CAPÍTULO 8

MAPEAMENTO DIGITAL DE SOLOS DO MARANHÃO: FUNDAMENTOS, BOAS PRÁTICAS E EXEMPLOS DE MAPEAMENTO DE CLASSES E ATRIBUTOS

Taciara Zborowski Horst
Professora - Universidade Tecnológica Federal do Paraná

Maria de Lourdes Mendonça-Santos
Pesquisadora – EMBRAPA SOLOS

Jean Michel Moura-Bueno
Professor - Universidade de Cruz Alta

Alessandro Samuel-Rosa
Professor - Universidade Tecnológica Federal do Paraná

Ana Caroline Pretto
Graduanda - Universidade Tecnológica Federal do Paraná

1. INTRODUÇÃO

O mapeamento de solos é uma ferramenta essencial para compreender a distribuição e as propriedades dos solos em uma região. Tradicionalmente, esse processo dependia da experiência do pedólogo, que utiliza observações de campo e a relação qualitativa solo-paisagem para delinear unidades de solo. Com o avanço das tecnologias, especialmente o sensoriamento remoto, os sistemas de informação geográfica (SIG) e as técnicas de modelagem matemática e estatística, o mapeamento digital de solos (MDS) emergiu como um método capaz de integrar grandes volumes de dados ambientais e pedológicos, gerando informações sobre os recursos do solo de forma mais precisa e em escalas antes inimagináveis. Essas inovações possibilitaram a criação de mapas, essenciais para o planejamento sustentável do uso da terra, a agricultura de precisão e a conservação ambiental. O trabalho de descrição dos perfis em campo e coleta das amostras para análises laboratoriais, continua sendo fundamental para treinar os modelos e representar o ambiente mapeado.

Neste capítulo, serão apresentadas as bases do MDS, seus fundamentos e sobre como os avanços tecnológicos e metodológicos transformaram a maneira como o solo é mapeado e como são explicitadas de forma quantitativa, as relações solo-paisagem. Será ainda apresentado e discutido o modelo SCORPAN (McBratney, Mendonça-Santos e Minasny, 2003) e sobre como ele pode ser utilizado para gerar informações de solo, tanto em termos de classes como de propriedades ou atributos dos solos, oferecendo uma visão abrangente de suas aplicações práticas. Os exemplos apresentados serão específicos para o Maranhão, uma região rica em diversidade ambiental e pedológica e de grande relevância para o estudo dos solos no Brasil. Serão explicitados os métodos utilizados na aplicação do modelo SCORPAN, desde o processamento dos dados iniciais até a modelagem e geração de mapas digitais. Além disso, serão demonstrados os produtos que podem ser gerados a partir desse método, como de propriedades específicas do solo e análises de distribuição espacial dos solos na paisagem.

Ao final deste capítulo, espera-se que os leitores tenham uma compreensão clara de como o MDS com base no modelo SCORPAN pode ser aplicado para transformar dados pontuais de solo em informações espacialmente explícitas de classes e atributos/propriedades dos solos para diversas finalidades, desde a agricultura até a gestão ambiental.

2. MAPEAMENTO DE SOLOS

O mapeamento de solos tem suas raízes na relação entre os solos e a paisagem, conceito que emergiu com os trabalhos de Dokuchaev, considerado o pai da Pedologia, em sua obra Russian Chernozem (1883). Essa abordagem foi expandida por outros cientistas do mundo ociedental, por Hans Jenny, que sistematizou os fatores de formação do solo em sua obra clássica Factors of Soil Formation: A System of Quantitative Pedology (1941). A partir de observações de campo nos Estados Unidos, Jenny propôs o modelo ClORPT, um acrônimo para Climate (Cl), Organisms (O), Relief (R), Parent material (P), Time (T), representado pela equação:

$$Solo = f(Cl, O, R, P, T, \dots)$$

Esse modelo estabelece que a formação e distribuição dos solos são condicionadas por fatores ambientais que interagem entre si ao longo do tempo. Jenny deixou a equação em aberto (reticências) para permitir a inclusão de novos fatores, reconhecendo a complexidade e variabilidade dos ambientes pedológicos.

A abordagem tradicional de mapeamento de solos baseia-se nesse modelo teórico e na experiência prática do pedólogo. O profissional analisa propriedades morfológicas do solo em campo – como cor, textura, estrutura e profundidade – e

as relaciona com elementos da paisagem, como relevo, vegetação e uso da terra. A partir dessas correlações, são delimitadas unidades de solo em mapas, mesmo em áreas não diretamente amostradas. Por exemplo, em regiões declivosas, a observação recorrente de solos rasos - Neossolos Litólicos com menos de 50 cm de profundidade – leva o pedólogo a associá-los a áreas com elevada declividade. Ao verificar essa relação em múltiplos pontos, o pedólogo projeta sua ocorrência para áreas similares usando dados topográficos, como os Modelos Digitais de Elevação (MDE), e delineia mapas de distribuição da classe com base nessa inferência. Esse método é fortemente dependente da capacidade do pedólogo em interpretar a paisagem e integrar conhecimento empírico aos fatores de formação. O processo decisório, desde a escolha dos pontos de amostragem até o delineamento de unidades de solo, é guiado por um julgamento técnico sem um modelo formal explicitamente definido. Por isso, o mapeamento tradicional é considerado uma prática interpretativa.

O avanço das tecnologias, como o sensoriamento remoto, os SIG e a crescente disponibilidade de dados ambientais em alta resolução, trouxe uma nova perspectiva ao estudo do solo. Essas ferramentas possibilitam representar os fatores de formação de forma contínua e quantitativa, ampliando a capacidade de análise e inferência espacial.

Foi nesse contexto que surgiu o MDS, consolidado pelo modelo SCORPAN, proposto por McBratney, Mendonça-Santos e Minasny (2003). Esse modelo expande o CLORPT ao incorporar a posição espacial (n) e o próprio solo (s) como variáveis preditoras, formando o acrônimo SCORPAN: solo - Soil (s), clima - Climate (c), organismos - Organisms (o), relevo - Relief (r), material parental - Parent material (p), idade - Age (a), e a posição espacial - Spatial position (n). Assim, o solo passa a ser descrito como função desses fatores e suas proxies digitais:

$$Solo = f(S, C, O, R, P, A, N)$$

Com isso, é possível modelar a distribuição de classes ou atributos do solo com base em algoritmos matemáticos e estatísticos, integrando grandes volumes de dados ambientais. O modelo SCORPAN tornou-se a base conceitual do MDS, permitindo a geração de mapas com maior precisão, escalabilidade e reprodutibilidade (Figura 1).

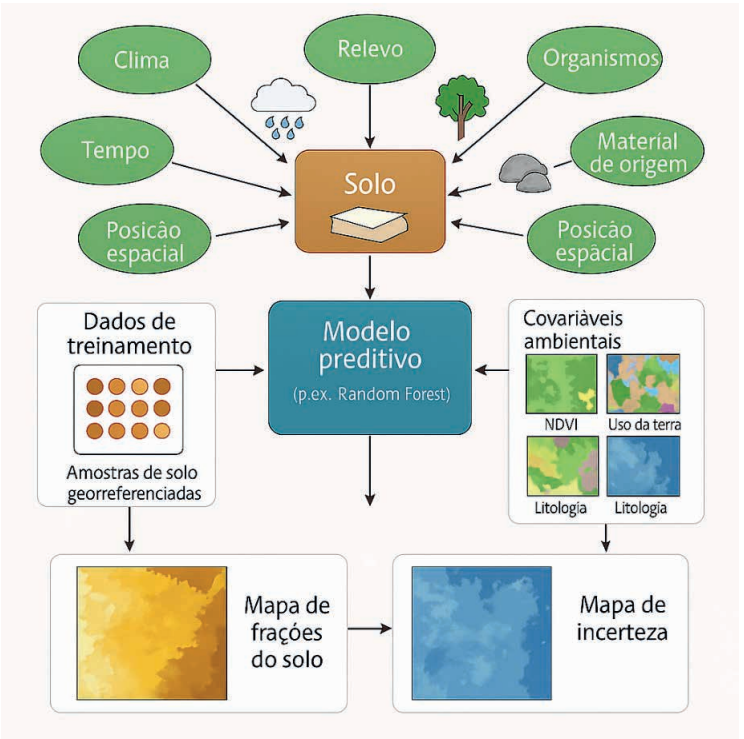

Figura 1. Fluxograma da modelagem preditiva do MDS.

