuma depender das características do solo sendo amostrado, especialmente a presença e proporção relativa da fração grossa. Os resultados obtidos por cada método não necessariamente são equivalentes, demandando alguma estratégia de harmonização. Contudo, como no caso dos dados de COS, os procedimentos adotados nesta versão não incorporam a harmonização dos dados de DS em função da indisponibilidade dos dados e informações necessários.

A determinação da DS é laboriosa e requer muito cuidado, independente do método utilizado. A maior dificuldade reside na necessidade de obter amostras de solo com a estrutura preservada e fazer seu transporte para o laboratório sem alterações. Esse tipo de amostra é especialmente difícil de obter em solos com fragmentos grossos ou cobertos com vegetação com muitas raízes e raízes de grande diâmetro. Tais dificuldades resultam na ausência de dados de DS na maioria dos estudos, característica refletida no SoilData. Contudo, a DS costuma ter forte correlação com outras variáveis pedológicas e ambientais que são de mensuração mais fácil ou comum (Palladino et al., 2022). Assim, para evitar o descarte de dados, os dados faltantes de DS de 34.589 camadas iniciando antes de 30 cm de profundidade do solo foram estimados usando um modelo de regressão do tipo Random Forest (Tabela 1).

Tabela 1. Parâmetros do modelo de regressão Random Forest treinado para estimar os dados faltantes de densidade do solo inteiro de 34.589 camadas iniciando antes de 30 cm de profundidade do solo.

Parâmetro	**Valor/Descrição**
Software	Package ranger version 0.14.1
Number of trees	755 = (n * 0.25)
Sample size (n)	3017
Number of independent variables (p)	84
Mtry	$9 = (p^{-0.5})$
Target node size	5
Variable importance mode	impurity
Splitrule	variance

O modelo de regressão consistiu de 755 árvores de regressão treinadas usando n = 3017 amostras de treinamento e p = 84 variáveis preditoras. A cada divisão em cada árvore de

regressão, nove variáveis preditoras foram selecionadas aleatoriamente como candidatas a serem usadas como separadoras (*Mtry*). O objetivo foi reduzir a correlação entre as árvores e assim tornar o modelo mais robusto. As divisões realizadas em cada árvore de regressão foram determinadas pela redução da variância nos dados (*Splitrule*). Assim, a importância de uma variável preditora foi definida pelo grau de redução da variância que ocorreu com a sua inclusão na divisão da árvore. O processo de divisão de uma árvore foi interrompido quando o número de amostras de treinamento em um nó foi menor que cinco (*Target node size*).

As p = 84 variáveis preditoras utilizadas no modelo de regressão consistem de covariáveis pedológicas e ambientais. As covariáveis pedológicas são variáveis preditoras numéricas e categóricas (bi e multivariadas) criadas a partir de dados de propriedades do solo (concentração de argila, classificação do solo, entre outros) disponíveis no SoilData. As covariáveis ambientais são variáveis preditoras extraídas de mapas de propriedades do solo e outras informações ambientais. Duas fontes de covariáveis ambientais foram usadas, ambas disponíveis no Google Earth Engine: SoilGrids e MapBiomas. Seus dados foram amostrados para 12.139 pontos (correspondentes a 19.141 camadas) que continham coordenadas espaciais.

O SoilGrids 250m v2.0 é uma coleção de mapas globais de propriedades do solo disponíveis para seis intervalos de profundidade, três dos quais foram usados para treinar o modelo de regressão: 0-5, 5-15 e 15-30 cm. As propriedades do solo utilizadas são concentração de argila, areia, COS e fragmentos grossos e densidade do solo inteiro. Já o MapBiomas oferece séries temporais de mapas de cobertura e uso da terra e temas relacionados para todo o território brasileiro. Os dados usados para treinar o modelo de regressão foram aqueles de cobertura e uso da terra para o período de 1985 a 2021. Após amostrar esses mapas, foi identificada e mantida para cada ponto de treinamento a classe de cobertura e uso da terra do ano em que o mesmo foi coletado no campo.

A segunda etapa da preparação das covariáveis pedológicas e ambientais foi o tratamento dos dados ausentes (NAs). Covariáveis ambientais apresentam dados ausentes em amostras de treinamento sem coordenadas espaciais e pontos que caem em vazios de dados dessas covariáveis (áreas urbanas, corpos d'água etc). Para variáveis categóricas como a classe de solo e a classe de cobertura e uso da terra, uma nova categoria chamada "UNKNOWN" foi criada para substituir os NAs. Para covariáveis numéricas como o pH e a capacidade de troca de cátions (CTC), três novas variáveis foram geradas para substituir a original:

1. Uma variável numérica onde os NAs são substituídos por "+Inf".
2. Uma variável numérica onde os NAs são substituídos por "-Inf".
3. Uma variável categórica com duas categorias, "ISNA" e "ISNOTNA", indicando se os dados de uma amostra são ausentes ou não.

A abordagem usada para tratar NAs é baseada no conceito de incorporar nas próprias covariáveis a ausência de dados (Twala et al., 2008). A exceção foram os dados de

concentração de COS de algumas camadas, as quais passaram por uma etapa preliminar de imputação utilizando a concentração média do tipo de horizonte ou camada do solo.

3.1.4 Profundidade e espessura da camada

Dados de propriedades do solo estão disponíveis no SoilData para camadas de diferentes espessuras. Essa variação acontece porque a camada de interesse varia de caso para caso. Em trabalhos de cunho pedológico, a espessura das camadas amostradas é definida em função da observação de características que evidenciem a atuação de processos de formação do solo. Assim, a espessura dessas camadas, denominadas de horizontes, pode variar entre dois e 50 cm, valor máximo de espessura recomendado em manuais de descrição e coleta de solo no campo.

Já em trabalhos de cunho edafológico, voltados à análise das relações entre o solo e uma ou mais das esferas terrestres (biosfera, atmosfera e hidrosfera), geralmente são amostradas camadas de solo de espessura fixa predeterminada. Contudo, a amostragem costuma ser limitada aos primeiros 10 ou 20 cm a partir da superfície do solo. Trata-se da zona com maior interação com as demais esferas terrestres, especialmente a biosfera, sendo ocupada pela maior parte do sistema radicular das plantas, contendo grande população de micro e macro organismos, sendo a camada mais afetada pelas mudanças de uso e cobertura da terra e do clima.

Além disso, em parte considerável do território nacional, o solo possui menos de 30 cm de profundidade. São solos classificados como Neossolo Litólico, muito comuns em áreas de relevo acidentado e nas adjacências de afloramentos rochosos. Amostras coletadas nesses locais, mesmo que abrangendo toda a extensão vertical do solo, terão espessura variável e muitas vezes menor do que 30 cm.

Nesta versão beta, a variação na espessura e profundidade das camadas superficiais do solo são ignoradas. Todas as amostras de solo coletadas entre zero e 30 cm de profundidade, mesmo que abrangendo apenas uma parte dessa camada, seja por limitações pedológicas ou experimentais, são utilizadas como são, sem a realização de qualquer tipo de transformação dos dados para uma profundidade padrão.

3.1.5 Cálculo do estoque de COS

A quantidade (massa) de COS estocada numa camada de solo num ponto foi calculada usando a seguinte equação:

$$ESTOQUE = CARBONO \times (1 - FG) \times DS \times ESP$$

em que:

- CARBONO é a concentração (g/kg) de carbono orgânico na fração terra fina (partículas de diâmetro ≤ 2 mm) seca em estufa (105 °C),

- FG é a proporção (volume) da massa do solo ocupada por fragmentos minerais de diâmetro maior do que dois milímetros (fração grossa),
- DS é a densidade (g/cm^3) do solo inteiro e
- ESP é a espessura (m) da camada de solo sendo considerada nos cálculos.

3.2 Covariáveis ambientais

As covariáveis ambientais espacialmente explícitas são utilizadas para predizer os estoques de COS em locais e momentos não amostrados. Essas covariáveis representam os fatores de formação do solo. Os fatores definem o solo em termos de controles da pedogênese, processos pedogenéticos e distribuição do solo na paisagem. Cada fator tem um determinado impacto sobre o solo através de uma variedade de processos físicos, químicos e biológicos.

Dentre os principais fatores de formação do solo que podem ser representados estão o clima, organismos, relevo e material parental. O clima pode ser representado por mapas de classificação climática. Mapas de cobertura e uso da terra, imagens de satélite e índices de vegetação podem ser usados para representar os organismos – ou mesmo o material parental, no caso do COS. Já o relevo pode ser representado por modelos digitais de elevação e índices morfométricos derivados. Outra fonte de dados espacialmente exaustivos utilizados como covariáveis são os mapas existentes de propriedades do solo, que podem indicar padrões espaciais de variação da propriedade de interesse e probabilidade de ocorrência de solos, que indicam as chances globais de ocorrência das classes taxonômicas. Essas covariáveis são úteis porque as propriedades do solo costumam ser correlacionadas.

Para a coleção beta, foram levantadas de bancos de dados espaciais abertos mais de uma centena de dados ambientais espacialmente explícitos para serem utilizados como covariáveis ambientais. A inserção destas no modelo preditivo se deu de modo gradual, com o intuito de avaliar sua influência nos mapas resultantes. Um total de 86 covariáveis foram incluídas na versão beta, selecionadas com base na plausibilidade de sua relação com a dinâmica espaço-temporal dos estoques de COS.

As covariáveis foram utilizadas na modelagem considerando a cobertura temporal dos dados, podendo ela ser estática (um único dado para toda série) ou dinâmica (resolução anual) (Tabela 2). Dentre as variáveis estáticas estão as variáveis morfométricas do relevo, derivados de modelos digitais de elevação. Além delas, mapas de propriedades do solo também foram utilizados para representar padrões espaciais de larga escala. Estes mapas, produzidos por iniciativas como SoilGrids, por ignorarem a dimensão temporal, também foram usados como variáveis estáticas. As covariáveis são apresentadas nos próximos itens. Os dados e códigos utilizados são de acesso aberto e podem ser consultados no material suplementar deste documento.

Tabela 2. Conjunto de covariáveis estáticas e dinâmicas utilizadas na modelagem conforme o fator de formação do solo que elas representam.

Fator preditivo	**Covariáveis**	**Tipo de variável**	**Resolução espacial (original)**	**Dimensão temporal**
Solo (s)	Probabilidade de ocorrência de classes ou tipos de solo	contínua	250 m	estática
	Propriedades do solo	contínua	250 m	estática
Clima (c)	Classificação do clima	categórica nominal, transformada para binária	-	estática
Organismos (o)	Classificação da vegetação primária (fitofisionomia)	categórica nominal, transformada para binária	-	estática
	Classificação da cobertura e uso da terra	categórica nominal, transformada para binária	-	dinâmica
	Classificação territorial (bioma)	categórica nominal, transformada para binária	-	estática
	Estoque de carbono em compartimentos terrestres	contínua	-	estática
	Índices de vegetação	contínua	30 m	dinâmica
Relevo (r)	Propriedades morfométricas do terreno	contínua	90 m	estática
Material parental (p)	-	-	-	
Idade (a)	Idade do tipo de cobertura e uso da terra	contínua	30 m	dinâmica
Posição espacial (n)	Coordenadas geográficas	contínua		estática

3.2.1 Covariáveis ambientais estáticas

3.2.1.1 Propriedades e classes de solo

A utilização de mapas existentes como covariáveis explora a correlação existente entre as próprias propriedades do solo. São mapas do conteúdo de argila, capacidade de troca de cátions e de classes taxonômicas. Na coleção beta, utilizamos os mapas do solo do conteúdo de argila, areia, silte, carbono, nitrogênio total, fragmentos grossos e capacidade de troca de cátions, pH em água e densidade do solo produzidos pelo SoilGrids 2.0 (Poggio et al., 2021). Como o objetivo foi predizer o estoque de COS para até 30 cm e os mapas disponíveis foram produzidos para as camadas 0-5, 5-15 e 15-30 cm, foi calculada média ponderada dos dados em cada pixel usando a seguinte equação:

$$mapa_{0\text{-}30\ cm} = mapa_{0\text{-}5\ cm} \times (1/6) + mapa_{5\text{-}15\ cm} \times (2/6) + mapa_{15\text{-}30\ cm} \times (3/6)$$

A resolução original dos dados de 250 m foi degradada para 1 km utilizando o método do vizinho mais próximo como forma de remover artefatos indesejados. Muitas vezes, a informação mais grossa da paisagem explica melhor a variação espacial do solo do que a resoluções mais finas, especialmente quando o número de amostras de treinamento é muito pequeno (Costa et al., 2018; Samuel-Rosa et al., 2015). Em seguida, os dados foram interpolados utilizando o método do vizinho mais próximo para a resolução espacial de 30 metros para alcançar a padronização dos dados de treinamento e preencher vazios, especialmente nos limites do território brasileiro e próximo de manchas de áreas urbanas e corpos d'água.

Os mapas de probabilidade de ocorrência das classes de solo do Word Reference Based for Soil Resources (WRB) (FAO, 2014) produzidos pelo SoilGrids 2.0 foram adicionados à lista de covariáveis estáticas. Dentre estes mapas estão, por exemplo, a probabilidade de ocorrência de solos orgânicos (Histosols), que apresentam uma camada de pelo menos 40 cm de material predominantemente orgânico.

Foram somadas as probabilidades de ocorrência de classes de solos com características pedológicas semelhantes que estejam correlacionadas com os estoques de carbono. A soma da probabilidade de ocorrência das classes Chernozems, Kastanozems, Phaeozems e Umbrisols deram origem a covariável Humisols, que representa a probabilidade de encontrarmos solos com alto teor de COS na camada superficial. A soma da probabilidade de ocorrência das classes Arenosols e Podzols gerou à covariável Sandysols, que representa a probabilidade de ocorrência de solos com horizonte superficial arenoso e, portanto, espera-se uma correlação negativa com os estoques de COS. A soma da probabilidade de ocorrência das classes Leptosols e Regosols gerou a covariável Thinsols, que representa a probabilidade de ocorrência de solos rasos. Solos de outras áreas úmidas, cujas classes são Gleisols, Planosols e Stagnosols, foram agrupadas em uma única covariável denominada de

Wetsols. A tabela 3 demonstra as covariáveis estáticas de classes ou propriedades do solo e sua relação presumida com os estoques de COS.

Tabela 3. Covariáveis estáticas de classes ou propriedades do solo e sua relação presumida com os estoques de COS.

Covariavéis	**Descrição (relação com os estoques de COS)**
Conteúdo de argila	A argila é responsável pela proteção físico-química da matéria orgânica no solo, reduzindo sua taxa de decomposição. Espera-se que solos argilosos possuam maior estoque de carbono em comparação a solos arenosos.
Conteúdo de silte	O silte, por ser uma partícula fina, associado à argila, auxilia na proteção físico-química do COS.
Conteúdo de areia	A areia possui papel importante na retenção de carbono quando as partículas mais finas, como silte e argila, já estão saturadas de matéria orgânica. É possível encontrar até 40% do COS na fração grossa do solo.
Capacidade de troca de cátions	A capacidade de troca de cátions é uma medida da quantidade de cargas elétricas negativas no solo. A matéria orgânica é uma das principais fontes dessas cargas no solo. Logo, espera-se que, quanto maior for a CTC, maior seja o estoque de carbono no solo.
pH em água	A humificação da matéria orgânica no solo produz ácidos orgânicos, o que reduz o pH do solo. Isso é especialmente importante em regiões tropicais (Amazônia) e temperadas (campos de altitude) úmidas, espera-se que o menor pH do solo tenha correlação positiva com o conteúdo de carbono.
Densidade do solo	A densidade do solo também varia com o conteúdo de matéria orgânica no solo: quanto mais matéria orgânica menor será a densidade dele.
Conteúdo de carbono	O conteúdo de carbono no solo refere-se à quantidade de COS.
Conteúdo de nitrogênio total	Relação C/N Transformaram-se os valores de centigrama por quilograma para grama por quilograma
Conteúdo de fragmentos grossos	Fragmentos grossos de solo referem-se a pedaços de material mineral com mais de 2 mm de diâmetro que estão presentes no solo. A presença de fragmentos grossos no solo pode afetar sua fertilidade e estrutura, uma vez que esses fragmentos podem dificultar o desenvolvimento das raízes das plantas e a penetração de água e nutrientes no solo.