2.1. Fundamentos e Bases Conceituais do Mapeamento Digital de Solos

O MDS foi definido por Lagacherie et al. (2006) como "a criação e inserção de dados em sistemas de informação espacial de solos por meio de modelos numéricos, para inferir as variações espaciais e temporais de classes de solos e de suas propriedades, a partir de observações, conhecimento pedológico e variáveis ambientais relacionadas". Essa abordagem produz mapas digitais contínuos de classes ou atributos do solo (como o teor de argila ou classe taxonômica, por exemplo), acompanhados de suas respectivas incertezas. Esses produtos são trabalhados em ambientes digitais, como SIG, permitindo análises integradas de dados em diferentes escalas e formatos. Assim, o MDS tornou-se uma ferramenta estratégica para o planejamento territorial, a agricultura de precisão, o monitoramento ambiental e a formulação de políticas públicas.

O raciocínio do MDS mantém o fundamento da relação solo-paisagem, formalizada inicialmente por Jenny (1941), com base nos fatores de formação descritos por Dokuchaev em 1883. No entanto, o avanço das tecnologias possibilitou

a evolução dessa formulação quantitativa para um modelo que integra a localização espacial através das coordenadas geográficas e permite a inclusão de outras variáveis ambientais que atuam na formação dos solos, inclusive, outras informações sobre os solos (McBratney, Mendonça-Santos e Minasny, 2003). No modelo SCORPAN, cada fator é representado por um conjunto de covariáveis ambientais, que podem ser contínuas (como elevação ou índice de vegetação por diferença normalizada - NDVI) ou categóricas (como classes de uso da terra ou litologia). Essas covariáveis são derivadas de diferentes fontes de dados, como MDE, imagens de satélite, mapas geológicos, dados climáticos interpolados ou informações de uso e cobertura do solo.

O termo "covariável" é utilizado porque essas variáveis explicativas covariam com a variável de interesse (alvo), ou seja, variam espacialmente de forma relacionada às mudanças observadas nas propriedades ou classes de solo. Por exemplo, a declividade (derivada de um MDE) pode aumentar à medida que a profundidade do solo diminui em uma determinada paisagem, evidenciando uma relação funcional entre relevo e desenvolvimento do solo. Essas covariáveis representam proxies quantificáveis dos fatores de formação do solo e permitem explicitamente modelar como o ambiente muda e influencia a pedogênese. Ao integrar essas informações em algoritmos de modelagem preditiva, o MDS permite a predição de atributos ou classes de solo em locais não amostrados, com base na variação espacial dessas covariáveis ambientais e nos dados de solos usados para treinar os modelos.

Em contraste com o mapeamento tradicional, que depende do julgamento do pedólogo, o MDS utiliza modelos matemáticos e estatísticos para combinar observações pontuais com essas covariáveis ambientais e gerar mapas digitais. Entre os algoritmos amplamente utilizados destacam-se técnicas de regressão, como a regressão múltipla, a regressão logística e outras e métodos de aprendizado de máquina, como o Random Forest, especialmente eficaz em contextos com múltiplos preditores e relações não lineares.

A transição conceitual e metodológica entre o mapeamento tradicional e o digital pode ser ilustrada pelo mesmo exemplo dos Neossolos Litólicos, apresentado no item anterior. No modelo tradicional, a associação entre solos rasos e áreas de elevada declividade é construída a partir da experiência de campo do pedólogo, que observa padrões recorrentes e projeta essa relação sobre áreas adjacentes (Figura 2A). No MDS, essa inferência é formalizada por meio de modelos numéricos que permitem quantificar e validar essas relações (Figura 2B). Por exemplo, em uma análise exploratória realizada em área declivosa com presença de Neossolos Litólicos, observou-se uma correlação negativa significativa ($r = -0,74$) entre a profundidade do solo e a declividade do terreno (Horst, 2017). Isso indica que, à medida que a declividade aumenta, a profundidade dos perfis tende a diminuir — uma evidência numérica da relação pedológica observada qualitativamente no campo.

A partir dessa relação, algoritmos de regressão ou modelos de aprendizado de máquina podem ser treinados para prever a probabilidade de ocorrência de solos rasos com base em dados contínuos de declividade. O resultado é um mapa digital que representa, em escala contínua, a distribuição potencial de Neossolos Litólicos em toda a paisagem — mesmo em áreas não amostradas.

Além disso, ao incluir outras covariáveis relevantes, como material de origem, índices de vegetação ou uso da terra, o modelo ganha robustez e capacidade de captar padrões complexos de formação do solo. Isso ilustra o principal diferencial do MDS: a transformação de relações qualitativas em funções matemáticas espacializadas, aplicáveis em grandes áreas e com controle explícito sobre a incerteza.

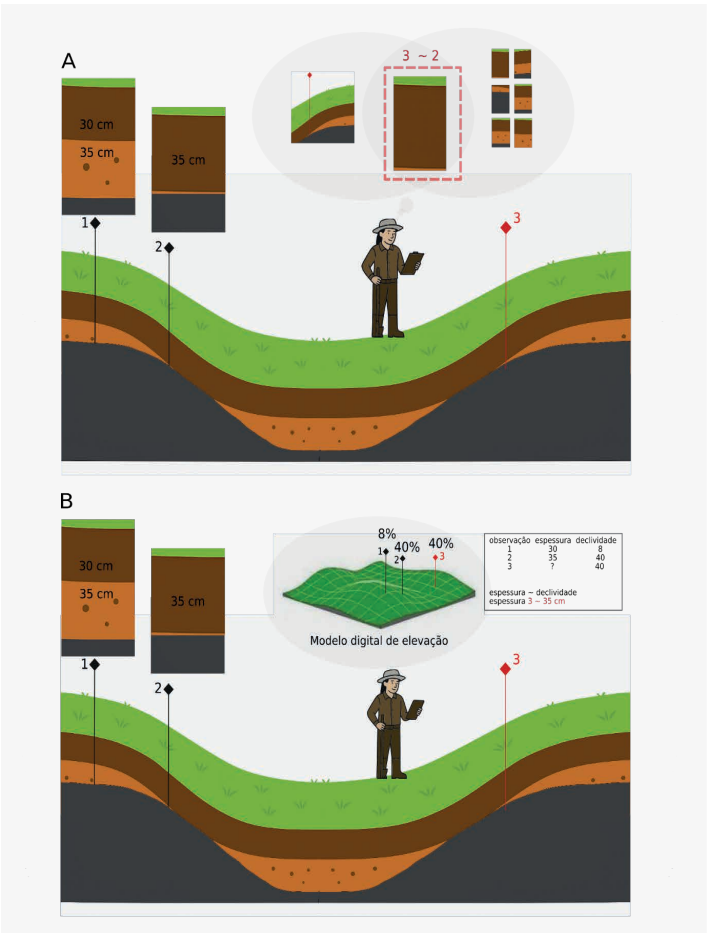

Figura 2. Comparação entre o raciocínio inferencial (qualitativo) no mapeamento tradicional (A), baseado na interpretação visual e extrapolação entre observações, e no mapeamento digital de solos (B), baseado na modelagem de relações quantitativas entre atributos do solo e variáveis ambientais como declividade.

Os primórdios do MDS, com base no modelo SCORPAN, foram estabelecidos pela publicação de McBratney, Mendonça-Santos e Minasny (2003), a qual descreveu as bases para o MDS no mundo. O primeiro Workshop global em MDS foi realizado em Montpellier, na França em 2004, resultando no livro "Digital Soil Mapping - An Introductory Perspective", publicado em 2007 (Lagacherie, McBratney e Voltz, 2007) e trazendo toda a base de trabalhos em MDS no mundo, incluindo um capítulo sobre a situação do MDS no Brasil "The state of the art of Brazilian Soil Mapping and prospects for Digital Soil Mapping" (Mendonça-Santos e Santos, 2007).

No Brasil, os primeiros trabalhos aplicados em MDS datam do início dos anos 2000, com destaque para as publicações do "2nd Global Workshop Digital Soil Mapping", realizado no Rio de Janeiro em 2006, culminando com a publicação dos trabalhos no livro intitulado "Digital Soil Mapping with Limited Data" (Hartemink, McBratney e Mendonca-Santos, 2008). Destaque para o trabalho de mapeamento de classes de solos do Rio de Janeiro, por Mendonça-Santos et al., (2006), no Rio de Janeiro. Nesse mesmo ano, Giasson et al. (2006), usaram o MDS para mapear os solos do Rio Grande do Sul, e depois muitos outros trabalhos seguindo estes protocolos foram realizados podendo ser citados (Ten Caten, 2012; Moura-Bueno et al., 2019; Samuel-Rosa et al., 2019).

Para atributos, como o estoque de carbono orgânico, destaca-se o trabalho de Mendonça-Santos et al. (2007), seguido por outras iniciativas nacionais com destaque para o mapa de carbono do solo até 2 m de profundidade, elaborado por Vasques et al. (2017) para compor o Mapa Global de Carbono Orgânico do Solo da FAO. No ano 2000, com o advento do MDS, foi criada a Rede Brasileira de Mapeamento Digital de Solos, por meio de um projeto financiado pelo CNPq, que permitiu os encontros, discussões e avanços em MDS e a cooperação entre membros da rede, bem como a consolidação no tema MDS no Brasil.

Cancian et al. (2018), ao conduzirem uma análise bibliométrica da produção científica em MDS entre 1996 e 2017, constataram o crescimento expressivo das publicações e o destaque crescente da pesquisa brasileira no cenário internacional, identificando aproximadamente 200 pesquisadores atuando com MDS no Brasil. Apesar desses avanços, uma revisão sistemática de 334 artigos de MDS conduzida por Coelho et al. (2021) revelou que, até a publicação de Mendonça-Santos et al. (2020), não havia trabalhos registrados com aplicação de MDS no estado do Maranhão, reforçando a relevância de novas iniciativas nesta região.

Atualmente, o Brasil avança com o Programa Nacional de Solos (PronaSolos), que visa mapear todo o território nacional em escalas de 1:25.000 a 1:100.000, que se tornou uma política pública coordenada pelo MAPA (Ministério da Agricultura e Pecuária) (https://www.gov.br/agricultura/pt-br/assuntos/sustentabilidade/

pronasolos), cuja plataforma com mapas e oinformacoes de solos do Brasil pode ser acessada em : https://www.embrapa.br/pronasolos e https://pronasolos.sgb.gov.br/. Nesse contexto, o MDS desponta como ferramenta indispensável para otimizar recursos, direcionar o planejamento amostral e ampliar a cobertura espacial dos levantamentos. O recente lançamento do Repositório de Dados de Solo (SoilData) (Samuel-Rosa et al., 2018) e estudos como os de Stumpf et al. (2016) e Hendriks et al. (2019) reforçam o papel estratégico do uso de dados legados no suporte à modelagem digital.

Nesse contexto, destaca-se a iniciativa MapBiomas Solo, que representa um avanço metodológico e institucional na consolidação do MDS em larga escala no Brasil. A partir da integração entre dados legados harmonizados — como aqueles disponíveis no SoilData —, imagens de sensoriamento remoto e algoritmos de aprendizado de máquina, o projeto viabiliza a geração de séries temporais de mapas de atributos do solo, com periodicidade anual e cobertura nacional (MAPBIOMAS, 2023). Essa abordagem espaço-temporal permite não apenas a análise estática das propriedades do solo, mas também o monitoramento de sua dinâmica ao longo do tempo, contribuindo para diagnósticos ambientais, gestão territorial e formulação de políticas públicas.

2.2. Componentes Operacionais do Mapeamento Digital de Solos

A aplicação do MDS envolve a integração entre dados observados em campo e técnicas de modelagem espacial. Operacionalmente, o MDS depende de três componentes fundamentais:

- Dados de treinamento: conjunto de amostras georreferenciadas com a variável-alvo (por exemplo, frações granulométricas, pH, classe de solo…), obtidas por amostragem de campo e análises laboratoriais;
- Modelo preditivo: algoritmo matemático ou estatístico que aprende os padrões espaciais da variável-alvo com base na correlação entre os dados amostrados ou covariáveis ambientais (fatores do modelo SCORPAN);
- Avaliação e validação: etapa que permite quantificar o desempenho do modelo, com métricas apropriadas conforme o tipo de variável (contínua ou categórica), além de gerar estimativas da incerteza associada à predição.

Essa estrutura torna o modelo explícito, reprodutível e passível de validação numérica, permitindo não apenas a geração de mapas, mas também a quantificação da confiança nas estimativas (mapa de erros ou incertezas do modelo). A natureza da variável-alvo — se contínua ou categórica —, define os métodos utilizados, tanto para a modelagem quanto para a avaliação e expressão da incerteza.

- Dados de treinamento

Os dados de treinamento correspondem ao conjunto de amostras de solo georreferenciadas que contêm a variável-alvo a ser modelada. Essa variável pode representar qualquer propriedade do solo (como teor de areia, silte, argila, carbono orgânico, pH, profundidade…) ou mesmo uma função do solo (como capacidade de retenção de água ou aptidão agrícola), desde que esteja relacionada aos fatores da paisagem — conforme o princípio da relação solo-paisagem que fundamenta o modelo SCORPAN.

As amostras podem ser coletadas especificamente para a finalidade de modelagem (atividade-fim), ou podem ser provenientes de reuso de dados legados, ou seja, coletados originalmente para outras finalidades. Ambas as abordagens são válidas, desde que os dados sejam confiáveis, bem documentados e compatíveis com os objetivos da modelagem.

A qualidade, variabilidade e distribuição espacial dos dados de solo têm influência direta sobre a robustez do modelo. Não há um número fixo ou mínimo universal de amostras: o mais importante é que o conjunto amostral represente adequadamente a complexidade ambiental da área de estudo. Em geral, quanto maior o número de amostras bem distribuídas na paisagem, maior é a chance de capturar os principais gradientes ambientais, o que favorece a generalização e reduz a incerteza das predições. Em ambientes altamente heterogêneos, essa representatividade torna-se ainda mais crítica para evitar extrapolações indevidas e interpretações incorretas.

Por fim, é fundamental compreender a natureza da variável-alvo — se contínua, categórica, derivada ou composta —, assim como o tipo e a quantidade de dados disponíveis. Essas informações guiam a escolha do algoritmo de modelagem mais adequado, influenciando desde a estrutura dos dados de entrada até as estratégias de validação e expressão da incerteza.

- Modelo preditivo

O modelo preditivo é o mecanismo matemático/estatístico ou computacional que estabelece a relação entre os dados de solo (variável-alvo) e os fatores ambientais representados por covariáveis. Ele é a peça central do MDS, pois transforma observações pontuais em inferências espaciais sobre toda a área de interesse. Existem dois grandes grupos de abordagens: modelos geoestatísticos e modelos baseados em covariaveis ambientais.

Modelos baseados em geoestatística exploram a autocorrelação espacial entre pontos amostrados, assumindo que locais próximos tendem a apresentar características similares. Técnicas como krigagem ordinária, cokrigagem ou krigagem com regressão externa são aplicadas principalmente para variáveis contínuas,

oferecendo não apenas a predição, mas também a variância da estimativa como medida de incerteza. Essas abordagens são mais indicadas quando a densidade amostral é razoável e a estrutura espacial pode ser modelada com semivariogramas confiáveis.

Modelos baseados em covariáveis ambientais: utilizam variáveis derivadas dos fatores do modelo SCORPAN (solo, clima, organismos, relevo, material parental, tempo e posição espacial) como proxies digitais da paisagem. Essas covariáveis podem ser contínuas, como declividade, NDVI, altitude, precipitação média, entre outras; ou categóricas, como uso da terra, tipo de vegetação, classes geológicas ou litológicas.

É essencial que as covariáveis selecionadas possuam relação causa-efeito plausível com a variável-alvo, ou seja, que façam sentido à luz dos processos de formação e distribuição dos solos. Essa relação pode ser linear ou não linear, o que influencia diretamente a escolha do algoritmo mais apropriado.

Diversos algoritmos de modelagem estão disponíveis para capturar essas relações, entre eles:

Modelos lineares: como regressão linear múltipla, indicados para relações diretas e conjuntos de dados com número limitado de variáveis;

Árvores de decisão e regressão: como Random Forest e Gradient Boosting, que lidam bem com grandes conjuntos de dados e relações complexas;

Máquinas de vetores de suporte (SVM): eficazes em cenários com alta dimensionalidade e separações não lineares;

Redes neurais artificiais: úteis para modelagens mais sofisticadas, ainda que exigentes em calibração e volume de dados.

Esses modelos podem ser utilizados de forma individual ou em conjunto (ensemble), combinando diferentes algoritmos para melhorar o desempenho geral. Um exemplo clássico é o próprio Random Forest, que opera como um conjunto de árvores de decisão, integrando múltiplas predições para gerar um resultado mais estável e preciso.

Alguns modelos têm desempenho superior com grandes volumes de dados (como Random Forest e SVM), enquanto outros, como a regressão linear, podem ser mais adequados em situações com menor número de variáveis ou menor complexidade de relacionamentos.

Em modelagens verticais (por camadas), é comum aplicar predição sequencial, utilizando os valores preditos da camada mais superficial (por exemplo, 0–10 cm) como covariável adicional para a predição da camada seguinte (10–20 cm), e assim por diante. Essa abordagem encadeada busca preservar a coerência vertical do perfil

do solo, o que é especialmente importante para atributos como granulometria, carbono orgânico ou umidade, cuja distribuição ao longo da profundidade depende fortemente das camadas superiores.