(Continua)

Tabela 3: Covariáveis estáticas de classes ou propriedades do solo e sua relação presumida com os estoques de COS. (Continuação)

Covariavéis	**Descrição (relação com os estoques de COS)**
Probabilidade de ocorrência de Sandysols	Espera-se que o estoque de carbono seja baixo em solos arenosos em função da ausência do efeito de proteção física da matéria orgânica.
Probabilidade de ocorrência de Humisols	Refere-se a solos com elevado conteúdo de carbono orgânico nas camadas superficiais.
Probabilidade de ocorrência da classe de solo Ferralsols	Refere-se a solos com elevado teor de óxidos de ferro e alumínio na camada subsuperficial
Probabilidade de ocorrência da classe de solo Histosols	Refere-se a solos composto principalmente de materiais orgânicos.
Probabilidade de ocorrência de Thinsols	Thinsols são solos caracterizados por uma camada superficial com menos de 10 cm de espessura.
Probabilidade de ocorrência de Wetsols	Refere-se a Solos úmidos, que apresentam umidade em excesso, seja na superfície ou em camadas mais profundas, apresentam características, como coloração escura, devido ao alto teor de matéria orgânica, textura argilosa e grande capacidade de retenção de água.
Óxidos de Ferro	Os óxidos de ferro são compostos químicos que contêm ferro e oxigênio em diferentes proporções, os mesmo interagem com o carbono no solo e promove interações importantes referente aos ciclos biogeoquímicos no solo.
Argilominerais	Os argilominerais têm um papel importante na fertilidade do solo, pois podem reter nutrientes e água em suas superfícies, tornando-os disponíveis para as plantas. Além disso, eles podem afetar a permeabilidade do solo e sua capacidade de reter água, o que é crucial para o desenvolvimento das raízes das plantas.
Carbono acima do solo	O carbono acima do solo refere-se ao carbono armazenado nas partes aéreas das plantas, como troncos, galhos, folhas e frutos, bem como nos resíduos vegetais depositados na superfície do solo

(Continua)

Tabela 3. Covariáveis estáticas de classes ou propriedades do solo e sua relação presumida com os estoques de COS. (Continuação)

Covariavéis	Descrição (relação com os estoques de COS)
Carbono abaixo do solo	Biomassa subterrânea refere-se à matéria orgânica que é encontrada abaixo da superfície do solo ou da água. Isso pode incluir raízes, tubérculos, rizomas, bulbos, entre outros tipos de tecidos vegetais que se desenvolvem abaixo do solo.
Madeira morta	Madeira morta são árvores ou partes de árvores que morreram, mas que permanecem em pé, geralmente devido a condições ambientais específicas ou ações de agentes biológicos.
Serapilheira	É a camada de matéria orgânica morta, composta por folhas, galhos, frutos, flores e outros detritos de plantas que se acumulam sobre o solo. Essa camada é importante porque atua como um importante recurso para a vida do solo e para a saúde do ecossistema.
Carbono Total	É o consolidado dos valores de estoque de carbono para os diferentes compartimentos: carbono acima do solo, carbono abaixo do solo, madeira morta e serapilheira.

O mapa de probabilidade de ocorrência de solos pretos (*black soils*) também foi usado como covariável (FAO, 2022). Solos pretos são solos minerais que têm um horizonte superficial escuro, enriquecido com COS com pelo menos 25 cm de profundidade.

Além disso, foram incluídos os mapas de carbono acima do solo, carbono abaixo do solo, madeira morta e serapilheira produzidos pelo Quarto Inventário Nacional de Emissões e Remoções Antrópicas de Gases de Efeito Estufa, que integra a Quarta Comunicação Nacional do Brasil à Convenção-Quadro das Nações Unidas sobre Mudança do Clima (Brasil, 2021) e pós-processados pelo Sistema de Estimativas de Emissões de Gases de Efeito Estufa (SEEG) (SEEG, 2022)

3.2.1.2 Morfometria de terreno

Diversos modelos digitais de elevação estão disponíveis para representar o relevo do território brasileiro (Figura 2). Sua cobertura temporal, entretanto, é esparsa, limitando-se a algumas poucas datas. A menor disponibilidade de dados de elevação da superfície do solo, comparada a imagem satelitais, se deve à dinâmica temporal menos pronunciada dessas superfícies. Séries temporais de modelos digitais de elevação da superfície do solo de altíssima precisão permitem quantificar a erosão hídrica e eólica. O processo erosivo é um dos principais responsáveis pelas perdas de carbono em áreas sob cultivo, com revolvimento do solo e em pastagens mal manejadas. Por outro lado, a erosão favorece o enriquecimento dos solos das áreas de várzea, com carbono, por meio da deposição do material transportado das áreas mais altas da paisagem.

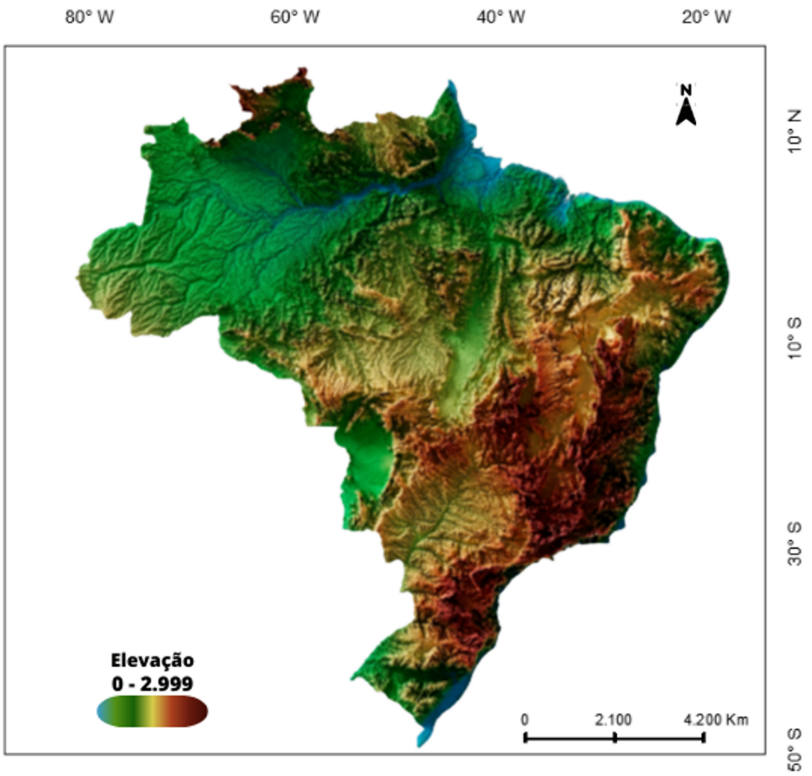

Figura 2. Mapa de elevação do Brasil Fonte: Adaptado de Alessandra Souza, 2020.

A solução mais razoável para utilizar covariáveis morfométricas no mapeamento espaço-temporal dos estoques de carbono no solo é ignorar a variação temporal do relevo. Contudo, precisam ser computadas covariáveis morfométricas estáticas que descrevem o contexto ambiental. Na coleção beta, foram utilizados os dados de elevação do modelo digital de elevação SRTM, a partir do qual foram derivadas 26 variáveis morfométricas na resolução espacial de 90 metros (Amatulli et al., 2020).

As covariáveis morfométricas foram processadas com a aplicação de preenchimento de vazios e depois reamostradas de 90 m para 30 metros usando o método do vizinho mais próximo. A seguir serão apresentados as covariáveis geomorfométricas utilizadas na versão beta, bem como sua importância para a estruturação do modelo (Tabela 4).

Tabela 4. Lista de covariáveis estáticas morfométricas e sua relação presumida com os estoques de COS.

Covariável	**Descrição (relação com os estoques de COS)**
Índice Topográfico Composto [-4; 11] * 10	Logaritmo da área de captação cumulativa a montante dividido pela tangente do local. Destaca os locais com maior potencial de acúmulo de água na paisagem. Áreas úmidas devem acumular o COS trazido das áreas elevadas por erosão. A saturação do solo reduz a decomposição do COS, promovendo seu acúmulo.
Índice de exposição a leste [-0,7; 0,6] * 100	O leste fornece medidas contínuas que descrevem a orientação em combinação com a inclinação. Tem sido usados na distribuição de espécies de plantas e mapeamento de florestas.
Índice de exposição a norte [-0,6; 0,6] * 100	O norte fornece medidas contínuas que descrevem a orientação em combinação com a inclinação. Tem sido usado na distribuição de espécies de plantas e mapeamento de florestas.
Elevação [0; 2647] m	Reprocessamento de dados STRM. A temperatura do ar diminui com a elevação, reduzindo a atividade microbiana. Deve haver acúmulo de COS e aumento dos estoques.
Índice de potência do fluxo [0; 8279] log10(spi +1) * 100 [0; 8279]	Reflete a potência erosiva associada ao fluxo e a tendência das forças gravitacionais de mover a água a jusante. Usado em modelos de erosão do solo, suscetibilidade a deslizamentos e estimativa de águas subterrâneas. Pode indicar perda de COS pelo escoamento superficial erosivo.
Declividade [0; 48] graus → % tan(3.141593/180*degrees)) *100 [0; 111] %	É a taxa de mudança de elevação na direção da linha de fluxo de água. É especialmente importante para a quantificação da erosão do solo, velocidade do fluxo de água ou aptidão agrícola.
Rugosidade [0; 316]	É expressa como a maior diferença absoluta inter-células de uma célula focal e suas oito células circundantes.
Convergência [-82; 81]	É uma variável do terreno que destaca as áreas convergentes como canais e áreas divergentes como cumes.
Curvatura do perfil (vertical) [-0,004; 0,001] * 10.000	A análise das curvaturas permite entender como a água e como os sedimentos se movem na paisagem e ajuda a quantificar sua acumulação ou dispersão.

3.2.1.3 Classificação climática

O clima, enquanto fator de formação, refere-se às condições de disponibilidade de água, ar e calor no interior e na atmosfera próxima da superfície do solo. Essas condições determinam a atividade biológica num local, regulando a taxa de acúmulo de fotoassimilados pela vegetação e sua respiração e da microbiota do solo. Dessa forma, os estoques de COS variam de uma região para a outra dependendo não apenas do uso e cobertura locais do momento atual, mas também das condições dos anos anteriores. Por isso, o clima atua como regulador das entradas e saídas de COS.

Dados espacialmente explícitos das condições meteorológicas diárias estão disponíveis para todo o território brasileiro cobrindo um longo período de tempo. Esses dados consistem de mapas produzidos utilizando dados de estações meteorológicas e sensores orbitais. Na versão beta, entretanto, optou-se por utilizar apenas dados de classificação climática como covariável, os quais auxiliam na modelagem do efeito dos padrões climáticos espaciais de larga escala sobre os estoques de COS. Para isso, a classificação climática de Köppen, considerada a primeira classificação quantitativa quanto ao clima das regiões de todo o mundo, foi utilizada (Figura 3) (Alvares et al., 2013).

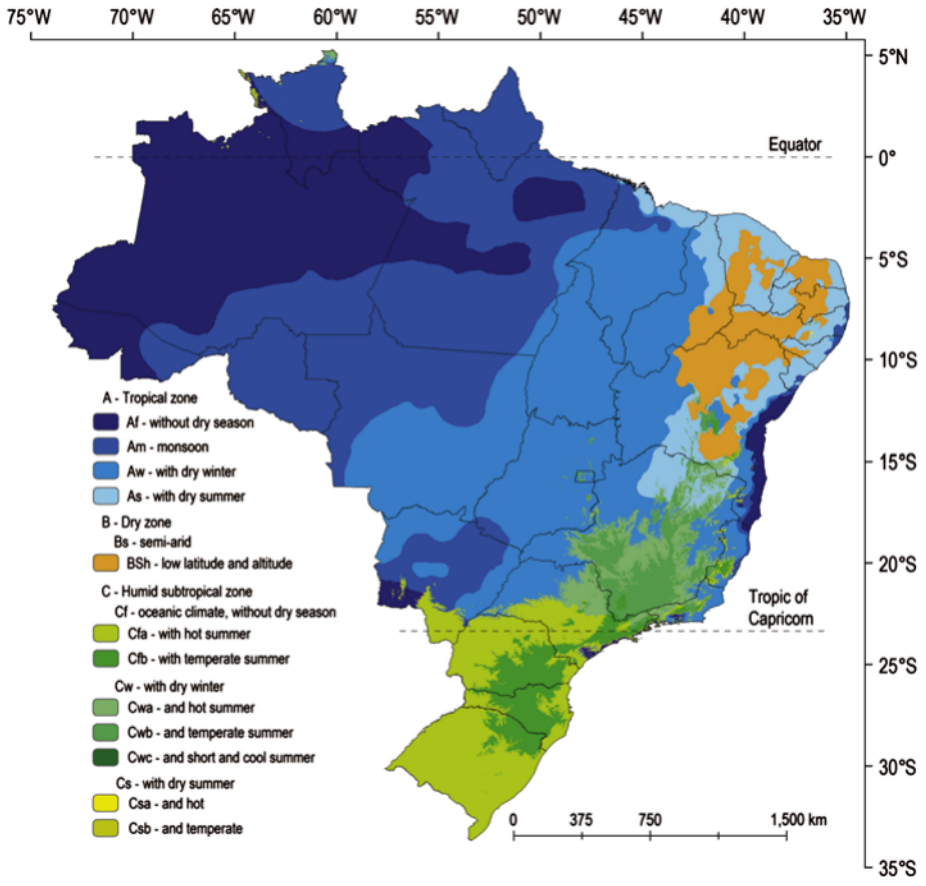

Figura 3. Classificação climática para o Brasil de acordo com os critérios de Köppen (1936). Fonte: (Alvares et al., 2013).

Por se tratar de dados multicategóricos, a classificação climática foi transformada para variáveis binárias (0 ou 1) antes de serem incluídas no modelo de regressão, conforme exemplificado na Tabela 5. Assim, a classificação climática de Köppen resultou em três covariáveis binárias para modelagem dos estoques de COS.

Tabela 5. Covariáveis binárias e equivalência com classes climáticas.

Covariável	Classe climática e valor recebido		
	Cfa	Cfb	Cwa
Cfa	1	0	0
Cfb	0	1	0
Cwa	0	0	1

3.2.1.4 Classificação da vegetação

Cada bioma possui características distintas e uma relação diferente com os estoques de COS. Nesse sentido foram utilizados dados da classificação do território brasileiro em biomas (Figura 4a) e da classificação da vegetação (fitofisionomia) (Figura 4b). Assim como para a classificação climática, essas covariáveis multicategóricas foram convertidas em covariáveis binárias antes de serem incluídas no modelo de regressão.

a) b)

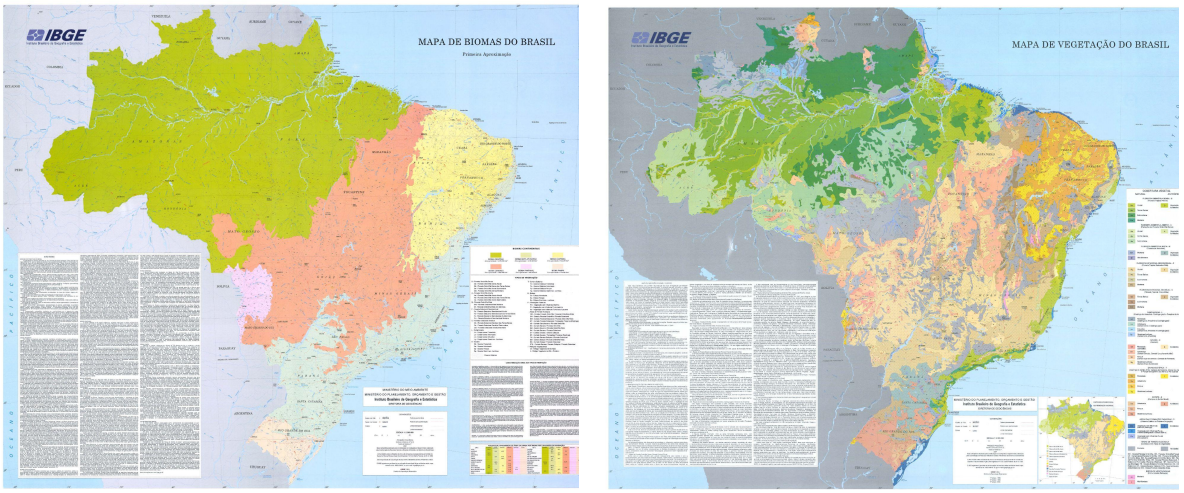

Figura 4. Mapa de Biomas (a) e da vegetação do Brasil (b). Fonte: IBGE, 2004.

3.2.2 Covariáveis ambientais dinâmicas

Os organismos vivos operam as entradas e saídas de COS no solo. Mudanças na cobertura vegetal e uso das terras podem resultar tanto em ganhos como perdas na concentração de COS. Essas mudanças também promovem alterações na densidade do solo e consequentemente nos estoques de COS. A redução da concentração de COS e da biomassa vegetal sobre esse, associadas a práticas de manejo não conservacionistas, aumentam a susceptibilidade do solo à erosão. Por isso, os mapas de cobertura e uso da terra são fundamentais para mapear a variação espaço-temporal dos estoques de COS.

3.2.2.1 Uso e cobertura da terra

A série de mapas de cobertura e uso da terra do MapBiomas (Coleção 7.1) foi utilizada para representar os efeitos dos organismos vegetais (Figura 5). Estas variáveis foram consideradas como covariáveis dinâmicas no modelo, com resolução temporal de um ano. As seguintes classes foram incluídas:

- Formação Florestal
- Formação Savânica
- Formação Campestre
- Campo Alagado e Área Pantanosa
- Pastagem
- Lavouras Temporárias e Perenes
- Silvicultura
- Mosaico de Agricultura
- Pastagem
- Outras Formações Naturais

As classes Mangue, Restinga Arborizada, Apicum e Outras Formações não Florestais foram agrupadas na classe “Outras Formações Naturais”. Na Figura 5 são apresentados os mapas de uso e cobertura da terra do ano de início (1985) e final (2021) da série, produzidos pelo MapBiomas (Coleção 7.1).

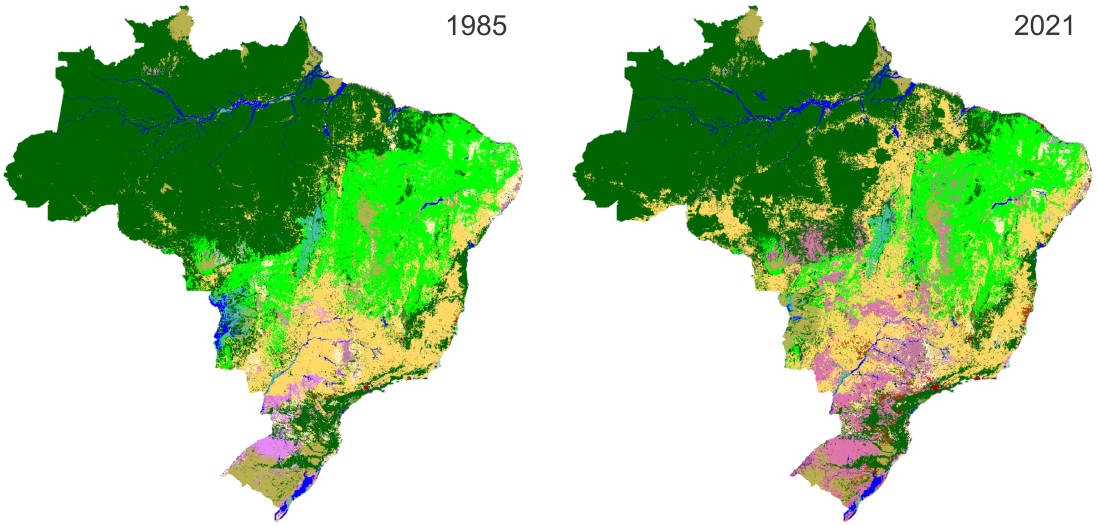

Figura 5. Mapas de uso e cobertura da terra dos anos de 1985 e 2021. Fonte: MapBiomas (Coleção 7.1).

Por se tratar de dados multicategóricos, os dados da classificação do uso e cobertura da terra foram transformados para covariáveis binárias antes de serem incluídos no modelo de regressão. Assim, uma covariável representa a presença ou ausência de uma dada classe de uso e cobertura para cada ano da série (Tabela 6).

Tabela 6. Exemplo de equivalência entre classes de uso e cobertura e a indicação binária de presença para cada pixel, em cada ano da série.

Classe de uso e cobertura da terra	1985	1986	1987
Cobertura florestal	1	0	0
Lavoura	0	1	1
Floresta	0	0	0

3.2.2.2 Persistência do uso e cobertura

O estoque de COS para um determinado ano não depende apenas da vegetação do mesmo ano, mas também daquela de anos anteriores, tendo em vista que a mudança da vegetação tem um efeito retardado na mudança no COS. Para melhorar a representação dos efeitos da vegetação nos estoques de COS ao longo do tempo, foi computado o histórico de uso e cobertura da terra em cada pixel. Esse cálculo resultou na medida de quanto tempo cada pixel esteve sob determinado uso ou cobertura. O resultado deste processamento foi uma série histórica da persistência do uso e cobertura da terra para o período de 1985-2021. O cálculo foi realizado conforme pseudocódigo explicitado abaixo.

```
JavaScript

LULC = Mapa de uso e cobertura mapbiomas
CLASS = Classe de uso e cobertura (2º nível)
ano = seq(from = 1985, to = 2021, by = 1)
x = valor do pixel

Para cada ano
      Para cada LULC
            Se LULC = CLASS,
                  Se ano = 1985 então,
                        pixel x = 20,
                  Senão ,
                        Se teve mudança de classe, então
                              Se ela durou dois anos, então
                                    x = 1,
                              Senão,
                                    x = x + 1,
                        Senão,
                              x = x + 1,
            Senão,
                 x = 0
      Fim.
Fim.
```

Em função da ausência de dados para o período anterior a 1985, assumimos que naquele ano todos os usos e coberturas da terra já haviam permanecido estáveis por 20 anos. Isso significa que, para o ano de 1985, todas as classes de uso e cobertura receberam um valor de permanência igual a 20.

3.2.2.3 Índices de vegetação

Três índices de vegetação foram utilizados como covariáveis dinâmicas:

- Índice de vegetação da diferença normalizada (NDVI)
- Índice de vegetação ajustado ao solo (SAVI)
- Índice de vegetação melhorado (EVI)

O NDVI foi calculado utilizando as imagens de satélites da coleção Landsat com a seguinte expressão:

NDVI = (NIR - RED) / (RED + NIR)

O SAVI leva em consideração o efeito do solo na reflectância da vegetação e é especialmente útil em áreas com vegetação densa e com solos muito escuros, onde outros índices de vegetação podem não ser tão precisos. O índice SAVI é calculado a partir da diferença entre

a reflectância do infravermelho próximo (NIR) e do vermelho (RED) das imagens de satélite, e é ajustado para levar em consideração o efeito do solo na reflectância da vegetação, com a seguinte expressão:

SAVI = ((NIR - RED) / (NIR + RED + L)) x (1 + L)

O EVI (Enhanced Vegetation Index) é um índice de vegetação que utiliza informações espectrais das imagens de satélite para estimar a densidade da vegetação em uma determinada área. O EVI é uma versão melhorada do NDVI (Normalized Difference Vegetation Index), que leva em conta o efeito do solo e das nuvens na reflectância da vegetação. O EVI é calculado a partir das bandas espectrais do vermelho, do infravermelho próximo e do azul das imagens de satélite, utilizando a seguinte fórmula:

EVI = 2.5 x ((NIR - RED) / (NIR + 6 x RED - 7.5 x BLUE + 1))

O valor resultante do EVI varia de -1 a 1, sendo que valores mais altos indicam uma vegetação mais densa e saudável. O índice é especialmente útil em áreas com vegetação densa e em regiões com presença de nuvens, que podem interferir na precisão de outros índices de vegetação.

Considerando a série temporal (1985 a 2021), os índices de vegetação foram calculados utilizando as imagens anuais de 1985 a 1999 e 2004 a 2011, do Landsat 5; imagens anuais de 2000 a 2003 e 2012 do Landsat 7; e imagens anuais de 2013 a 2021 do Landsat 8. Como algumas áreas do Brasil estão permanentemente cobertas por nuvens (dados faltantes), especialmente para o início da série histórica na região norte do país, os dados faltantes foram preenchidos utilizando uma estratégia de preenchimento baseada nos anos imediatamente anterior e posterior ao ano com dado faltante. Para minimizar os efeitos da utilização de dados provenientes de distintos sensores Landsat ao longo da série temporal, a tendência dos mesmos foi removida.

O estoque de COS para um determinado ano não depende apenas da vegetação desse mesmo ano, mas também depende muito dos anos anteriores. Por isso, os dados dos índices de vegetação foram utilizados para gerar médias ponderadas ao longo de vários anos, voltando no tempo. Essa estratégia já foi utilizada anteriormente para produzir os mapas de estoque de carbono (0-30 cm) para a Argentina, de 1982 a 2017 (Heuvelink et al., 2021). O decaimento exponencial ajuda a representar o decréscimo do efeito da vegetação pretérita ao longo de um período de tempo. Para isso, utilizamos uma função de decaimento exponencial, com alfa decaimento igual a 0,7 (Figura 6) que resultou em uma lista de pesos para os anos anteriores ao de interesse.

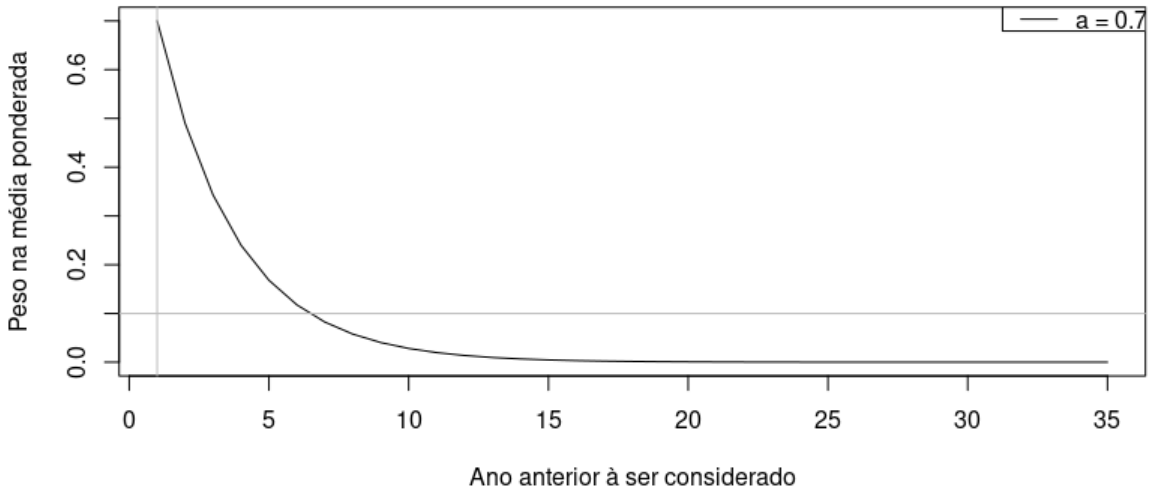

Figura 6. Decaimento exponencial para cálculo de peso dos índices de vegetação ao longo da série histórica

A aplicação da função de decaimento exponencial requer que as covariáveis estejam disponíveis para anos antes do ano inicial da série temporal. Frente a essa limitação, assumimos que o NDVI, EVI e SAVI para os anos anteriores a 1985 eram iguais a 1985. Para reduzir o tempo de computação, aplicamos um limite de 0,1. Isso significa que incluímos apenas contribuições do passado para as quais a função de decaimento exponencial está acima do limite, resultando em seis anos.

3.3 Modelo preditivo

O método de regressão adotado para predizer o estoque de COS foi o Random Forest. Um modelo Random Forest consiste, basicamente, de uma coleção de *M* árvores de regressão randomizadas. Isso significa que, internamente, o Random Forest opera construindo árvores de regressão.

Uma árvore de regressão é construída por meio de um processo iterativo de particionamento de dados, chamado de particionamento recursivo binário, que visa minimizar o erro quadrático médio em cada partição. O algoritmo começa selecionando a melhor forma de dividir ao meio o conjunto de dados de treinamento com base nos valores das covariáveis. O critério de seleção é que o erro quadrático médio do subconjunto a ser formado seja o menor possível, comparado à média dos dados. O algoritmo faz isso de maneira iterativa, formando uma estrutura semelhante a uma árvore. O processo continua, até que uma partição final seja gerada, a partir da qual outras divisões não reduzem o erro quadrático médio do subconjunto. A predição então é realizada com base na média dos dados do subconjunto após o último particionamento.

Como o objetivo do algoritmo é reduzir o erro quadrático médio, uma árvore de regressão pode dividir o conjunto de dados em um grande número de subconjuntos, até o ponto em

que um subconjunto contenha pouquíssimos dados. Como consequência, o modelo fica sobre-ajustado aos dados de treinamento e perde a sua capacidade de generalização (chamado de overfitting). Ou seja, ele se ajusta muito bem ao conjunto de dados de treinamento, mas pode ser ineficiente para fazer predições em dados desconhecidos, por fazer predições pouco viesadas mas com alta variância.

O Random Forest faz o ajuste de diversas árvores de regressão randomizadas, portanto, independentes entre si, usando o conjunto de dados de treinamento. Para isso, cada árvore de regressão do modelo é ajustada utilizando uma reamostragem com reposição dos dados de treinamento (usando o método *bootstrap*), e algumas covariáveis, o que resulta em árvores independentes umas das outras. Ao final, um valor predito pelo Random Forest é resultante da média da predição de todas as árvores de regressão em cada local. Com isso, ambos o viés e a variância das predições tendem a ser reduzidos, diminuindo a tendência de overfitting das árvores de regressão.

Para ajustar o modelo, dois parâmetros precisam ser definidos (Tabela 7). O primeiro é o número de árvores de regressão que serão ajustadas (parâmetro ntree) e o número de covariáveis preditoras que serão utilizadas em cada árvore (parâmetro mtry). Como hiperparâmetros do modelo foram definidos o número de árvores de regressão ntree = 1/10 do número de amostras e o número de covariáveis selecionadas em cada divisão mtry = 1/3 do número de covariáveis. O modelo foi treinado no Google Earth Engine com a função ee.Classifier.smileRandomForest com o modo de saída "*Regression*". Os códigos utilizados são de acesso aberto e podem ser consultados no material suplementar deste documento.

Tabela 7. Parâmetros do modelo Random Forest utilizado para mapeamento do estoque de COS no território brasileiro.

Parâmetro	**Descrição**	**Valor**
ntree	número de árvores de regressão	1000
mtry	número de covariáveis selecionadas em cada divisão das árvores	25
minLeafPopulation	número mínimo de observações nos nós terminais	5
bagFraction	fração das observações do conjunto de treinamento selecionadas aleatoriamente para compor o conjunto de avaliação	0,632

As coordenadas espaciais e temporais (ano de coleta) dos dados pontuais de solo foram utilizados para extrair as covariáveis ambientais estáticas dinâmicas, para compor a matriz

de treinamento. A matriz de treinamento foi utilizada para treinar um único modelo preditivo, utilizado para fazer a predição no espaço e tempo. Para isso, foi realizada uma predição espacial para cada ano de interesse (1985-2021), e toda vez, a camada contendo as covariáveis dinâmicas de interesse é inserida para indicar ao modelo uma referência de tempo. Todos os procedimentos, demonstrados no fluxograma (Figura 7) e detalhados nos itens a seguir, foram realizados no workspace MapBiomas no Google Earth Engine e Google Cloud Storage Platform.