Além da dimensão vertical, o MDS pode incorporar também a dimensão temporal, permitindo a modelagem espaço-temporal de atributos do solo. Para isso, são utilizados dados de solo e covariáveis com referência temporal (ex.: séries históricas de NDVI, precipitação, uso da terra), bem como mapas anteriores da variável-alvo, que podem ser incluídos como preditores para anos subsequentes. Essa abordagem possibilita analisar e prever mudanças nos atributos do solo ao longo do tempo, desde que haja dados suficientes para representar essas variações. A predição espaço-temporal depende, portanto, da mudança das covariáveis ao longo do tempo, sendo uma estratégia poderosa para estudos de dinâmica do solo sob diferentes cenários de uso da terra, clima e manejo.

- Avaliação das estimativas

A avaliação das estimativas produzidas por um modelo de MDS é indispensável para garantir sua confiabilidade e utilidade prática. Essa avaliação é realizada por meio da validação, que consiste na comparação entre os valores observados e medidos (dados de campo ou laboratório) e os valores preditos pelo modelo. A partir dessa comparação, obtêm-se métricas quantitativas que descrevem a qualidade das predições em termos de acurácia, precisão, tendência e consistência. As métricas utilizadas variam, conforme a natureza da variável-alvo, se contínua ou categórica.

Para variáveis contínuas, como teores de argila, profundidade do solo ou teor de carbono orgânico, comumente se utiliza o ME (erro médio, que revela a existência de tendência sistemática nas predições), o MAE (erro absoluto médio), O MSE é raiz quadrada no RMSE; RMSE (erro quadrático médio), o MEC (que indica a proporção da variância explicada pelo modelo). Além disso, é possível calcular o 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. As métricas são calculadas pelas seguintes equações:

$$\text{ME} = \sum_{i=1}^{n} \frac{yi - xi}{n} \quad (1)$$

$$\text{MAE} = \frac{\sum_{i=1}^{n} |yi - xi|}{n} \quad (2)$$

$$MSE = \frac{(yi - xi)^2}{n} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (yi - xi)^2}{n}} \quad (4)$$

$$MEC = 1 - \frac{MSE}{\frac{(\underline{xi} - xi)^2}{n}} \quad (5)$$

em que:

yi é o valor predito;

xi é o valor observado;

n é o número de observações;

$\underline{xi}$ 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;

Já para variáveis categóricas, como classes texturais ou classes taxonômicas de solo, utilizam-se métricas derivadas da matriz de confusão, como a acurácia global (AG), a precisão do usuário (AU), relacionada ao erro de comissão. Em MDS a AU considera a proporção de cada classe predita que concorda com os dados de referência, ou seja, representa a acurácia individual de predição de cada classe.

$$AG = N\sum i = 1\ nii / k \quad (6)$$

$$UAi = nii / n + i \quad (7)$$

em que:

nii é o número de acertos na classe ii,

n+i é o total de amostras preditas na classe ii,

N é o total geral de amostras,

k é o número de classes,

A forma como os dados são utilizados na validação influencia a interpretação das métricas obtidas. A validação externa, também chamada de holdout, é considerada a abordagem mais rigorosa. Nela, o modelo é testado com um conjunto de dados completamente independente, que não foi utilizado em nenhuma etapa do treinamento. Esse conjunto idealmente deve ter sido obtido por amostragem probabilística, de modo a representar fielmente a variabilidade da paisagem e permitir avaliar a capacidade do modelo de generalizar para novos locais. No entanto, essa abordagem exige um volume suficiente de dados, para que parte deles possa ser reservada exclusivamente para o teste.

Quando não é possível dispor de dados independentes, aplica-se a validação cruzada (cross-validation), na qual o mesmo conjunto de dados é subdividido em partes e utilizado alternadamente para treino e teste. Essa técnica permite maximizar o uso de conjuntos de dados limitados, sendo especialmente útil quando há poucas observações ou quando os dados são espacialmente dependentes. No entanto, por envolver algum grau de sobreposição entre os dados utilizados para treinar e para testar o modelo, a validação cruzada tende a fornecer estimativas de desempenho ligeiramente mais otimistas, o que deve ser considerado na interpretação dos resultados.

Além da avaliação por métricas de erro, um componente fundamental da modelagem preditiva é a incerteza associada às estimativas. Diferentemente do erro — que depende da existência de observações para comparação direta — a incerteza é uma propriedade interna do modelo, relacionada ao grau de confiança nas predições feitas em cada ponto da área mapeada. A incerteza representa, portanto, a variabilidade esperada das estimativas, considerando a estrutura dos dados, a densidade amostral, a adequação das covariáveis e a complexidade do ambiente modelado.

Para variáveis contínuas, a incerteza pode ser expressa por meio do desvio padrão da predição, intervalos de confiança (por exemplo, 90%) ou quantis de predição (como P5, P50, P95), que indicam a dispersão dos valores possíveis para cada pixel. Já para variáveis categóricas, a incerteza é representada geralmente pela probabilidade de ocorrência de cada classe (por exemplo, 70% de chance de uma área pertencer a uma determinada classe textural ou classe de solo), ou por medidas de entropia, como o índice de confusão (IC), que indicam o grau de indecisão do modelo em cada local (Burrough et al., 1997). O IC pode ser caracterizado por valores de probabilidade produzidos como um subproduto da classificação preditiva. O IC traz uma medida da confusão que o modelo preditivo faz entre as duas classes de solo mais prováveis. O IC varia entre 0 e 1, onde 1 significa máxima confusão (mínima precisão) e 0 a mínima confusão (máxima precisão).

$$IP = percentil\ 90 - percentil\ 10 \quad (6)$$

$$IC = 1 - (\mu maxi - \mu(max\text{-}1)i) \quad (7)$$

em que:

μmaxi = valor de associação da classe i com a máxima probabilidade de ocorrência (μk) no pixel i,

μ(max -1)i = segundo maior valor de associação no mesmo pixel i,

μk = valor de probabilidade da classe i atribuído em k classes.

A comunicação clara da incerteza é fundamental para o uso ético e eficiente dos produtos do MDS. Ela permite ao usuário avaliar a robustez local das predições, identificar regiões com maior ou menor confiabilidade e, quando necessário, planejar novas coletas de dados para reduzir a incerteza nas áreas mais críticas. Assim, a incerteza é não apenas uma limitação, mas também uma ferramenta estratégica de interpretação e tomada de decisão, sendo indispensável em qualquer aplicação prática de mapas digitais de solo.

3. APLICAÇÃO DE MDS NO MAPEAMENTO DE ATRIBUTOS DOS SOLOS DO MARANHÃO

Neste item, será apresentado um exercício prático de mapeamento digital da granulometria dos solos do Estado do Maranhão, com foco na predição das frações areia, silte e argila, propriedades físicas fundamentais para o entendimento do funcionamento e uso do solo. A atividade é estruturada para apresentar, de forma sequencial, os componentes operacionais do MDS — dados de treinamento, modelo preditivo, validação e geração de mapas, conforme apresentado neste capítulo — e acompanha um código em Google Earth Engine (GEE) que permite a reprodução completa do processo.

Serão introduzidas as definições básicas de granulometria e de cada etapa envolvida na modelagem. Será utilizado o algoritmo Random Forest para gerar mapas contínuos das frações texturais nas camadas de 0–30 cm, com resolução espacial de 30 metros. As predições são feitas com base em covariáveis ambientais representativas dos fatores de formação do solo, e o fluxo será exemplificado passo a passo no GEE, permitindo que o leitor compreenda e aplique o MDS com base em princípios conceituais sólidos e ferramentas acessíveis.

Para orientações detalhadas sobre como acessar e utilizar os dados e scripts necessários para reproduzir este exercício, consulte o item “Acesso aos dados” apresentado ao final deste capítulo.

3.1. Dados de treinamento

A base de dados utilizada para o exercício de modelagem corresponde a uma coleção de amostras de solo georreferenciadas, compilada, harmonizada e armazenada sob o nome "matriz_psd_maranhao_v3", acessível via asset no GEE. Essa base representa um subconjunto do repositório SoilData, ajustado para a área do Maranhão, contendo informações laboratoriais das frações granulométricas para 489 amostras (Figura 3), na profundidade de 0–30 cm.