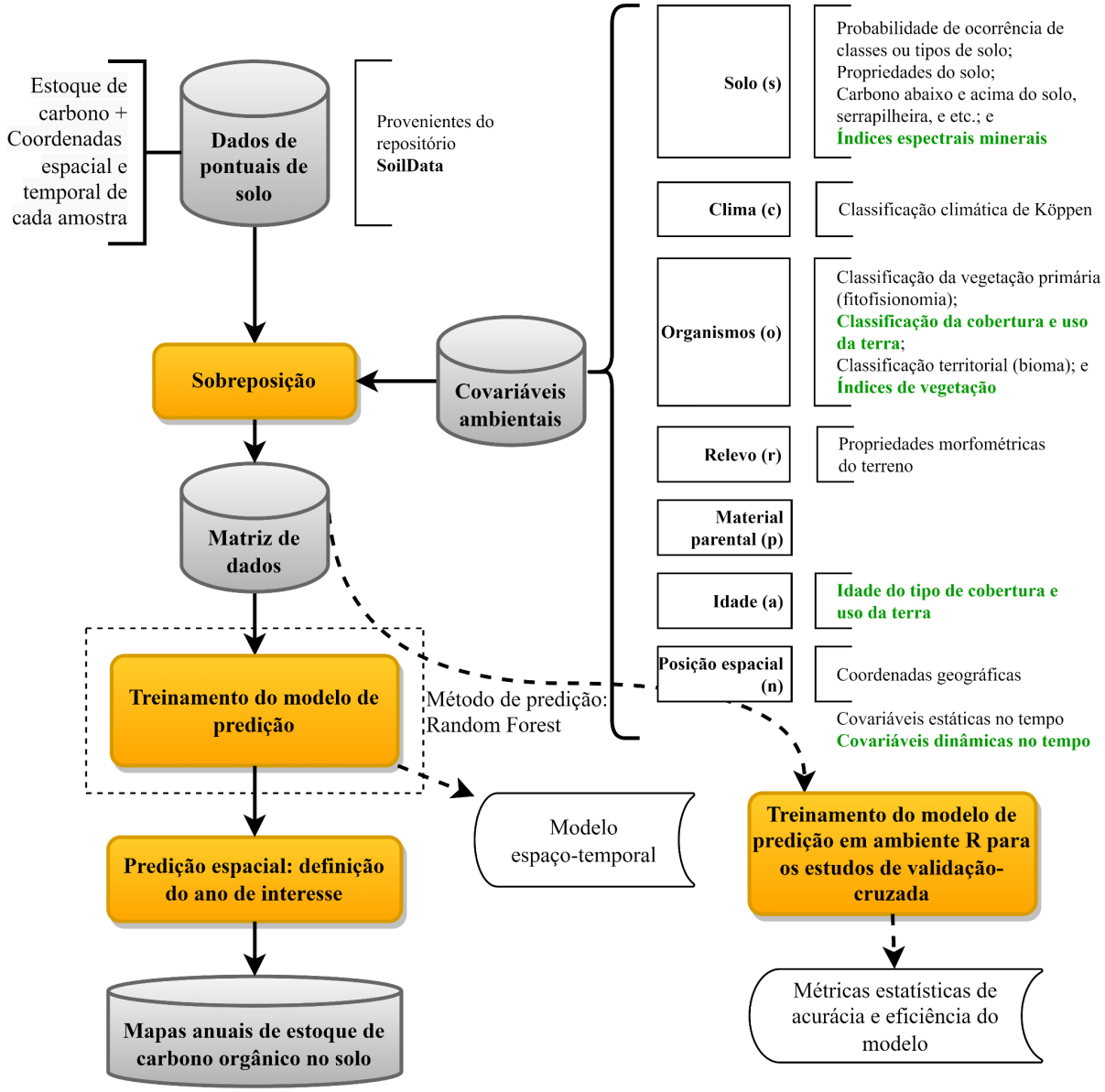

Figura 7. Fluxograma das etapas para a modelagem espaço-temporal dos estoques de carbono orgânico do solo no território brasileiro.

3.4 Pós-processamento

3.4.1 Máscaras

Os mapas de uso e cobertura da terra do MapBiomas (Coleção 7.1) foram usados para delinear as áreas de terras emersas a serem mapeadas ao longo da série temporal. A classe 33, que designa rios, lagos e oceanos, e a classe 31 referente a aquicultura foram utilizadas para essa finalidade. A união dessas classes foi utilizada para identificar as áreas com cobertura de água permanente (frequência maior que 35 anos) e áreas alagáveis, ou seja, aquelas que estiveram emersas em algum ano da série. As áreas de água permanente foram excluídas do universo de predição, recebendo o valor -1.

Algumas classes de cobertura e uso da terra são reconhecidas por possuírem estoque de COS muitíssimo baixo ou nulo. As classes usadas para delinear essas áreas estão listadas na Tabela 8. Para essas áreas, o estoque de COS foi considerado como sendo igual a zero.

Tabela 8. Classes de uso e cobertura da terra do MapBiomas (Coleção 7.1) com estoque de COS muitíssimo baixo ou nulo.

ID	Classes Coleção 7.1	Natural/Antrópico	Uso e Cobertura da Terra
23	4.1. Praia, duna e areal	Natural	Cobertura
24	4.2. Infraestrutura urbana	Antrópico	Uso
30	4.3. Mineração	Antrópico	Uso
29	2.4. Afloramento rochoso	Natural	Cobertura

3.4.2 Filtro temporal

Algumas covariáveis utilizadas nas predições apresentaram pixels com dados faltantes mesmo após o seu processamento inicial. Isso ocorreu com maior frequência em regiões como o Amapá, que possuem uma densa cobertura de nuvens ao longo do ano, dificultando a obtenção de imagens da superfície. Como resultado, alguns pixels foram retornados pelo modelo sem valor predito. Na etapa de pós-processamento, todos estes pixels com dados faltantes foram preenchidos usando um filtro temporal. Caso não houvesse informação de COS para um determinado ano, foi utilizado o valor do pixel mais próximo nos anos anteriores. Se não houvesse valores preditos nos anos anteriores, foi utilizado o valor do pixel do ano seguinte mais próximo. Os pixels que permaneceram sem predições foram mapeados como -2.

3.5 Estatísticas pontuais e zonais

Foram calculadas estatísticas zonais dos estoques de COS no Brasil. As áreas de interesse (feições) foram biomas, unidades federativas, estados, municípios, unidades de conservação, áreas de uso e cobertura da terra (MapBiomas Coleção 7.1, nível 1). Em cada área de interesse, o estoque total e médio de COS foi computado. O código utilizado é de acesso aberto e pode ser consultado no material suplementar deste documento.

Considerando o suporte de predição pontual do mapeamento, que reporta um valor único para cada pixel do território (médio em t/ha de COS para determinado pixel) foi realizado o cômputo do estoque de COS em cada feição. A soma total dos estoques em cada feição foi obtida multiplicando o valor do pixel (estoque de COS) pela área do pixel (ee.Image.pixelArea()). Por exemplo, um dado pixel, de coordenada (*i, j, t),* tem o seu valor de estoque de COS dado em t/ha. Considerando que cada pixel representa a média do estoque de COS na área que lhes confere a sua resolução espacial, que é de aproximadamente 30 × 30 m (ou cerca de 900 m^2), a equação empregada para obter a média dos pixels seria:

$$\sum COS_{estoque}\ (ano) = \sum_{(i,j)=1}^{n} (COS_{estoque,\ n}) \bullet 900\ m^2$$

Para o cálculo médio, a soma dos estoques (estoque total da feição) foi dividida pela soma da área ocupada pela feição (área da feição).

4. Estratégias de avaliação

4.1 Dados de treinamento

O cálculo do estoque de COS estocada numa camada de solo requer dados de quatro propriedades do solo:

- concentração (g/kg) de COS na fração terra fina (diâmetro ≤ 2 mm) seca em estufa (105 °C),
- proporção (volume) da massa do solo ocupada por fragmentos minerais de diâmetro > 2 mm,
- densidade (g/cm) do solo inteiro e
- espessura (m) da camada de solo sendo considerada nos cálculos.

O uso de dados de estoques de COS para produzir uma série de mapas anuais requer que os mesmos sejam acompanhados de coordenadas do local (espaço) e momento (tempo) da coleta no campo. Em geral, quanto mais representativa da realidade modelada forem os dados pontuais de solo, maiores são as chances de se obter mapas mais realísticos. Assim, nesta Coleção Beta, a qualidade dos dados pontuais de solo foi avaliada utilizando gráficos de distribuição no tempo e espaço geográfico e de atributos do solo. A meta dessa avaliação gráfica foi verificar o quão bem os dados cobrem as dimensões modeladas do território e solo brasileiros.

Os dados de densidade do solo inteiro receberam atenção especial na avaliação de qualidade dos dados de treinamento. Isso porque essa propriedade do solo precisou ser estimada para um grande número de amostras de solo usando um modelo Random Forest. Neste caso, a avaliação consistiu do cômputo de um conjunto de estatísticas do erro de predição do modelo treinado. São elas:

ME $= \sum_{i=1}^{n} \frac{yi-xi}{n}$

MAE $= \frac{\sum_{i=1}^{n} |yi-xi|}{n}$

MSE $= \frac{(yi-xi)^2}{n}$

RMSE $= \sqrt{\frac{\sum_{i=1}^{n} (yi-xi)^2}{n}}$

NSE $= 1 - \frac{MSE}{\frac{(\overline{xi}-xi)^2}{n}}$

Em que:

- yi é o valor predito;
- xi é o valor observado;
- n é o número de observações;
- $\overline{xi}$ é a média dos valores observados;
- *ME* é calculado através da média dos erros;
- *MAE* é o erro calculado a partir da média dos erros absolutos;
- *MSE* é o erro calculado a partir da média dos erros elevado ao quadrado;
- *RMSE* é o erro calculado através da raiz quadrada do mse;
- *NSE* considera a eficiência do modelo, sendo uma medida de ajuste do modelo em relação a variabilidade dos dados observados. É calculado como 1 menos a razão entre o MSE e a média dos resíduos elevado ao quadrado;
- *slope* é o coeficiente de declividade do modelo de regressão entre os valores observados e os valores preditos. O resultado é o segundo coeficiente retornado pela função *lm*, que realiza a regressão linear.

Um valor negativo de ME indica subestimativas, e um valor positivo indica superestimativas. O MAE é uma métrica que avalia a precisão de um modelo de previsão, levando em consideração a diferença absoluta entre os valores observados e os valores preditos. Ele é expresso na mesma unidade de medida dos dados. Um MAE menor indica um melhor desempenho do modelo em termos de precisão (Hyndman and Koehler, 2006; Karunasingha, 2022).

O MSE e o RMSE são medidas que caracterizam o desempenho do modelo. O MSE é calculado em uma unidade de leitura que pode ser difícil de entender, mas a raiz quadrada no RMSE permite que ele seja expresso na mesma unidade de medida das observações analisadas, facilitando sua interpretação. Tanto o RMSE quanto o MAE penalizam erros muito grandes, mas tendem a se aproximar um do outro quando os erros são menores (Karunasingha, 2022).

NSE é uma estatística normalizada que determina a magnitude relativa da variância residual em comparação com a variação nos dados de medição. A escala varia de menos infinito a 1, onde 1 representa uma correspondência perfeita entre o modelo e os dados observados. Um valor próximo a 0 indica um desempenho pior do que a média dos dados observados (Lopes, 2020).

As estatísticas do erro de predição do modelo Random Forest foram calculadas usando as predições fora-do-saco (out-of-bag) e de uma validação cruzada com 10 partições construídas de maneira inteiramente aleatória. A comparação dos resultados obtidos usando os dois procedimentos permite verificar a sensibilidade do modelo Random Forest à redução de 10% no número de pontos de treinamento. Por fim, a plausibilidade pedológica do modelo Random Forest treinado foi avaliada ordenando as covariáveis em função de sua importância na redução da variância nas partições das árvores de regressão.

4.2 Validação cruzada do modelo preditivo

A validação cruzada comum de mapas digitais de solo pode gerar estimativas de incerteza (das predições) muito otimistas nos cenários em que as amostras de solo apresentam agrupamentos espaciais, comum em bases de dados legados. Nesse sentido, a validação cruzada espacial é uma alternativa para inferir sobre a magnitude de erro dos modelos nas regiões com amostras esparsas, ajudando a refletir sobre as extrapolações dos modelos no espaço.

Para validação do modelo foi utilizado os pacotes randomForest e Caret no software R, configurados com os mesmos hiperparâmetros utilizados para a modelagem em nuvem no Google Earth Engine, garantindo a consistência dos resultados. Na validação cruzada comum, do tipo k-Fold, as amostras foram divididas aleatoriamente em 10 subconjuntos e as amostras foram aleatoriamente sorteadas para uso no treinamento e validação do modelo. Na validação cruzada espacial, os dados de treinamento passaram primeiro por um processo de agrupamento espacial e temporal usando k-*Means* (30 grupos). Diferentes grupos de amostras foram aleatoriamente sorteados, também em 10 subconjuntos, para treinar ou validar o modelo. A Figura 8, ilustra o fluxograma das estratégias de validação cruzada e validação cruzada espacial adotadas neste estudo. Os dados e códigos utilizados são de acesso aberto e podem ser consultados no material suplementar deste documento.

Durante a validação, foram explorados diferentes números de clusters (10, 20 e 30) para considerar a influência das amostras agrupadas e das regiões sem agrupamento. Essa abordagem busca encontrar um equilíbrio entre a representação de padrões gerais e as particularidades locais. A variação no número de clusters permite obter previsões mais próximas aos dados de treinamento, considerando a heterogeneidade dos dados e a generalização do modelo.

Ao considerar um menor número de clusters, como 10, foi possível identificar padrões mais gerais nas previsões, abrangendo uma área maior. Por outro lado, ao aumentar o número de clusters para 20 ou 30, é possível capturar nuances mais específicas para diferentes regiões. Portanto, optou-se por utilizar 30 clusters como mais apropriado para capturar resultados. Uma observação importante é avaliar o desempenho do modelo nas regiões onde temos menor número de amostras. Para avaliar a qualidade do modelo, utilizou-se métricas de ajuste do modelo, tais como ME, MAE, MSE, RMSE e NSE. Estas métricas foram obtidas tanto em nível nacional quanto por bioma. Para o nível de bioma, apenas os dados de validação cruzada comum são apresentados.

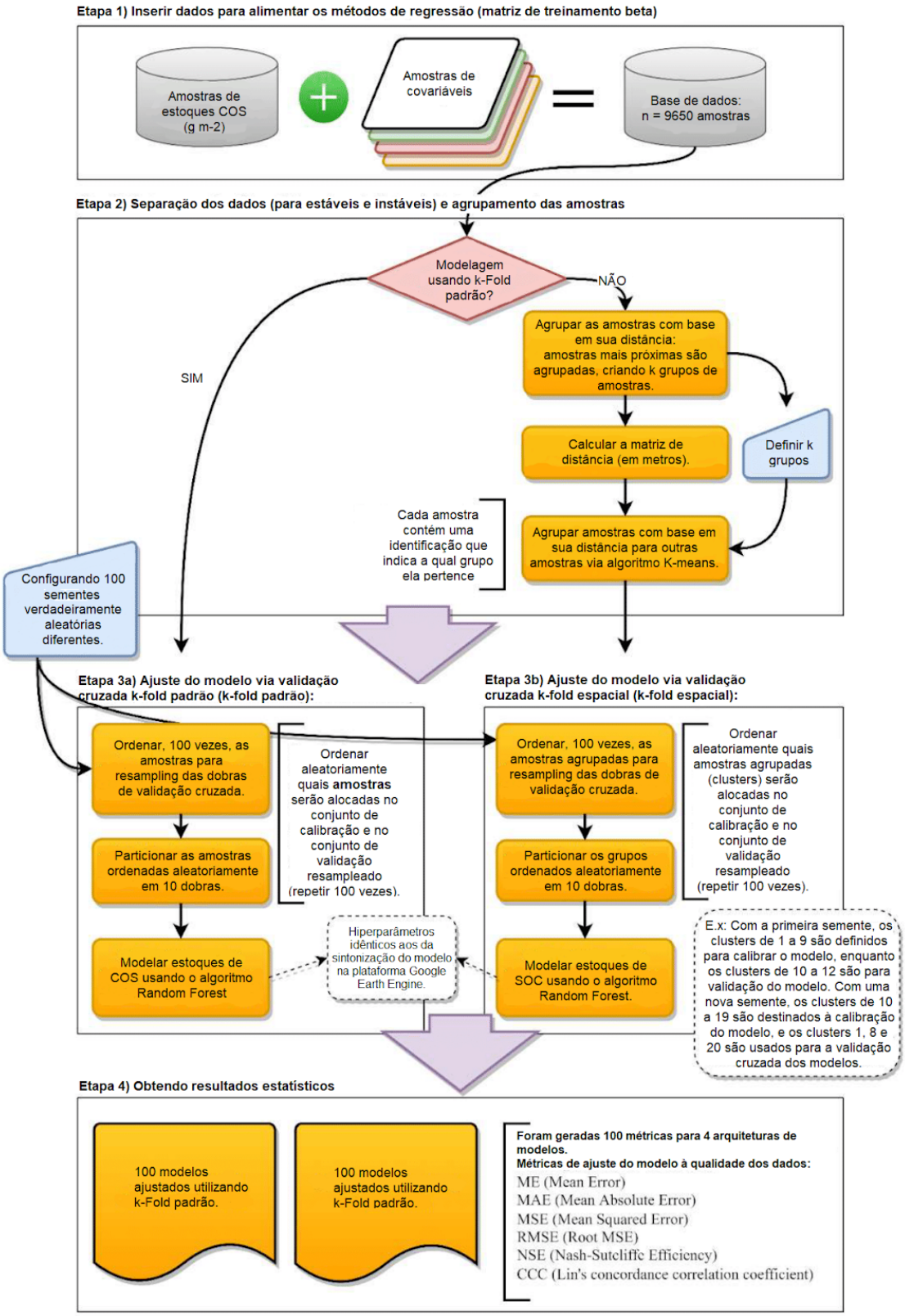

Figura 8. Fluxograma das estratégias de validação cruzada e validação cruzada espacial.

5. Resultados da coleção e sua análise

5.1 Disponibilidade de dados de treinamento

5.1.1 Densidade do solo

Os dados faltantes de densidade do solo nos pontos de amostragem foram estimados por meio de um modelo Random Forest. A Figura 9 demonstra a frequência dos dados disponíveis para o treinamento do modelo Random Forest pelo SoilData para predição da densidade do solo. Nota-se um volume maior de dados com densidade do solo acima de 1 g/cm^3 até 1,7 g/cm^3.

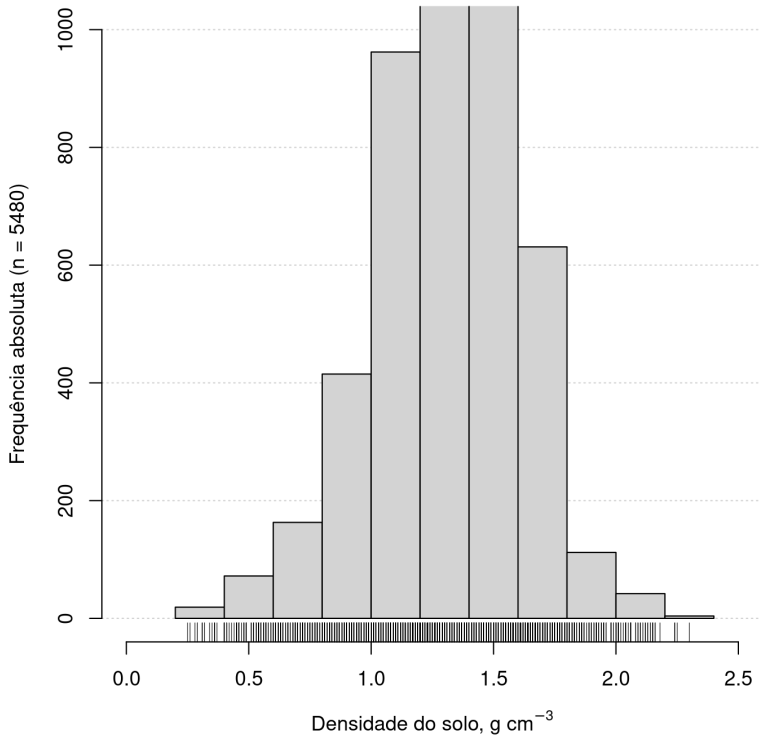

Figura 9. Dados de treinamento do *Random Forest* para densidade do solo (n = 5480) disponíveis no SoilData.

A Figura 10 demonstra a importância de cada variável utilizada no modelo de regressão Random Forest para predição da densidade do solo de 34.589 camadas com dados faltantes para essa propriedade do solo. As frações granulométricas e o teor de carbono orgânico são as propriedades de maior influência na predição da densidade do solo.

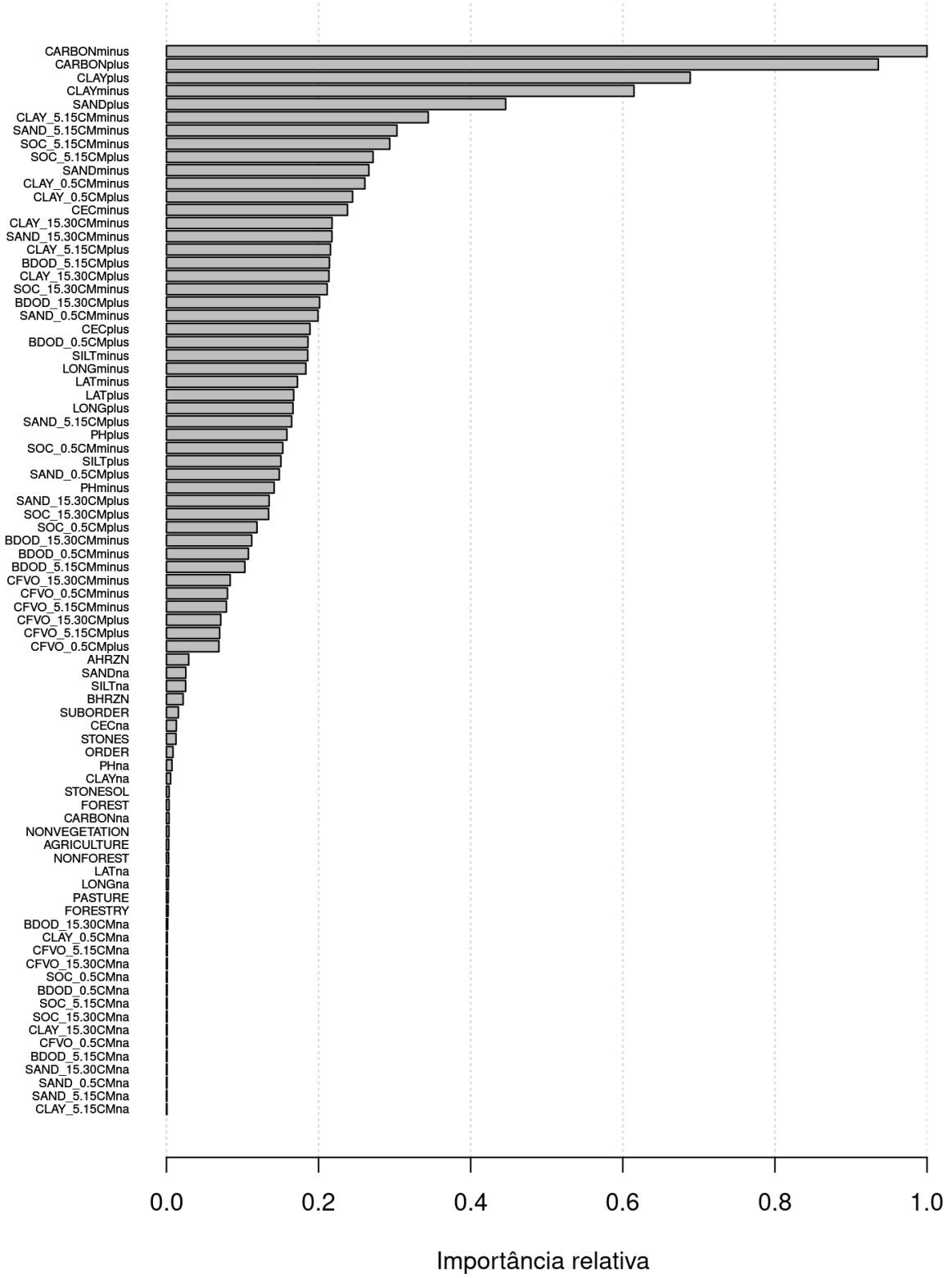

Figura 10. Importância relativa das variáveis preditoras na redução da variância nas partições das árvores de regressão do modelo Random Forest para densidade do solo.

A Tabela 9 apresenta os resultados estatísticos para o modelo de regressão Random Forest da estimativa da densidade do solo, utilizando-se o modelo OOB e CV para validação. Ambos os métodos demonstraram a efetividade do modelo com a mesma magnitude, podendo-se

confirmar com os valores de NSE de 0,6904 e 0,6808 para a validação com o OOB e CV, respectivamente.

Tabela 9. Estatísticas de desempenho do modelo de regressão *Random Forest* para a estimativa da densidade do solo.

Método	Estatísticas					
	ME (g/cm^3)	MAE (g/cm^3)	MSE (g/cm^6)	RMSE (g/cm^3)	NSE	slope
Out-of-bag (OOB)	0	0,1159	0,027	0,1642	0,6904	1,0641
k-fold = 10 (CV)	0	0,1191	0,0278	0,1667	0,6808	1,0888

A Figura 11 apresenta a dispersão dos dados observados de densidade do solo versus os dados preditos pelo modelo do Random Forest para ambos, validação por OOB e CV. Nota-se que a distribuição dos pontos de predição pelo modelo é semelhante entre os métodos de validação. O modelo apresenta uma melhor estimativa para valores de densidade entre 0,8 e 1,7 g/cm^3, uma vez que é a faixa onde se concentram o maior número de amostras de treinamento. Os esforços em reunir dados de solos que cobrem as diferentes regiões do país auxiliarão nos modelos de treinamento para estimativa da densidade do solo.

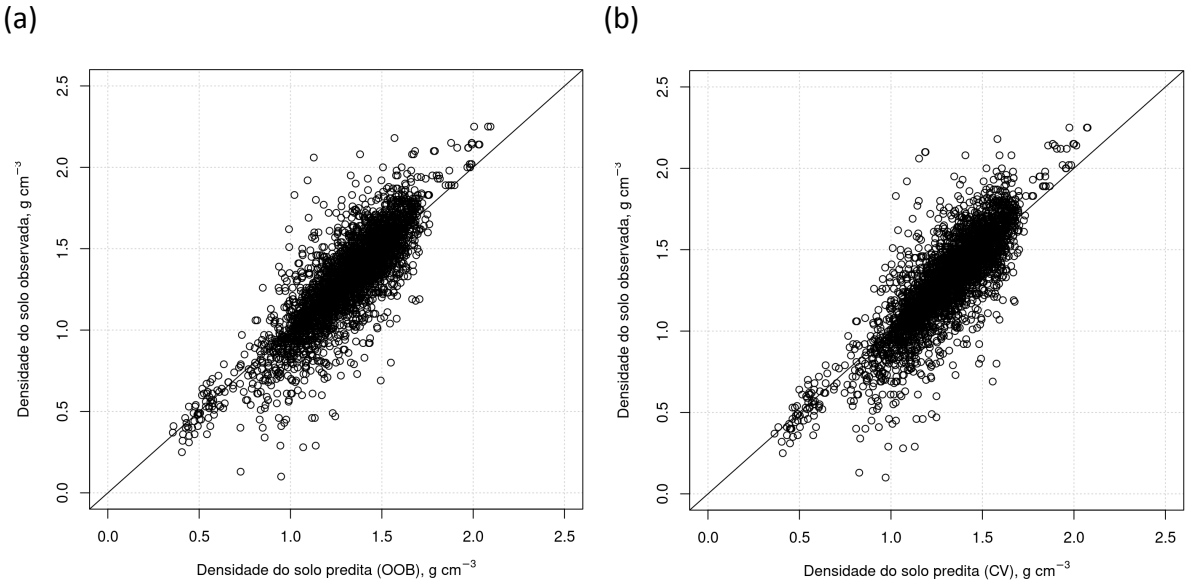

Figura 11. Dispersão dos valores de densidade do solo (g/cm^3) observados no campo e preditos nas amostras mantidas (a) fora da bolsa (*out-of-bag samples*) e preditos pela (b) validação cruzada (CV) durante o treinamento do modelo de regressão *Random Forest*.

5.1.2 Profundidade e espessura da camada

Dados de propriedades do solo estão disponíveis para camadas de diferentes espessuras.

Essa variação acontece porque a camada de interesse varia de caso para caso. Em alguns casos, como em áreas agrícolas, são amostradas camadas de espessura fixa predeterminada. Contudo a amostragem costuma ser limitada aos primeiros 10 ou 20 cm a partir da superfície do solo. Além disso, em parte considerável do território nacional, o solo possui menos de 30 cm de profundidade. Para estimar os estoques de carbono no solo, foram consideradas apenas amostras de solo de profundidade de até 30 cm (Figura 12). Porém, podemos estar subestimando os estoques caso haja solo com mais de 30 cm de profundidade.

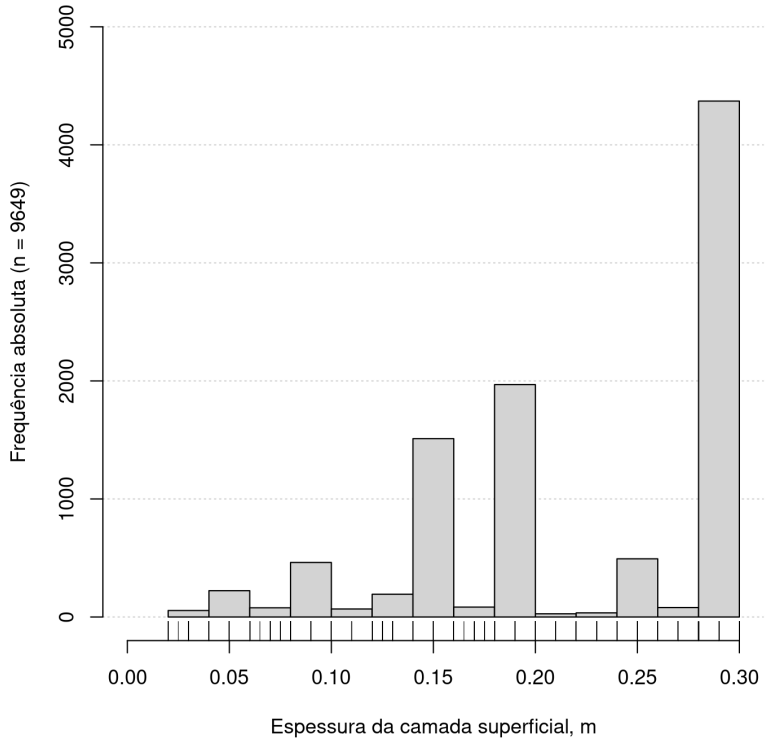

Figura 12. Distribuição de frequência empírica da profundidade/espessura da camada dos dados de solo (n = 9649) disponíveis no SoilData para modelagem espaço-temporal dos estoques de carbono orgânico no território brasileiro.

5.1.3 Fração grossa e raízes

Dados concretos sobre o volume ocupado pelas raízes ainda são escassos. Mas vale lembrar que as raízes finas constituem menos 1% da biomassa total das florestas, porém sua produção anual (de raízes finas) pode contribuir com mais de 50% na produção primária líquida total das florestas. Nas florestas, menos de 20% da biomassa total está abaixo do solo, embora mais de 50% do carbono absorvido anualmente pelas plantas pode estar alocado abaixo do solo. Cerca de 50 a 80% das raízes são encontradas nos primeiros 30 cm do solo. Portanto, é necessário reunir informações da densidade das raízes nos locais amostrados, e melhorar a precisão do estoque de COS em regiões de interesse.

5.1.4 Distribuição espacial e temporal

O volume de dados disponíveis no SoilData para mapeamento dos estoques de carbono no solo é considerável. Sua distribuição espaço-temporal, contudo, é heterogênea (Figura 13)**.** A maioria dos dados são provenientes de amostras de solo coletadas entre as décadas de 1970 e 1990 (Figura 14), sendo que 4336 pontos referem-se a coletas de amostras realizadas entre 1958 e 1984. Dados das décadas de 2000 e 2010 são escassos e concentrados em pequenas regiões onde trabalhos mais detalhados foram realizados por universidades e instituições de pesquisa estaduais. Dados produzidos por empresas privadas, e que estejam disponíveis para reuso, são praticamente inexistentes. Esforços são necessários para incentivar a abertura de dados das instituições públicas e privadas.

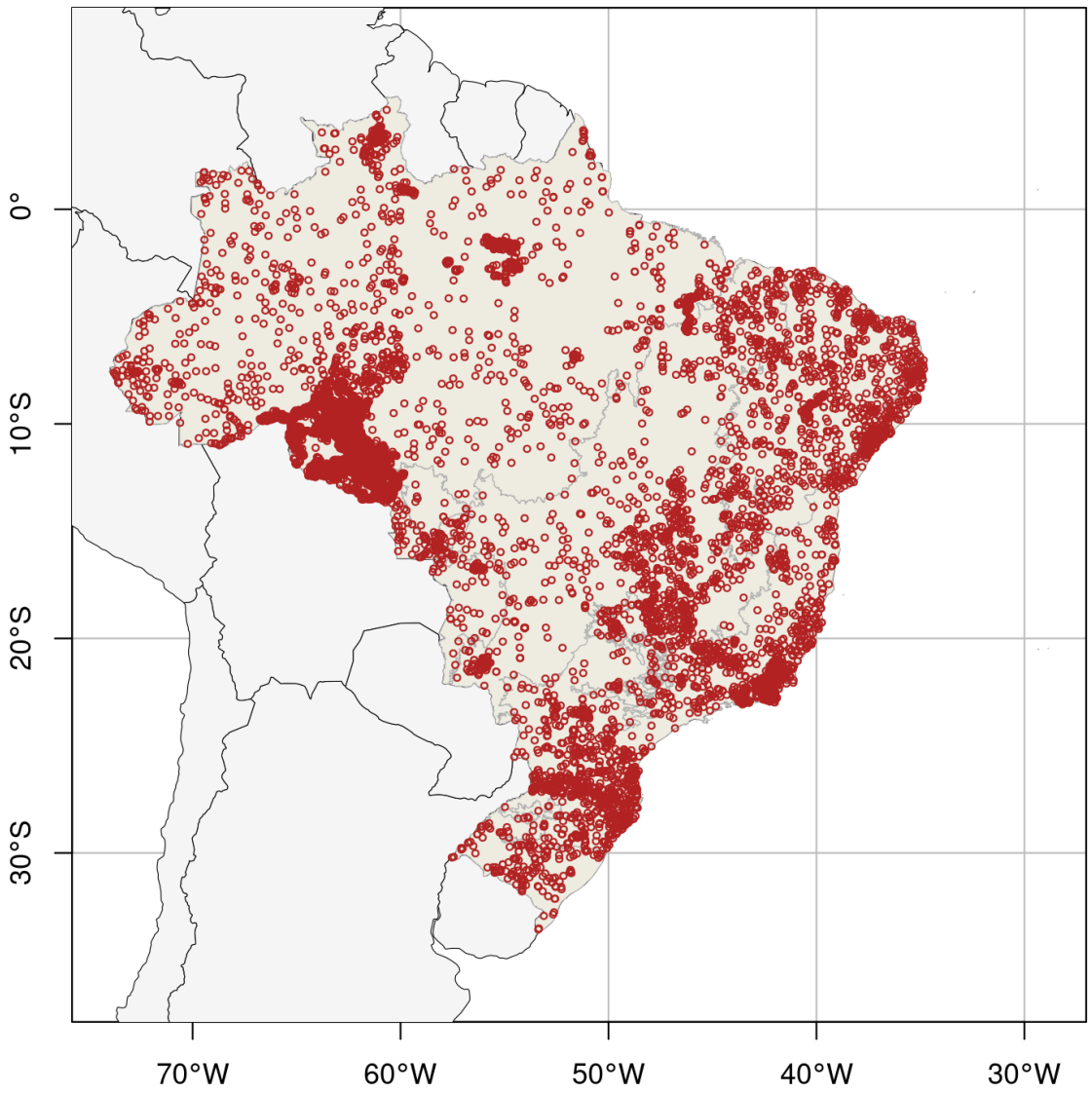

Figura 13. Distribuição de frequência empírica no espaço geográfico dos dados de solo (n = 9650) disponíveis no SoilData para modelagem espaço-temporal dos estoques de carbono orgânico no território brasileiro.