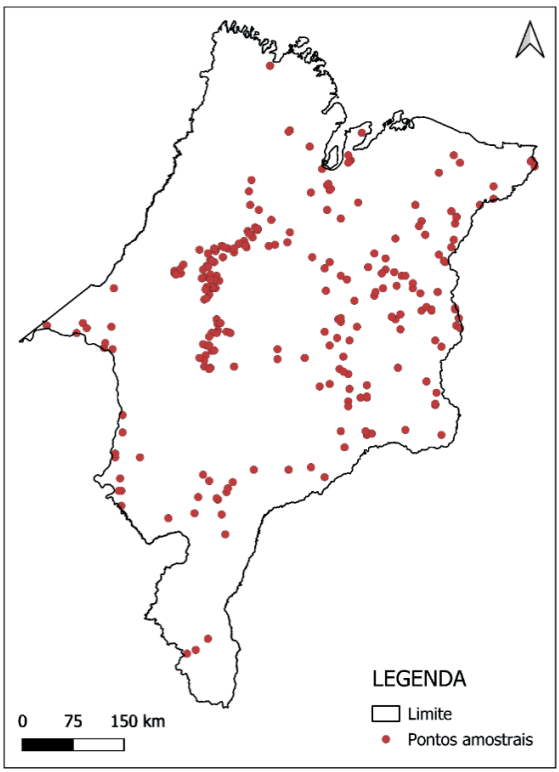

Figura 3. Distribuição espacial dos pontos amostrais utilizados como dados de treinamento no mapeamento da granulometria do estado do Maranhão (n = 489).

A variável-alvo do exercício é a granulometria do solo, definida pelas frações de:

- Areia (sand): partículas com diâmetro entre 2,0 mm e 0,05 mm;
- Silte (silt): partículas entre 0,05 mm e 0,002 mm;
- Argila (clay): partículas com diâmetro inferior a 0,002 mm.

Essas três frações compõem, em conjunto, 100% da massa mineral do solo e, por isso, constituem dados composicionais— isto é, interdependentes. Para evitar redundâncias estatísticas e garantir validade na modelagem, os valores foram transformados em razão logarítmica (Aitchison, 1982; Samuel-Rosa, 2012), utilizando as seguintes variáveis derivadas:

- log_clay_sand: logaritmo da razão entre argila e areia.
- log_silt_sand: logaritmo da razão entre silte e areia;

Essas variáveis derivadas são oriundas de um pré-processamento pela implementação da transformação log-ratio aditiva (ALR), sendo iniciado com a escolha de um denominador comum entre as frações texturais do solo. Neste estudo, a fração areia foi selecionada como referência para a transformação. Em seguida, são calculadas duas razões para cada amostra de terra: a razão entre argila e areia, e entre silte e areia. A essas razões aplica-se o logaritmo natural, resultando nos componentes ALR1 e ALR2, conforme a equação:

$$ALR1 = ln\frac{Argila}{Areia}$$

$$ALR2 = ln\frac{Silte}{Areia}$$

Assim, ALR1 e ALR2 passam a ser utilizadas como variáveis-alvo em modelos preditivos, cada um ajustado para predição separadamente ALR1 e ALR2. Os resultados da modelagem geram dois mapas intermediários: um para ALR1 e outro para ALR2. Para interpretar os resultados em termos de proporções originais de areia, silte e argila, é necessário reverter a transformação ALR. Onde, a primeira etapa consiste na exponenciação dos valores preditos convertendo ALR1 e ALR2 de volta às razões originais:

$$Seja\ R_1 = e^{ALR1}\ \ e\ R_2 = e^{ALR2}$$

Então,

$$Areia = \frac{1}{R_1 + R_2 + 1} x100$$

$$Silte = \frac{R_2}{R_1 + R_2 + 1} x100$$

$$Argila = \frac{R_1}{R_1 + R_2 + 1} x100$$

A seguir, é apresentado um exemplo da aplicação de dados composicionais em uma amostra de solo com porcentagens de 50% (500 g/kg) de argila, 30% (300 g/kg) de silte e 20% (200 g/kg) de areia. Primeiramente, calcula-se o logaritmo natural das razões das frações texturais em relação à areia:

$$ALR1 = ln\frac{0{,}50}{0{,}20} = ln(2{,}5) \approx 0{,}9163$$

$$ALR2 = ln\frac{0{,}30}{0{,}20} = ln(1{,}5) \approx 0{,}4055$$

Por meio da função exponencial os valores preditos serão convertidos de ALR1 e ALR2 para as razões originais, valores para os três mapas (percentagem de areia, silte e argila).

$$onde,\ R_1 = e^{0{,}9163} \approx 2{,}5$$

$$R_2 = e^{0{,}4055} \approx 1{,}5$$

Então,

$$Areia = \frac{1}{2{,}5 + 1{,}5 + 1} x100 = 20\%$$

$$Silte = \frac{1{,}5}{2{,}5 + 1{,}5 + 1} x100 = 30\%$$

$$Argila = \frac{2{,}5}{2{,}5 + 1{,}5 + 1} x100 = 50\%$$

Essa transformação é uma prática recomendada em modelagem de dados composicionais, pois elimina a restrição da soma constante e permite o uso de algoritmos tradicionais de regressão e classificação. Os dados de entrada foram filtrados para garantir consistência mínima, removendo registros com valores nulos ou fora do intervalo lógico, e os identificadores únicos (id) e índices foram preservados para fins de rastreabilidade.

3.2. Modelo preditivo

Para a predição das frações granulométricas dos solos do estado do Maranhão, foi adotado o algoritmo Random Forest, uma técnica de aprendizado de máquina baseada em um conjunto de árvores de decisão. O modelo foi escolhido por sua capacidade de lidar com conjuntos de dados complexos, com múltiplas covariáveis e

relações não lineares, além de sua robustez frente a dados ruidosos e sua flexibilidade para modelar variáveis contínuas. Ele captura a relação entre os dados pontuais e as covariáveis ambientais.

- Covariáveis Ambientais do modelo SCORPAN

O desempenho do modelo preditivo depende fortemente da escolha das covariáveis ambientais, que devem representar proxies digitais plausíveis dos fatores de formação do solo. Abaixo, são descritas as covariáveis utilizadas no modelo SCORPAN, sua natureza (contínua ou categórica), sua origem e a relação pedologicamente esperada com a granulometria, que está sendo aqui exemplificada:

- Covariáveis topográficas (fator R – relevo)

Elevação (elevation) – (Yamazaki et al., 2017) variável contínua, Modelo Digital de Elevação (MDE); relação com a litologia, influencia os processos de erosão, deposição e estabilidade dos solos. Solos em áreas mais elevadas tendem a ser mais erodidos, mais rasos. Dependendo do material de origem, com maior proporção de areia.

Declividade (slope) – (Amatulli et al., 2020, 2018) variável contínua: reflete o grau de inclinação do terreno. Áreas mais inclinadas tendem a apresentar menor acúmulo de material fino (silte e argila), favorecendo a ocorrência de solos rasos e arenosos.

Índice de posição topográfica (TPI) – (Amatulli et al., 2020, 2018): variável contínua: indica se um ponto está em crista, encosta ou vale. Vales favorecem deposição e acúmulo de materiais finos.

Curvatura do relevo (curvature) – (Amatulli et al., 2020, 2018): variável contínua: identifica zonas de concavidade/convexidade que influenciam a concentração de água e sedimentos, o que afeta a distribuição textural.

Rugosidade do relevo (roughness) – (Amatulli et al., 2020, 2018): variável contínua: indica a variação da elevação em uma vizinhança, refletindo a complexidade da superfície. Superfícies mais rugosas tendem a ter maior variabilidade nos processos de erosão, infiltração e deposição, afetando a redistribuição de partículas do solo.

Índice de convergência do relevo (convergence) – (Amatulli et al., 2020, 2018): variável contínua: representa a tendência de convergência ou divergência do fluxo superficial da água. Regiões com valores positivos favorecem o acúmulo de água e sedimentos (vales), enquanto valores negativos indicam áreas de escoamento (cristas e encostas), influenciando diretamente a distribuição textural do solo.

Índice de potência do fluxo (SPI – Stream Power Index) – (Amatulli et al., 2020, 2018): variável contínua: relaciona declividade e área de contribuição, estimando o poder erosivo do fluxo superficial. Áreas com SPI elevado tendem a apresentar maior remoção de partículas finas, influenciando negativamente no acúmulo de silte e argila.

- Covariáveis espectrais e vegetação (fatores O – organismos e N – localização)

Biomas brasileiros – (IBGE, 2019b): variável categórica, identifica os biomas oficiais do Brasil (ex: Amazônia, Cerrado, Mata Atlântica), que estão associados a diferentes regimes climáticos, tipos de vegetação e uso da terra, influenciando indiretamente a formação e distribuição dos solos.

Fitofisionomias – (IBGE, 2023, 2012): variável categórica, descreve a fisionomia da vegetação (ex: floresta estacional, savana, campo limpo), sendo um indicador da cobertura vegetal potencial e histórica, que pode refletir o grau de intemperismo e o estágio de desenvolvimento do solo.