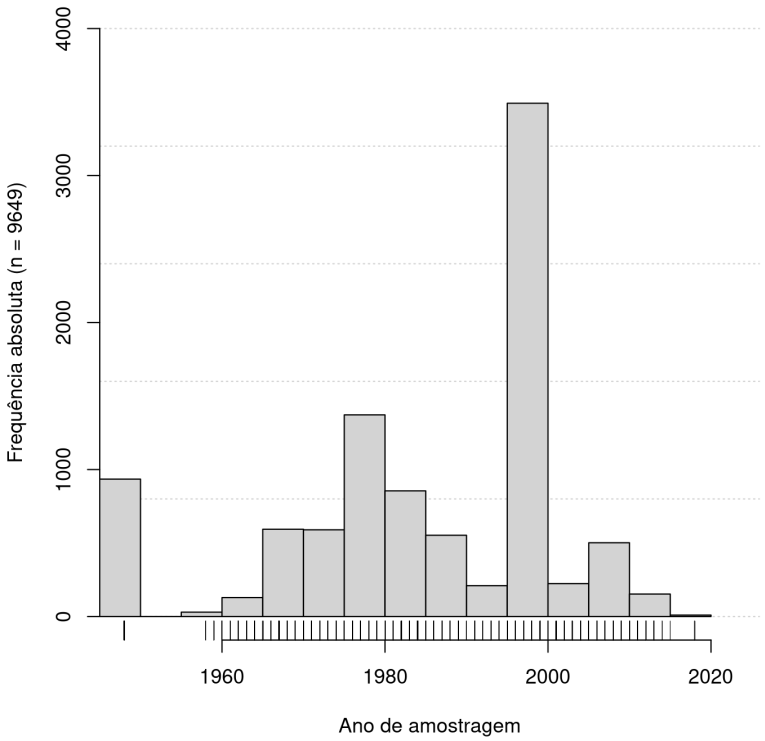

Figura 14. Distribuição de frequência empírica no tempo dos dados de solo (n = 9650) disponíveis no SoilData para modelagem espaço-temporal dos estoques de carbono orgânico no território brasileiro.

O número de amostras por bioma é bem variado, o que acaba desfavorecendo o modelo de treinamento de predição dos estoques de COS, subestimando ou superestimando valores (Tabela 10). O bioma Amazônia é o que apresenta o maior número de amostras (46,3%), seguido da Mata Atlântica (~30%). O Pantanal é o menor bioma brasileiro e também o que apresenta menor número de dados (~0,96%) pontuais dentre o conjunto de dados disponíveis no SoilData. Outro bioma com menor número de amostras é o Pampa, com representação de 1,75% dos dados. A discrepância do número de dados de treinamento entre os diferentes biomas pode prejudicar o modelo. As amostras dos biomas Pantanal e Pampa apresentam-se pouco distribuídos pelo seu território. Portanto, é de grande certeza a necessidade em reunir mais amostras de todas as regiões do país, a fim de melhorar o modelo de predição nas próximas versões.

Tabela 10. Distribuição dos dados de treinamento entre os biomas brasileiros.

Bioma	Número de pontos por ano (média)	Número de pontos a cada 1000 km^2	Número absoluto de pontos
Amazônia	121	1.1	4468
Pantanal	3	0.6	93
Caatinga	23	1.0	864
Cerrado	32	0.6	1116
Mata Atlântica	78	2.6	2888
Pampa	5	0.9	169

5.1.5 Estimativa pontual dos estoque de COS

Para o cálculo do estoque de COS o primeiro passo consistiu em transformar os dados de conteúdo de terra fina (g/kg) para volume (cm^3/cm^3). Isso foi feito assumindo que a média da densidade das frações fina e grossa é 2,65 g/cm^3. Abaixo, a Figura 15 apresenta a frequência absoluta dos dados de COS. A grande maioria dos dados referem-se a amostras com baixa concentração de carbono, mais precisamente abaixo de 20 t/ha.

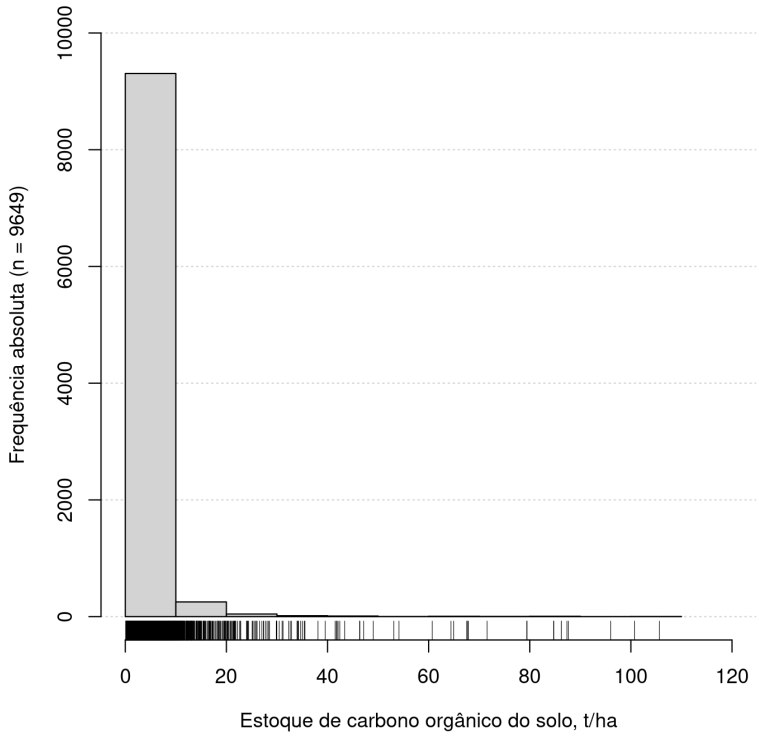

Figura 15. Distribuição de frequência empírica dos dados de estoque de carbono orgânico do solo (n = 9649) disponíveis no SoilData para modelagem espaço-temporal no território brasileiro.

5.2 Desempenho do modelo espaço-temporal

5.2.1 Desempenho geral do modelo

As estatísticas de validação cruzada são apresentadas por meio de diagramas de caixa (boxplot), que fornecem uma representação visual das métricas de desempenho dos modelos na validação cruzada comum e validação cruzada espacial, que permitem uma comparação entre diferentes métricas como o MSE, RMSE e NSE, para identificar padrões e tendências nas estimativas (Figura 16).

O desempenho do modelo espaço-temporal variou conforme a abordagem de validação cruzada utilizada. Na validação cruzada comum, em que as amostras são sorteadas individualmente para compor o conjunto de validação, o modelo apresentou um RMSE de 3,67 kg/m^2 (36,7 t/ha) e um NSE de 0,37. Esse tipo de validação é um bom indicador de qualidade das predições em locais em áreas onde há maior densidade de amostras, fornecendo portanto, estimativas mais precisas da qualidade dos mapas nessas áreas.

Por outro lado, na validação cruzada espacial o desempenho do modelo foi pior. Os valores de RMSE chegaram a 4,18 kg/m^2 (41,8 t/ha) com NSE de 0,08. Isso revela um desafio maior para o modelo em fazer estimativas precisas nas áreas com menor densidade amostral.

De modo geral, em lugares onde há vazios de dados, é provável que o desempenho do modelo seja mais próximo ao da validação cruzada espacial. Para obter uma estimativa global mais precisa do modelo espaço-temporal, a média aritmética das duas validações pode ser considerada, sendo RMSE = 3,9 kg/m^2 (39 t/ha) e um NSE = 0,22. Essa abordagem proporciona uma visão mais abrangente da qualidade do modelo, levando em conta tanto as áreas com densidade amostral significativa quanto às regiões com amostras mais escassas, contribuindo para uma análise mais precisa e abrangente dos resultados.

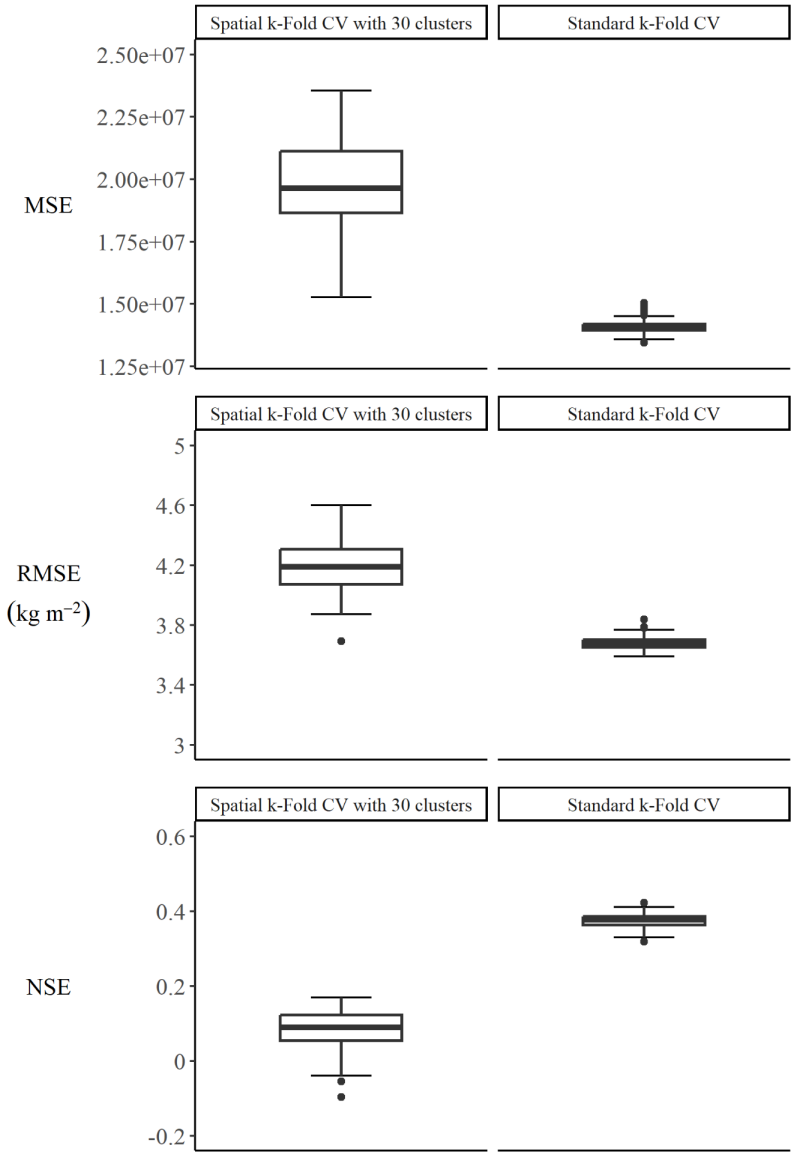

Figura 16. Diagrama de caixa das métricas de validação cruzada comum (Standard k-Fold CV) e validação cruzada espacial com 30 grupos (Spatial k-Fold CV with 30 clusters). Em que: MSE: erro médio absoluto, RMSE: erro quadrático médio, NSE: eficiência do modelo.

5.2.2 Desempenho do modelo por Bioma

Para um bom desempenho preditivo, o algoritmo precisa capturar as complexas relações entre solo-covariáveis nos dados de treinamento. O número de amostras necessárias para isso varia e elas devem estar alocadas de forma a contemplar os componentes da paisagem que afetam a dinâmica de cada bioma como o tipo de vegetação, o clima e a biodiversidade. De modo geral, quanto maior o número de amostras disponíveis para um bioma, maior as chances do modelo preditivo resultante ter um bom desempenho.

A análise das métricas de validação cruzada comum, agrupando os resultados por bioma permitem uma compreensão mais detalhada de como o modelo se comporta em diferentes

regiões do país. Os resultados permitem identificar diferenças expressivas, demonstrando a influência do número de amostras e do contexto de cada bioma na qualidade das predições (Figura 17)**.**

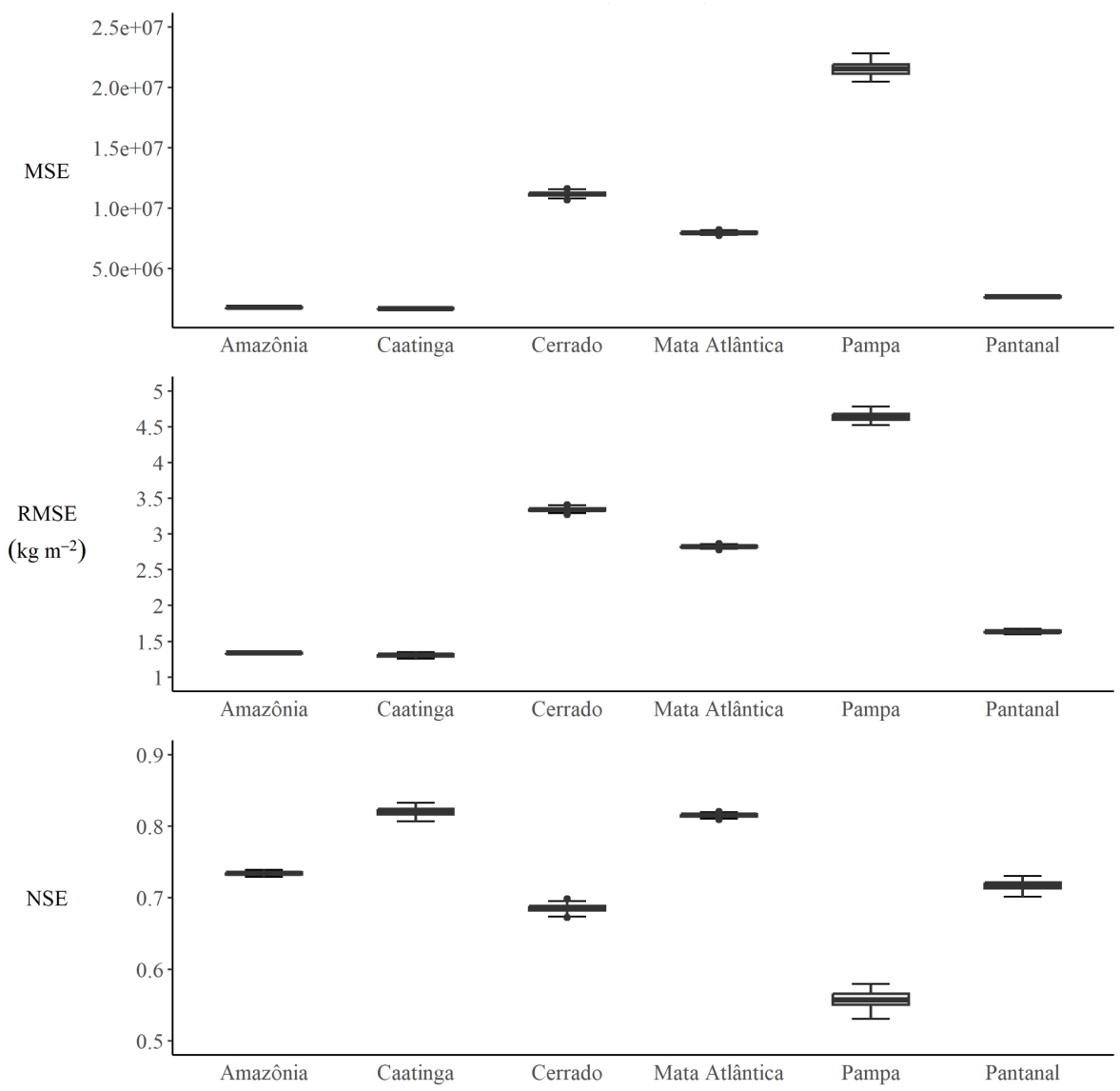

Figura 17. Diagrama de caixa das métricas de validação cruzada comum (Standard k-Fold CV). Em que: MSE: erro médio absoluto, RMSE: erro quadrático médio, NSE: eficiência do modelo.

O modelo apresentou melhor desempenho no bioma Caatinga e Mata Atlântica. Na Mata Atlântica, mesmo com 2888 amostras, registrou o segundo melhor desempenho, devido ao RMSE superior ao da Caatinga, com NSE de 0,49 e um RMSE de 4,67 kg/m^2 (46,7 t/ha). Enquanto a Caatinga teve 864 amostras, o modelo obteve o melhor resultado, com o NSE de 0,46 e menor RMSE de 2,26 kg m^{-2} (22,6 t/ha). É possível que a Caatinga, apesar de ter

menor número de amostras, possua amostras mais representativas em relação às características que o modelo está tentando estimar.

Ao comparar a Amazônia e o Pantanal, podemos observar diferenças significativas em relação à disponibilidade de amostras de treinamento e ao desempenho do modelo de previsão. A Amazônia, com 4468 amostras de treinamento, apresentou NSE de 0,24 e o RSME de 2,26 kg/m^2 (22,6 t/ha). Por outro lado, o Pantanal teve apenas 93 amostras de treinamento, mas o modelo ainda obteve resultados semelhantes, com NSE de 0,23 e um RMSE 2,69 kg/m^2 (26,9 t/ha). Essa comparação indica que, apesar da quantidade limitada de amostras no Pantanal, as similaridades entre os biomas permitiu métricas semelhantes.

No caso do Cerrado e do Pampa, observou-se que a representatividade das amostras é um fator crítico. O Cerrado, com 1166 amostras, apresentou baixa eficiência com NSE de 0,19 e alto RSME 5,33 kg/m^2 (53,3 t/ha), demonstrando baixa representatividade das amostras no bioma, resultando em maiores erros nas estimativas. De maneira similar, o Pampa, com apenas 169 amostras, registrou o pior desempenho, com NSE de -0,17, o que indica que o modelo não conseguiu explicar a variação dos dados melhor que sua média aritmética, resultando no maior RMSE de 7,54 kg/m^2 (75,4 t/ha) entre todos os biomas. Esses resultados destacam a possível baixa representatividade amostral possivelmente associada às mudanças na dinâmica de uso do solo.

6. Considerações práticas

6.1 Processamento e armazenamento de dados em nuvem

Para essa coleção foram armazenadas cerca de 1,5 terabytes de covariáveis estáticas e dinâmicas, incluindo ativos proprietários e disponíveis publicamente, como o SoilGrids, no workspace do MapBiomas na Google. O tempo de processamento de toda coleção foi de aproximadamente uma semana, a depender das flutuações do número de usuários e outros fatores inerentes à computação em nuvem.

A coleção beta completa de mapas anuais de carbono orgânico do solo do Brasil (1985-2021) tem aproximadamente 36 gigabytes, equivalente a uma média de 1 gigabyte por mapa/ano. Ela foi produzida utilizando aproximadamente 79.450 EECU/h Unidades de Computação do Earth Engine (EECU), que é uma medida de poder computacional da plataforma.

6.2 Nota de precaução

A coleção beta apresenta uma estimativa dos estoques globais e relativos de COS, bem como o padrão espacial de larga escala esperado desses estoques. A plausibilidade pedológica dos resultados, no espaço e no tempo, deverão ser avaliadas considerando as condições ambientais presentes e pretéritas de cada região do país. Além disso, é a partir desta coleção que a disponibilidade e qualidade dos dados de solo em cada região do país será avaliada. Isso irá direcionar esforços para o resgate e compartilhamento de dados, permitindo uma melhoria contínua na qualidade e precisão das estimativas.

O arcabouço teórico-conceitual utilizado para gerar as informações contidas nesses mapas é baseado em modelos matemáticos que integram informações sobre as características do solo, clima e vegetação. Os modelos levam em conta os principais processos que controlam o estoque de carbono no solo no espaço, como variáveis morfométricas de terreno, e no tempo, como mudanças no uso da terra. Entretanto, outros vetores importantes de mudança, como variáveis climáticas e regimes de manejo, não foram incluídos no modelo. Isso acontece devido a baixa disponibilidade de dados de campo e sua distribuição heterogênea no espaço e no tempo. Estes vetores serão considerados à medida que novos dados foram incluídos no SoilData e novas estratégias de modelagem forem exploradas.

Os impactos mais críticos do uso da terra no Brasil não são contemplados por essa série. Isso porque os maiores problemas de degradação do solo registrados no país, precedem 1985, período para o qual não temos informações satelitárias disponíveis. Nesta época, uma porção significativa da terra em uso antrópico agropecuário encontrava-se degradada ou em processo de degradação. Quando técnicas de manejo e conservação tornaram-se populares no país, a partir da década de 90, as mudanças no sistema de cultivo, como o cultivo mínimo e plantio direto, aumentaram os estoques de COS em algumas regiões do país, especialmente quando elas estavam abaixo do esperado pelas condições ambientais. Nestes

casos, o aumento dos estoques de COS se deve às condições iniciais do solo no início da série, o que não sugere que o uso seja eficiente em termos de sequestro de COS.

O modelo utilizado para fazer as predições é uma primeira aproximação do que esperamos que a distribuição dos estoques de COS nos primeiros 30 centímetros do solo brasileiro seja. Isso porque os modelos são representações simplificadas da realidade. Nele, processos complexos de adição, perda, transformação e translocação das propriedades do solo e, por consequência, dos estoques de COS, são representados de maneira simplista no espaço e no tempo. Ele contempla um determinado conjunto de dados de campo, associadas a um conjunto de covariáveis ambientais, cuja relação é computada por um modelo. As estimativas apresentadas são, portanto, a melhor aproximação dos valores verdadeiros, compostos por erros e incertezas inerentes ao processo de mapeamento do solo.

Os mapas de estoques de COS da coleção beta não apresentam os estimativas espacialmente explícitas de incerteza associadas ao valor predito, porém destacamos aqui as principais limitações dos dados utilizados e as potenciais fontes de incerteza do produto disponibilizado, dentre as incertezas associadas aos dados pontuais de solo, as covariáveis e ao modelo preditivo. Essas incertezas são combinadas (se somam, se anulam, se multiplicam…) e são transmitidas ao mapa final. Esse processo, chamado de propagação, causa uma distorção em relação à verdade de campo, ou do que esperamos que ela seja. Essa variação pode, por vezes, ter a mesma magnitude da própria variação natural dos estoques de COS na paisagem e no tempo.

6.2.1 Limitações inerentes aos dados pontuais de estoque de COS

As amostras disponíveis para coleção beta possuem distribuição espacial e temporal heterogênea ao longo da série e praticamente não possuem repetição no tempo.

Em cada dado pontual de solo, a quantificação das propriedades do solo contém erros, isso porque toda medição, por mais criteriosa que seja, é sempre uma aproximação do valor real. Na determinação do conteúdo de carbono de uma amostra de solo, os métodos analíticos e equipamentos têm precisão limitada distinta. Como os dados utilizados são oriundos de diferentes projetos, produzidos para diferentes propósitos, a combinação dos dados de COS obtidos por diferentes métodos analíticos exigiu uma etapa de harmonização para serem combinados.

Para parte das amostras de campo, a densidade do solo precisou ser estimada por um modelo de regressão, com base em outras propriedades do solo e covariáveis ambientais. Outro aspecto importante é que dados do volume do solo que é ocupado por raízes não é considerado no cálculo do estoque de COS. Apesar de serem muito importantes, especialmente em áreas de floresta, o volume normalmente é ignorado por falta de dados disponíveis.

Todas as amostras pontuais de solo estão associadas a um dado de posição no espaço geográfico, suas coordenadas geográficas. A qualidade destes dados posicionais depende principalmente da precisão dos equipamentos utilizados para obtê-las. Grosso modo, quanto mais antigo o dado de campo disponível no SoilData, pior a qualidade da estimativa posicional, chegando em casos extremos, na casa de quilômetros. Isso significa que por vezes, ainda que a estimativa do estoque de COS em um ponto esteja muito próxima do valor real esperado, esse dado pode ser erroneamente correlacionado com condições ambientais completamente distintas. Este risco é maior quanto maior a resolução das covariáveis.

6.2.2 Limitações inerente às covariáveis

As covariáveis são utilizadas em representações aproximadas da paisagem.

A predição temporal é particularmente dependente da qualidade dos dados de uso e cobertura da terra, ou seja, os erros de quantificação e alocação nos dados são propagados para os mapas de estoque de COS. Além disso, não temos conhecimento sobre a dinâmica do uso e cobertura da terra pré 1985.

Considerando que algumas áreas do Brasil estão permanentemente cobertas por nuvens (dados faltantes), especialmente para o início da série histórica na região norte do país, os dados faltantes foram preenchidos utilizando filtros, o que também podem afetar os resultados finais.

Finalmente, mapas de cobertura e uso da terra existentes bem como as imagens de satélite utilizadas para cálculo do NDVI, entretanto, não são suficientes para representar os tipos de manejo do solo nas áreas de produção agrícola, pecuária e florestal. O período de tempo coberto por esses mapas também é pequeno quando comparado ao período de tempo pelo qual os solos brasileiros são utilizados em atividades produtivas. A indisponibilidade de dados para as décadas anteriores aos anos 80 dificulta a modelagem do efeito de médio e longo prazo da dinâmica de cobertura e uso das terras sobre as propriedades do solo. Soluções serão desenvolvidas para melhorar a representação do manejo agrícola durante a série histórica.

6.2.3 Limitações inerentes ao modelo preditivo

Random Forest é um modelo baseado em dados, o que significa que seus resultados são diretamente influenciados pela qualidade e quantidade dos dados disponíveis.

Apesar da alta capacidade de computação do Google Earth Engine, fundamental para a aplicação de um modelo de predição em larga escala como este, foram identificadas limitações na implementação do algoritmo Random Forest na plataforma. A principal limitação é que o algoritmo implementado (através da função ee.Classifier.smileRandomForest) não permite maior controle do treinamento do modelo

além da configuração dos parâmetros mtry, ntree, minLeafPopulation (tamanho mínimo dos nós terminais) e bagFraction (tamanho(s) da amostra a ser retirada). Soluções estão sendo buscadas para melhorar a flexibilidade dos parâmetros do algoritmo e, especialmente para o cômputo das estatísticas globais e locais de incerteza do modelo.

6.3 Isenção de responsabilidade

As informações apresentadas na coleção beta são baseadas nos melhores, e às vezes únicos, dados de solo disponíveis, informações sobre o meio ambiente e técnicas de mapeamento digital do solo. Apesar de este produto ser criado com o máximo cuidado, o(s) autor(es) e/ou editor(es) e/ou MapBiomas não podem ser responsabilizados por quaisquer danos causados pelo uso destes dados ou de qualquer conteúdo nele sob qualquer forma, sejam ou não causados por possíveis erros ou falhas, nem por quaisquer consequências dos mesmos.

As designações empregadas e a apresentação do material neste produto informativo não implicam a expressão de qualquer opinião por parte da MapBiomas sobre a situação legal dos estoques de carbono do solo de qualquer território, cidade ou área ou de suas autoridades.

7. Considerações Finais e Perspectivas

O feedback dos usuários é importante para o aprimoramento contínuo das futuras versões do mapeamento, com melhorias na precisão e qualidade dos dados para atender às necessidades e expectativas dos usuários. Os desenvolvedores do MapBiomas Solo estão comprometidos em corrigir artefatos e inconsistências relatados pelos usuários da coleção beta, a fim de gerar uma nova coleção de mapas para disponibilização ao público. Essa colaboração entre usuários e desenvolvedores é fundamental para ampliar a utilidade da coleção de mapas do MapBiomas Solo para pesquisadores, cientistas e tomadores de decisão em todo o país.

Atualizações estão previstas para as próximas coleções. A primeira delas, é a estimativa espacialmente explícita da incerteza do modelo ao fazer predições, que serão calculadas ao nível do pixel. Junto com as avaliações globais de qualidade, essas estimativas de incerteza permitirão que a equipe tome decisões informadas sobre onde concentrar esforços para melhorar a série de mapas. Já os usuários de dados, uma vez informados e capacitados, utilizarão essas estatísticas para decidir como melhor empregar os dados em suas aplicações.

A segunda é o mapeamento dos estoques de COS na camada subsuperficial do solo, de 30 a 100 cm de profundidade. Nessa camada são esperados estoques de COS significativos, especialmente em biomas como Pantanal e Cerrado. A terceira é a expansão do universo de variáveis dinâmicas explicativas para contemplar outros vetores de mudança nos estoques de COS. Assim, a dinâmica temporal dos estoques de COS passará a contemplar, não só a mudança na cobertura e uso da terra, mas também as mudanças no clima, nos regimes de uso e manejo do solo, nos regimes de inundação e frequência, intensidade e cicatriz das queimadas do país.

Finalmente, se espera que a iniciativa fomente o desenvolvimento de uma ciência solo mais colaborativa e aberta, com compartilhamento de ideias, dados e algoritmos. A colaboração é essencial para produzir uma série temporal que represente mais fielmente a realidade do solo. Os dados coletados no campo por especialistas que conhecem as relações entre o solo e a paisagem dos biomas brasileiros, sejam eles resgatados de iniciativas pretéritas ou obtidas no campo por iniciativas contemporâneas, estão no cerne da evolução dos produtos. Iniciativas como o Programa Nacional de Levantamento e Interpretação de Solos do Brasil (PronaSolos) e o serão importantes parceiros nessa empreitada.

8. Referências

Alvares, C.A., Stape, J.L., Sentelhas, P.C., de Moraes Gon, J.L., Sparovek, G., 2013. Köppen's climate classification map for Brazil. Meteorol. Z. 22, 711–728. https://doi.org/10.1127/0941-2948/2013/0507

Amatulli, G., McInerney, D., Sethi, T., Strobl, P., Domisch, S., 2020. Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. Sci. Data 7, 162. https://doi.org/10.1038/s41597-020-0479-6

Arrouays, D., Leenaars, J.G.B., de-Forges, A.C.R., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., Heuvelink, G., Batjes, N., Carvalho, E., Hartemink, A., Hewitt, A., Hong, S.-Y., Krasilnikov, P., Lagacherie, P., Lelyk, G., Libohova, Z., Lilly, A., McBratney, A., McKenzie, N., Vasquez, G.M., Mulder, V.L., Minasny, B., Montanarella, L., Odeh, I., Padarian, J., Poggio, L., Roudier, P., Saby, N., Savin, I., Searle, R., Solbovoy, V., Thompson, J., Smith, S., Sulaeman, Y., Vintila, R., Rossel, R.V., Wilson, P., Zhang, G.-L., Swerts, M., Oorts, K., Karklins, A., Feng, L., Navarro, A.R.I., Levin, A., Laktionova, T., Dell, M., Suvannang, N., Ruam, W., Prasad, J., Patil, N., Husnjak, S., Pásztor, L., Okx, J., Hallett, S., Keay, C., Farewell, T., Lilja, H., Juilleret, J., Marx, S., Takata, Y., Kazuyuki, Y., Mansuy, N., Panagos, P., Liedekerke, M.V., Skalsky, R., Sobocka, J., Kobza, J., Eftekhari, K., Alavipanah, S.K., Moussadek, R., Badraoui, M., Silva, M.D., Paterson, G., da Conceição Gonçalves, M., Theocharopoulos, S., Yemefack, M., Tedou, S., Vrscaj, B., Grob, U., Kozák, J., Boruvka, L., Dobos, E., Taboada, M., Moretti, L., Rodriguez, D., 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ 14, 1–19. https://doi.org/10.1016/j.grj.2017.06.001

Batjes, N.H., Ribeiro, E., van Oostrum, A., 2020. Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). Earth Syst. Sci. Data 12, 299–320. https://doi.org/10.5194/essd-12-299-2020

Batjes, N.H., Ribeiro, E., van Oostrum, A., Leenaars, J., Hengl, T., Jesus, J.M. de, 2017. WoSIS: providing standardised soil profile data for the world. Earth Syst. Sci. Data 9, 1–14. https://doi.org/10.5194/essd-9-1-2017

Bernoux, M., da Conceição Santana Carvalho, M., Volkoff, B., Cerri, C.C., 2002. Brazil's Soil Carbon Stocks. Soil Sci. Soc. Am. J. 66, 888–896. https://doi.org/10.2136/sssaj2002.8880

Botelho, R.G.M., Brilha, J., 2022. Principles for Developing a National Soil Heritage Inventory. Geoheritage 14, 7. https://doi.org/10.1007/s12371-021-00643-y

Brasil, 2021. Quarta Comunicação Nacional do Brasil à Convenção Quadro das Nações Unidas sobre Mudança do Clima. Ministério da Ciência, Tecnologia e Inovações, Brasília, Distrito Federal, Brasil.