Índices minerais espectrais – (Landsat 5, 7, and 8): variável contínua, derivados de imagens Landsat, são índices espectrais desenvolvidos para detectar a presença relativa de minerais específicos na superfície do solo.

Minerais de argila (clay minerals) - (Landsat 5, 7, and 8): relacionado à presença de minerais de argila na superfície do solo, geralmente associados a maior intemperismo e acúmulo de partículas finas.

Óxidos de ferro (oxides) - (Landsat 5, 7, and 8): indica concentração superficial de óxidos (ex: hematita, goethita), que estão relacionados à coloração, drenagem e grau de intemperismo do solo.

- Covariáveis geológicas e pedológicas (fator P – material de origem)

Classe litológica (Províncias geológicas) – (IBGE, 2019): variável categórica proveniente de mapeamentos geológicos, que identifica o tipo de rocha ou sedimento que deu origem ao solo. Rochas ou materiais mais intemperizados, como os arenitos, tendem a originar solos arenosos. Já rochas máficas (como basaltos) ou sedimentos argilosos geralmente formam solos de textura mais fina, como os argilosos.

Mapas de solos pré-existentes (World Reference Base for Soil Resources) – (Hengl et al., 2017): variável categórica: pode ser utilizado como covariável proxy do fator "s", fornecendo conhecimento prévio incorporado ao modelo.

Black Soils (Probabilidade de ocorrência) – (FAO, 2022): variável categórica: representa áreas com maior chance de ocorrência de solos escuros e ricos em matéria orgânica.

- Covariáveis climáticas (fator C – clima)

Classificação climática de Köppen – (Alvares et al., 2013) variável categórica: sistema climático que classifica regiões com base em temperatura e precipitação ao longo do ano. No código, são utilizadas três resoluções: Köppen L1: classes principais (ex: Af, Cfb, Aw); Köppen L2 e L3: subcategorias mais detalhadas. Essa variável é útil para capturar os efeitos do clima no desenvolvimento dos solos, como intemperismo, lixiviação e acúmulo de matéria orgânica.

- Covariáveis espaciais (fator N – posição)

Latitude e longitude – variáveis contínuas: utilizadas como proxies da localização geográfica, que capturam gradientes espaciais não explicitamente representados pelas outras covariáveis.

Recorrência de água – (MAPBIOMAS PROJECT, 2025): variável contínua: representa a frequência com que um local apresentou presença de água ao longo da série histórica disponível de 39 anos (1985–2023), indicando áreas sujeitas a alagamentos, planícies aluviais ou presença de corpos d'águas intermitentes, que influenciam a deposição de materiais e a textura do solo.

3.3. Parametrização do modelo preditivo

A modelagem foi realizada separadamente para cada uma das duas variáveis-alvo transformadas: log_silt_sand e log_clay_sand, referentes às camadas de 0-30 cm. O modelo foi parametrizado com um número fixo de árvores (estimadores) e profundidade máxima controlada para evitar sobre-ajuste. Os hiperparâmetros ajustados incluíram: ntree, número total de árvores na floresta (100), mtry, quantidade de variáveis testadas em cada divisão (16), nodesize, número mínimo de amostras necessárias para dividir um nó (2), maxNodes, limite máximo de nós por árvore (30), e sampsize, proporção da amostra utilizada para treinar cada árvore (0,632). O conjunto de covariáveis ambientais — previamente harmonizado e ajustado à resolução espacial de 30 metros — foi utilizado como entrada para o treinamento. A etapa de modelagem foi executada inteiramente na plataforma GEE, utilizando os recursos de computação paralela para processar grandes volumes de dados ambientais.

3.4. Predição espacial e mapeamento da granulometria do solo do estado do Maranhão

Após o treinamento, o modelo foi aplicado sobre a área de interesse para gerar mapas contínuos das razões logarítmicas entre frações granulométricas em resolução de 30 metros (Figura 4). Em seguida, os valores preditos foram transformados de volta para as frações originais (areia, silte e argila) por meio da inversão da log-ratio, respeitando a composição fechada dos dados e garantindo que as frações somem 100% (Figuras 5, 6 e 7). Essa abordagem permitiu integrar dados de solos e informações ambientais e algoritmos modernos de predição para gerar produtos espacialmente explícitos, com potencial de aplicação em diferentes escalas de gestão do território.

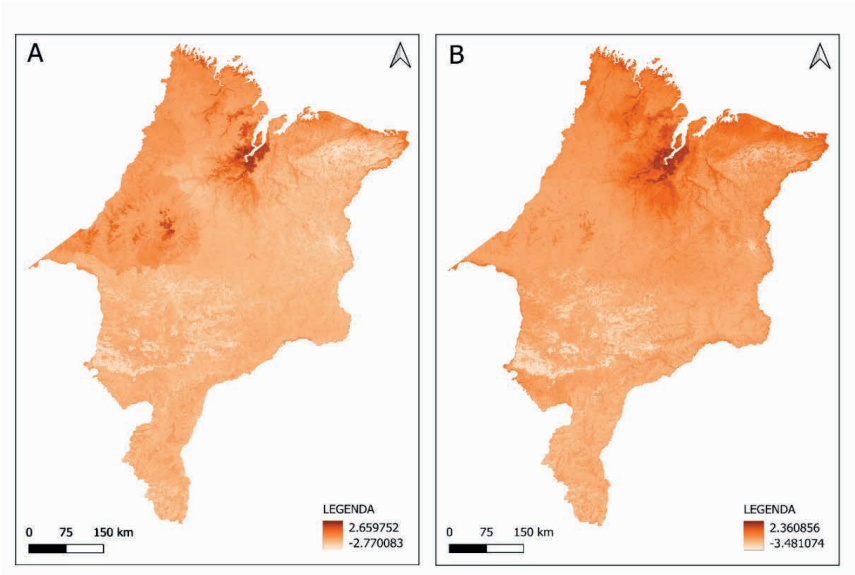

Figura 4. Predição espacial das log ratio - razões entre areia e argila (A) e areia e silte (B)

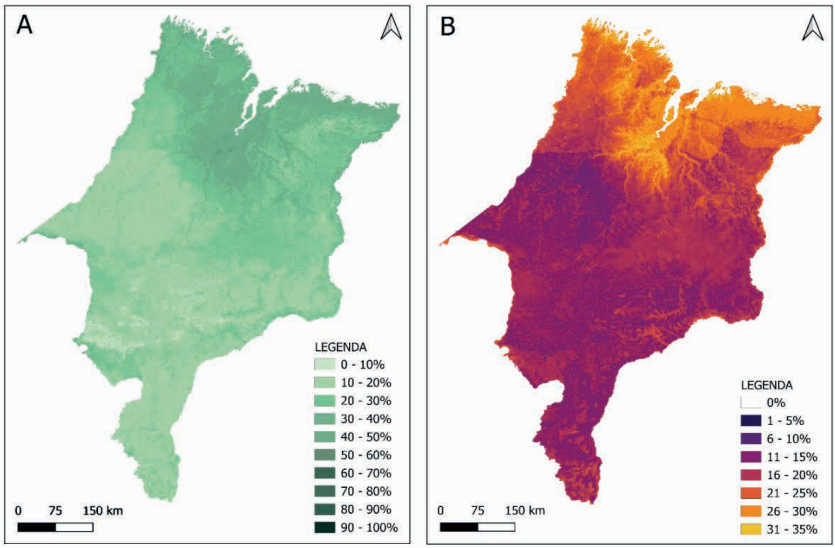

Figura 5. Predição espacial do teor de silte (A) e estimativa das incertezas associadas, representada pelo desvio padrão (B).

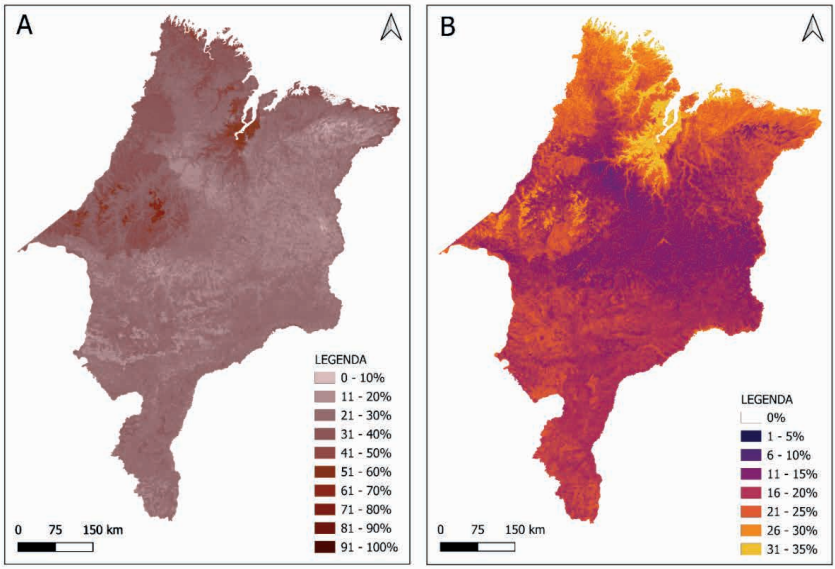

Figura 6. Predição espacial do teor de argila (A) e estimativa das incertezas associadas, representada pelo desvio padrão (B).

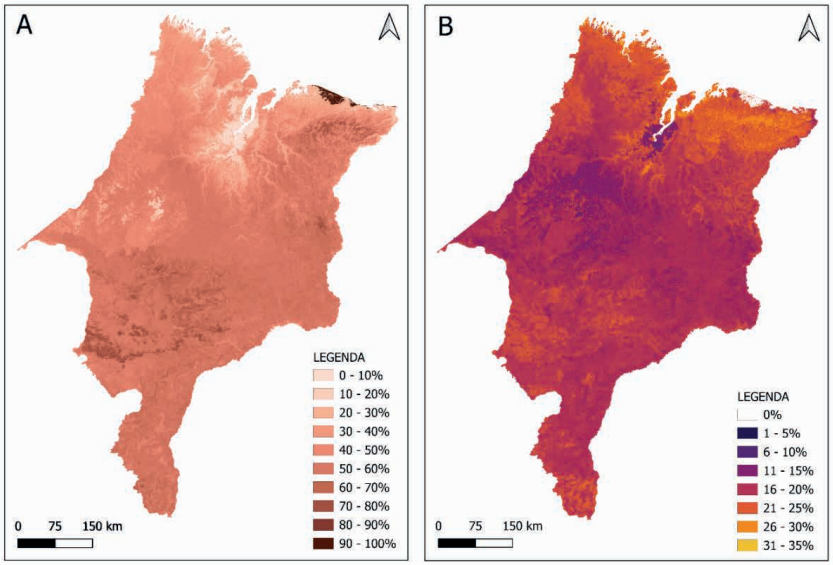

Figura 7. Predição espacial do teor de areia (A) e estimativa das incertezas associadas, representada pelo desvio padrão (B).

3.5. Avaliação dos modelos e mapas da granulometria

A avaliação do modelo de predição da granulometria foi realizada por meio de validação cruzada, técnica escolhida em função da quantidade limitada de dados disponíveis para o Maranhão, o que inviabilizaria a separação de um conjunto independente para a validação externa. A validação cruzada consiste em particionar os dados em subconjuntos de treino e teste utilizados de forma alternada, permitindo que todas as amostras contribuam tanto para o ajuste quanto para a avaliação do modelo. Apesar de fornecer estimativas ligeiramente mais otimistas, essa abordagem é amplamente aceita quando se busca aproveitar ao máximo os conjuntos de dados escassos ou heterogêneos disponíveis.

As métricas de desempenho dos modelos foram calculadas com base nas variáveis transformadas (log_silt_sand e log_clay_sand), e incluíram ME, MAE, MSE, RMSE, MEC e slope. As avaliações foram realizadas tanto por camada (Tabela 1) quanto pela média dos perfis de solo (Tabela 2). Esses indicadores permitiram quantificar a precisão das estimativas das razões entre as frações texturais, assegurando consistência na avaliação do desempenho do modelo.

Tabela 1. Estatísticas de validação por camada de solo (%) — n = 404 camadas de 235 perfis de solo

	ME	MAE	MSE	RMSE	MEC	Slope
argila	0,0	7,1	110,7	10,5	0,8	1,0
silte	0,4	6,3	105,8	10,3	0,5	0,9
areia	-0,4	10,1	213,7	14,6	0,7	1,0

Em que: ME: erro médio; MAE: erro médio absoluto; MSE: erro quadrático médio; RMSE: raiz quadrada do erro médio; MEC: coeficiente de desempenho do modelo, slope: declividade da regressão.

Tabela 2. Estatísticas de validação médias por perfil de solo (%) — n = 235 perfis de solo

	ME	MAE	MSE	RMSE	MEC	Slope
argila	0,6	6,7	97,6	9,9	0,8	1,1
silte	0,0	0,8	111,6	10,6	0,5	1,0
areia	-0,6	10,5	222,0	14,9	0,6	1,0

Em que: ME: erro médio; MAE: erro médio absoluto; MSE: erro quadrático médio; RMSE: raiz quadrada do erro médio; MEC: coeficiente de desempenho do modelo, slope: declividade da regressão.

Para a geração dos mapas, o algoritmo Random Forest produziu múltiplas árvores de decisão, e cada árvore gerou uma predição independente para cada pixel da área de interesse. A média das predições foi utilizada como valor final da estimativa em cada pixel. O desvio padrão entre as predições individuais das árvores foi utilizado como uma medida de incerteza associada à predição. Esse desvio padrão expressa a variabilidade interna do modelo em relação à mesma entrada de dados, indicando em quais regiões do espaço as predições são mais dispersas em relação às médias e, portanto, mais incertas.

Assim, os **produtos** finais do exemplo incluem tanto os mapas preditos das frações granulométricas, quanto os respectivos mapas de incerteza (Figuras 4, 5 e 6), permitindo uma análise espacialmente explícita da confiabilidade das estimativas em cada ponto da paisagem.

4. APLICAÇÃO DO MDS NO MAPEAMENTO DE CLASSES DE SOLOS DO MARANHÃO

Neste item, será apresentado um exemplo prático de mapeamento digital de Classes de Solo para o estado do Maranhão. As etapas envolveram, de forma sequencial, os componentes operacionais do MDS — dados de treinamento, covariáveis preditoras, modelo preditivo, validação e geração de mapas, conforme amplamente descrito no item 3. As predições foram feitas com base em covariáveis

ambientais representativas dos fatores de formação do solo (conforme item 3.2.1). O processamento das covariáveis ambientais e as predições foram realizadas em ferramentas acessíveis, Quantum GIS v3.4.11 (https://qgis.org/) e ambiente R (R Core Team, 2021).

Para orientações detalhadas sobre como acessar e utilizar os dados e scripts necessários para reproduzir este exercício, consulte o item "Acesso aos dados" apresentado ao final deste capítulo.

4.1. Dados de treinamento

A base de dados utilizada para modelagem de classes de solos corresponde a uma coleção de perfis de solo georreferenciados, compilada do repositório SoilData de estudos pedológicos realizados entre os anos de 1973 a 2024 (Figura 8). A informação de cada perfil de solo referente à classificação taxonômica, foi atualizada até o segundo nível categórico, conforme o Sistema Brasileiro de Classificação de Solos, SiBCS (Santos et al., 2018).

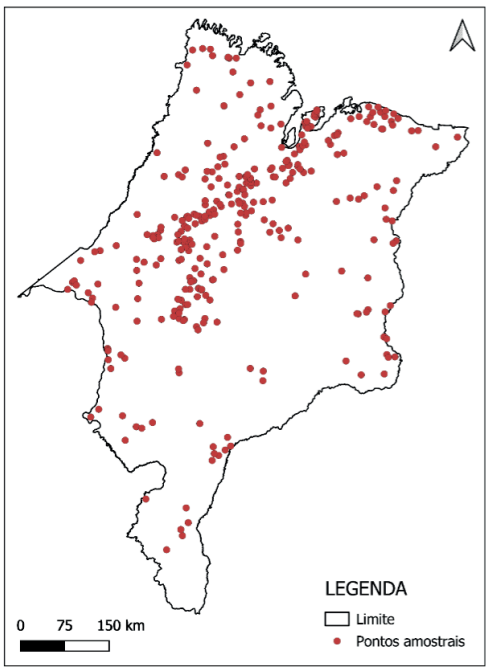

Figura 8. Distribuição espacial dos pontos amostrais utilizados como dados de treinamento no mapeamento de classes de solo do estado do Maranhão (n = 364).

Para o estado do Maranhão, foram compilados 222 perfis de solos, obtidos de diversas fontes. Foi observada pouca representatividade das seguintes classes de solos dentre os 222 perfis: Cambissolos Háplico (1 perfil), Gleissolos Melânicos (2 perfis), Gleissolos Tiomórficos (2 perfis), Latossolo Vermelho (1 perfil), Luvissolo Háplico (2 perfis), Plitossolos Pétricos (1 perfil), Organossolos Háplico (1 perfil), Vertissolos Hidromórfico (3 perfis). Então, optou-se por eliminar os perfis dessas classes, antes de proceder com a etapa de desenvolvimento do modelo preditivo, por não haver número suficiente de perfis para o processamento matemático. Isso resultou em um conjunto de 209 perfis de solos. Somado a essa base de dados, foram realizadas 155 pseudo-amostragens.