Brasil, 2015. Intended Nationally Determined Contribution Towards Achieving the Objective of the United Nations Framework Convention on Climate Change. República Federativa do Brasil, Brasília, Distrito Federal, Brasil.

Camargo, F.A.O., Alvarez, V.H., Baveye, P.C., 2010. Brazilian soil science: from its inception to the future, and beyond. Rev. Bras. Ciênc. Solo 34, 589–599. https://doi.org/10.1590/S0100-06832010000300001

Chagas, C.S., Junior, W.C., Bhering, S.B., Tanaka, A.K., Baca, J.F.M., 2004. Organization and structure of the Brazilian soil information system (SigSolos – version 1.0). Rev. Bras. Ciênc. Solo 28, 865–876. https://doi.org/10.1590/S0100-06832004000500009

Cooper, M., Mendes, L.M.S., Silva, W.L.C., Sparovek, G., 2005. A national soil profile database for Brazil available to international scientists. Soil Sci. Soc. Am. J. 69, 649–652. https://doi.org/10.2136/sssaj2004.0140

Costa, E.M., Samuel-Rosa, A., Anjos, L.H.C. dos, 2018. Digital elevation model quality on digital soil mapping prediction accuracy. Ciênc. E Agrotecnologia 42, 608–622. https://doi.org/10.1590/1413-70542018426027418

FAO, 2022. Global status of black soils. FAO, Rome, Italy. https://doi.org/10.4060/cc3124en

FAO, 2020. Technical specifications and country guidelines for Global Soil Organic Carbon Sequestration Potential Map (GSOCseq), Global Soil Partnership. Food and Agriculture Organization of the United Nations, Rome.

FAO, 2014. World reference base for soil resources 2014: international soil classification system for naming soils and creating legends for soil maps. FAO, Rome.

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

Hanson, B., Sugden, A., Alberts, B., 2011. Making data maximally available. Science 331, 649–649. https://doi.org/10.1126/science.1203354

Hengl, T., Jesus, J.M. de, MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km–global soil information based on automated mapping. PLoS ONE 9, e105992. https://doi.org/10.1371/journal.pone.0105992

Hengl, T., Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLOS One 12, e0169748. https://doi.org/10.1371/journal.pone.0169748

Heuvelink, G.B.M., Angelini, M.E., Poggio, L., Bai, Z., Batjes, N.H., van den Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G.F., Sanderman, J., 2021. Machine learning in space and time for modelling soil organic carbon change. Eur. J. Soil Sci. 72, 1607–1623. https://doi.org/10.1111/ejss.12998

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22, 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

IPCC, C.C., 2014. Mitigation of climate change. Contrib. Work. Group III Fifth Assess. Rep. Intergov. Panel Clim. Change.

Kämpf, N., 1971. Mineralogia e gênese de alguns solos da região nordeste do planalto riograndense. Universidade Federal do Rio Grande do Sul, Faculdade de Agronomia, Porto Alegre.

Karunasingha, D.S.K., 2022. Root mean square error or mean absolute error? Use their ratio as well. Inf. Sci. 585, 609–629. https://doi.org/10.1016/j.ins.2021.11.036

Lal, R., 2013. Soil carbon management and climate change. Carbon Manag. 4, 439–462. https://doi.org/10.4155/cmt.13.31

Lopes, T.R., 2020. Modelagem hidrológica e estudos de pagamentos por serviços ambientais para bacia hidrográfica do Rio Piracicaba (text). Universidade de São Paulo. https://doi.org/10.11606/T.11.2020.tde-08012021-111754

Ottoni, M.V., Filho, T.B.O., Schaap, M.G., Lopes-Assad, M.L.R.C., Filho, O.C.R., 2018. Hydrophysical Database for Brazilian Soils (HYBRAS) and pedotransfer functions for

water retention. Vadose Zone J. 17, 1–17. https://doi.org/10.2136/vzj2017.05.0095
Ottoni, M.V., Lopes-Assad, M.L.R.C., Pachepsky, Y., Filho, O.C.R., 2014. A hydrophysical database to develop pedotransfer functions for Brazilian soils: challenges and perspectives, in: Teixeira, W.G., Ceddia, M.B., Ottoni, M.V., Donnagema, G.K. (Eds.), Application of Soil Physics in Environmental Analyses. Springer International Publishing, Basel, pp. 467–494. https://doi.org/10.1007/978-3-319-06013-2_20
Palladino, M., Romano, N., Pasolli, E., Nasta, P., 2022. Developing pedotransfer functions for predicting soil bulk density in Campania. Geoderma 412, 115726. https://doi.org/10.1016/j.geoderma.2022.115726
Peralta, G., Di Paolo, L., Luotto, I., Omuto, C., Mainka, M., Viatkin, K., Yigini, Y., 2022. Global soil organic carbon sequestration potential map (GSOCseq v1.1) - Technical manual. FAO, Rome. https://doi.org/10.4060/cb2642en
Poggio, L., de Sousa, L.M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. SOIL 7, 217–240. https://doi.org/10.5194/soil-7-217-2021
Rahmati, M., Weihermüller, L., Vanderborght, J., Pachepsky, Y.A., Mao, L., Sadeghi, S.H., Moosavi, N., Kheirfam, H., Montzka, C., Looy, K.V., Toth, B., Hazbavi, Z., Yamani, W.A., Albalasmeh, A.A., Alghzawi, M.Z., Angulo-Jaramillo, R., Antonino, A.C.D., Arampatzis, G., Armindo, R.A., Asadi, H., Bamutaze, Y., Batlle-Aguilar, J., Béchet, B., Becker, F., Blöschl, G., Bohne, K., Braud, I., Castellano, C., Cerdà, A., Chalhoub, M., Cichota, R., Císlerová, M., Clothier, B., Coquet, Y., Cornelis, W., Corradini, C., Coutinho, A.P., de Oliveira, M.B., de Macedo, J.R., Durães, M.F., Emami, H., Eskandari, I., Farajnia, A., Flammini, A., Fodor, N., Gharaibeh, M., Ghavimipanah, M.H., Ghezzehei, T.A., Giertz, S., Hatzigiannakis, E.G., Horn, R., Jiménez, J.J., Jacques, D., Keesstra, S.D., Kelishadi, H., Kiani-Harchegani, M., Kouselou, M., Jha, M.K., Lassabatere, L., Li, X., Liebig, M.A., Lichner, L., López, M.V., Machiwal, D., Mallants, D., Mallmann, M.S., de Oliveira Marques, J.D., Marshall, M.R., Mertens, J., Meunier, F., Mohammadi, M.H., Mohanty, B.P., Pulido-Moncada, M., Montenegro, S., Morbidelli, R., Moret-Fernández, D., Moosavi, A.A., Mosaddeghi, M.R., Mousavi, S.B., Mozaffari, H., Nabiollahi, K., Neyshabouri, M.R., Ottoni, M.V., Filho, T.B.O., Pahlavan-Rad, M.R., Panagopoulos, A., Peth, S., Peyneau, P.-E., Picciafuoco, T., Poesen, J., Pulido, M., Reinert, D.J., Reinsch, S., Rezaei, M., Roberts, F.P., Robinson, D., Rodrigo-Comino, J., Filho, O.C.R., Saito, T., Suganuma, H., Saltalippi, C., Sándor, R., Schütt, B., Seeger, M., Sepehrnia, N., Moghaddam, E.S., Shukla, M., Shutaro, S., Sorando, R., Stanley, A.A., Strauss, P., Su, Z., Taghizadeh-Mehrjardi, R., Taguas, E., Teixeira, W.G., Vaezi, A.R., Vafakhah, M., Vogel, T., Vogeler, I., Votrubova, J., Werner, S., Winarski, T., Yilmaz, D., Young, M.H., Zacharias, S., Zeng, Y., Zhao, Y., Zhao, H., Vereecken, H., 2018. Development and analysis of the soil water infiltration global database. Earth Syst. Sci. Data 10, 1237–1263. https://doi.org/10.5194/essd-10-1237-2018
Samuel-Rosa, A., Dalmolin, R.S.D., Miguel, P., Zalamena, J., Dick, D.P., 2013. The effect of intrinsic soil properties on soil quality assessments. Rev. Bras. Ciênc. Solo 37, 1236–1244. https://doi.org/10.1590/S0100-06832013000500013
Samuel-Rosa, A., Dalmolin, R.S.D., Moura-Bueno, J.M., Teixeira, W.G., Alba, J.M.F., 2020. Open legacy soil survey data in Brazil: geospatial data quality and how to improve it. Sci. Agric. 77. https://doi.org/10.1590/1678-992x-2017-0430
Samuel-Rosa, A., Heuvelink, G.B.M., Vasques, G.M., Anjos, L.H.C., 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243–244,

214–227. https://doi.org/10.1016/j.geoderma.2014.12.017
Samuel-Rosa, A., Vasques, G.M., 2017. Dados para aplicações pedométricas em larga escala no Brasil. Bol. Inf. SBCS 43, 22–24.
Sato, M.V., 2015. Primeira aproximação da biblioteca espectral de solos do Brasil: caracterização de espectros de solos e quantificação de atributos (Dissertação de Mestrado). Universidade de São Paulo, Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, SP. https://doi.org/10.11606/d.11.2015.tde-15102015-152045
SEEG, 2022. Sistema de Estimativas de Emissões e Remoções de Gases de Efeito Estufa: Setor Mudança de Uso da Terra e Florestas, 4th ed, Nota Metodológica. Instituto de Pesquisa Ambiental da Amazônia, Belém, Pará, Brasil.
Twala, B.E.T.H., Jones, M.C., Hand, D.J., 2008. Good methods for coping with missing data in decision trees. Pattern Recognit. Lett. 29, 950–956. https://doi.org/10.1016/j.patrec.2008.01.010
Vasques, G. de M., Dart, R. de O., Baca, J.F.M., Ceddia, M.B., Mendonça-Santos, M. de L., 2017. Mapa de estoque de carbono orgânico do solo (COS) a 0-30 cm do Brasil.

Material Suplementar

Dados de treinamento

O conjunto final de dados de treinamento e o código R para o seu processamento estão disponíveis em https://doi.org/10.58053/MapBiomas/4FJWZC**.**

Covariáveis

O código de processamento dos dados de propriedades dos solos está disponível em: https://code.earthengine.google.com/d1434ce0161e91a0ce9c7291efd329fa
O conjunto final de dados de propriedades dos solos processado está disponível no workspace do MapBiomas no Google Earth Engine em:
https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/SOLOS/ISRIC_SOILGRIDS_30M-v2_gapfill

O código de processamento dos dados de probabilidade de classes de solos está disponível em: https://code.earthengine.google.com/3bfe9d169290bc05240b59284f5a76e8
O conjunto final de dados de probabilidade de solos processado está disponível no workspace do MapBiomas no Google Earth Engine em:
https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/SOLOS/WRB_ALL-SOILS_SOILGRIDS_30M-v4

O código de processamento dos dados geomorfométricos está disponível em:
https://code.earthengine.google.com/fca5297060be8f223d9df73f84f3e3cd
O conjunto final de dados geomorfométricos processados está disponível no workspace do MapBiomas no Google Earth Engine em:
https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/SOLOS/OT_GEOMORPHOMETRY_30m-v1

O código de processamento dos dados de classificação climática está disponível em:
https://code.earthengine.google.com/1dc4fc46136cd63e695978778349190f
O conjunto final de dados processados está disponível no workspace do MapBiomas no Google Earth Engine em
https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/SOLOS/IPEF_KOPPEN_30M_DUMMY

O código de processamento dos dados *dummy* de fitofisionomias está disponível em:
https://code.earthengine.google.com/ec8744622a491f840f5c2434e720e35f
O conjunto final de dados *dummy* de fitofisionomias processados está disponível no workspace do MapBiomas no Google Earth Engine em:
https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/SOLOS/IBGE_FITOFISIONOMIAS_30M_DUMMY

O código de processamento dos dados *dummy* dos biomas está disponível em:
https://code.earthengine.google.com/294d9265cc4db274d36d656d83cef4f3
O conjunto final de dados *dummy* dos biomas processados está disponível no workspace do MapBiomas no Google Earth Engine em:
https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/SOLOS/IBGE_BIOMAS_30M_DUMMY

O código de processamento dos dados de persistência das classes de uso e cobertura está disponível em:
https://code.earthengine.google.com/3a0fa46d0b61d571a5a43dd16f190095
O conjunto final de dados processados está disponível no workspace do MapBiomas no Google Earth Engine:
https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/SOLOS/AGE-LULC-STARTING-AT-20-v5-col7

Os código de processamento dos mosaicos anuais de NDVI, EVI e SAVI, respectivamente:
https://code.earthengine.google.com/db81a57f99eb7d1ae4b5093828dbc666?
https://code.earthengine.google.com/818e5bb2b27a7bceb2d6877937d2fbc1?https://code.earthengine.google.com/d416534ed0a5637889d4b714729e0233?
Os mosaicos anuais dos índices NDVI, EVI e SAVI estão disponíveis no workspace do MapBiomas no Google Earth Engine em:
https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/SOLOS/LANDSAT_NDVI_BY_BYTE
https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/SOLOS/LANDSAT_EVI_BY_BYTE
https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/SOLOS/LANDSAT_SAVI_BY_BYTE

Modelagem espaço-temporal

O código de treinamento dos pontos está disponível neste:

https://code.earthengine.google.com/e56a92137acffe4a5eac4695f4bcd014

O conjunto de treinamento, treinamento com informações pontuais sobre carbono e covariáveis preditoras:

https://code.earthengine.google.com/?asset=projects/mapbiomas-solos-workspace/assets/points/carbonStocks-beta

O código de predição do carbono orgânico do solo está disponível em:

https://code.earthengine.google.com/4f6976ae4866b2fc155a17ee803f1eeb

O mapa de toneladas por hectare de COS entre 0 e 30 cm, antes da aplicação do filtro temporal está disponível em:

https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/SOLOS/PRODUTOS_BETA/soil_organic_carbon-0_30_cm_t_ha-beta_2_0

O código de aplicação do filtro temporal e mascara de água está disponível a seguir:

https://code.earthengine.google.com/52d4c9283a936d2af3e99c31a0f5ea46?noload=1

Os mapas finais de carbono orgânico do solo de 0 a 30 cm em t/ha e uma versão de valores arredondados em kg/m² e estão disponíveis em:

https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/SOLOS/PRODUTOS_BETA/soil_organic_carbon-0_30_cm_t_ha-beta_2_1

https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/SOLOS/PRODUTOS_BETA/soil_organic_carbon-0_30_cm_kg_m2-beta_2_1

O mapa utilizado na visualização da plataforma e do MapBiomas em kg/m² foi multiplicado por 100 e somado aos mapas de uso e cobertura da coleção 7.1, desta forma os valores de dezena e unidade são referentes a classe de uso e cobertura e os valores de milhar e centena são referentes ao quantidade de kg/m² de carbono orgânico no solo:

https://code.earthengine.google.com/?asset=projects/mapbiomas-workspace/public/collection7_1/mapbiomas_soil_collection1_carbon_coverage_v1

Código de acesso e visualização dos mapas finais de COS:

https://code.earthengine.google.com/cecfc60d40a4e6a8082d878e7cb1a3f2

Validação cruzada

O código de processamento da validação cruzada e seus resultados estão disponíveis em: https://github.com/eupassarinho/soc_stocks-random_forest-prediction

Toolkit

Toolkit MapBiomas de download dos mapas em TIFF e estatísticas da coleção beta:

https://code.earthengine.google.com/?scriptPath=users%2Fmapbiomas%2Fuser-toolkit%3Amapbiomas-user-toolkit-soil.js

Tutorial de uso dos toolkits MapBiomas:

https://github.com/mapbiomas-brazil/user-toolkit