O processo de pseudo-amostragens foi realizado manualmente com base no mapa de solos do Maranhão, escala 1:1.000.000 (Embrapa, 1986) e na relação solo-paisagem do estado do Maranhão apresentada e discutida por Mendonça-Santos et al. (2020). Essas observações, somadas aos 209 perfis, totalizaram 364 observações contendo as classes taxonômicas de solos (2º nível categórico do SiBCS), as quais foram utilizadas para o exercício de MDS de classes de solo apresentado neste capítulo. Na tabela 3 é apresentada a distribuição do número de observações de perfil de solo e a pseudo-amostragem para cada classe de solo.

Tabela 3. Simbologia das classes de solos no 2º nível categórico do SiBCS e distribuição do número de perfis de solo e pseudo-amostragens de cada classe de solo das 364 observações utilizadas na predição e mapeamento das classes de solo do estado do Maranhão.

Subordem do SiBCS	Número de perfis/pseudo-amostragem (total)	% de observações no conjunto de dados
Argissolo Amarelo - PA	11/22 (33)	9
Argissolo Vermelho - PV	10/12 (22)	6
Argissolo Vermelho-Amarelo - PVA	52/13 (65)	18
Espodossolo - E	2/5 (7)	2
Gleissolo Háplico GX	19/15 (34)	9
Latossolo Amarelo LA	20/31 (51)	14
Neossolo Litólico - RL	5/13 (18)	5
Neossolo Flúvico - RY	4/18 (22)	6
Neossolo Quartzarênico - RQ	12/26 (38)	10
Plintossolo Háplico - FX	17/8 (25)	7
Plintossolo Argilúvico - FT	27/22(49)	13
Total	364	100

4.2. Modelo preditivo e covariáveis ambientais

O modelo preditivo adotado foi o mesmo algoritmo adotado na predição da granulometria – Random Forest. As covariáveis utilizadas como preditoras foram: fator R – topográficas (elevação, declividade, índice de posição topográfica, rugosidade do relevo, índice de convergência, índice de potência do fluxo, índice de umidade do terreno, curvatura planar, curvatura de perfil, curvatura geral, distância da rede de drenagem, profundidade do vale, formas do terreno, textura do terreno), fator organismos – vegetação (índice de vegetação por diferença normalizada - NDVI), fator material de origem – geológicas (classe litológica - Províncias geológicas), fator clima – climáticas (classificação climática de Köppen), fator N – posição espacial (latitude e longitude, recorrência de água).

4.3. Parametrização do modelo preditivo e mapeamento das classes de solos do estado do Maranhão

Na implementação do modelo de RF, foram ajustados os seguintes hiperparâmetros: ntree, número de árvores na floresta (500), mtry, quantidade de variáveis testadas em cada divisão (12), nodesize, número mínimo de amostras necessárias para dividir um nó (3), e sampsize, proporção da amostra utilizada para treinar cada árvore (7). O conjunto de covariáveis ambientais foi utilizado como entrada para o treinamento. Após essa etapa o modelo foi aplicado sobre a área de interesse para gerar o mapa de classes de solo em resolução de 30 metros (Figura 9). O resultado é o mapa de probabilidade individual de ocorrência de cada classe de solo (Figura 9), computado pela classe de maior probabilidade de ocorrência no píxel para gerar o mapa de classes de solos (Figura 10). Toda a etapa de modelagem foi executada inteiramente no ambiente R (R Core Team, 2021).

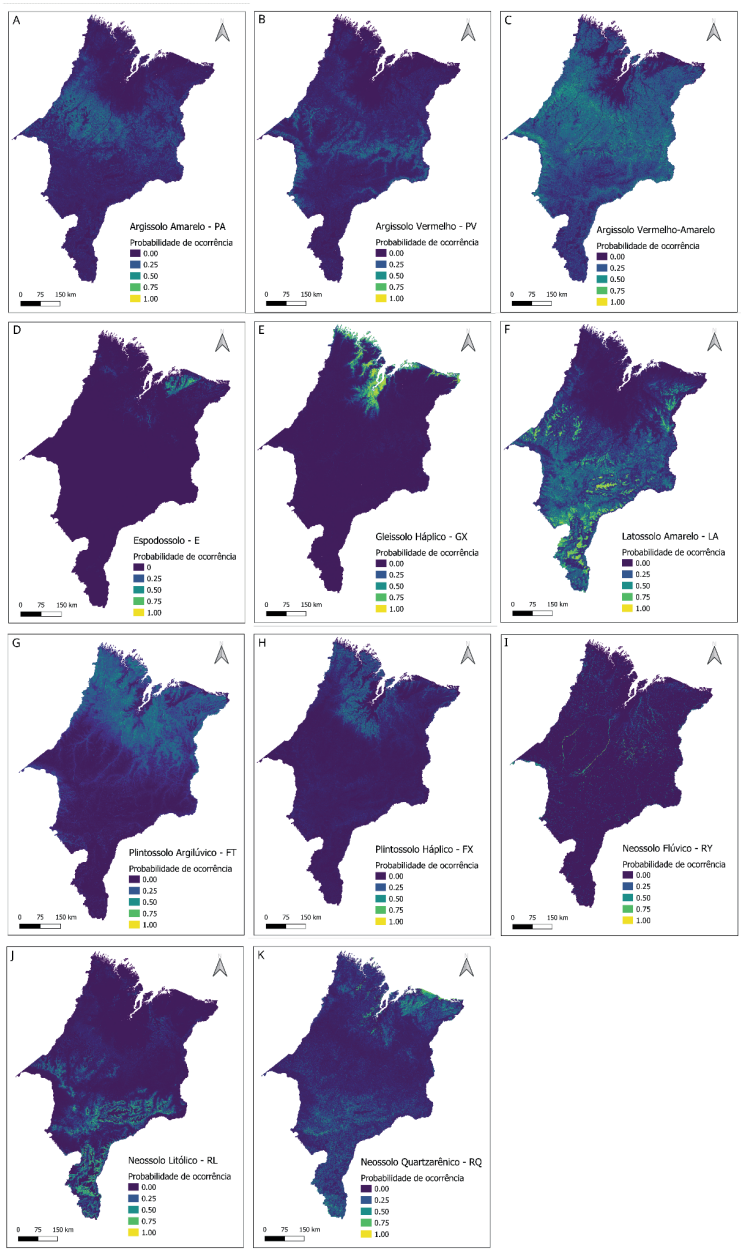

Figura 9. Probabilidade de ocorrência de cada classe de solo. A) Argissolo Amarelo, B) Argissolo Vermelho, C) Argissolo Vermelho-Amarelo, D) Espodossolo, E) Gleissolo Háplico, F) Latossolo Amarelo, G) Plintossolo Argilúvico, H) Plintossolo Háplico, I) Neossolo Litólico, J) Neossolo Flúvico, K) Neossolo Quartzarênico.

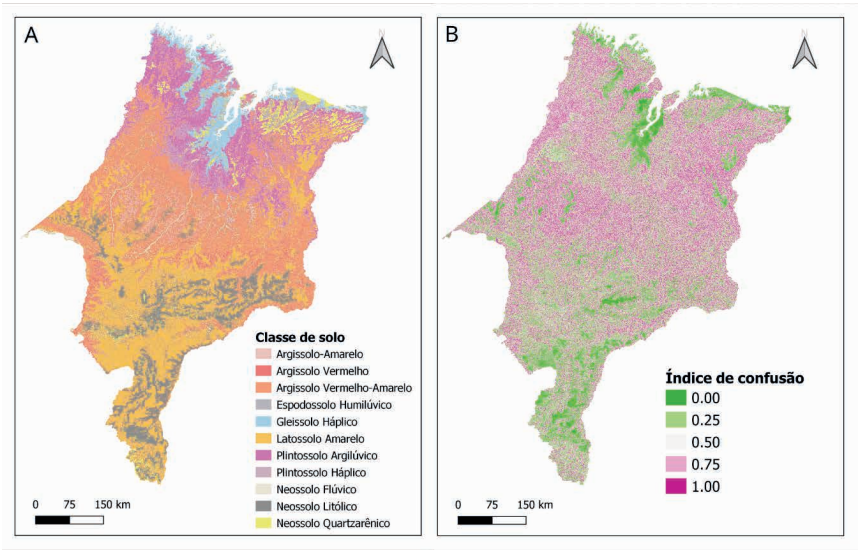

Figura 10. Predição e mapeamento de classes de solos do estado do Maranhão (A) e mapa de incertezas associadas – índice de confusão (B).

5. AVALIAÇÃO DO MODELO DE PREDIÇÃO E DO MAPA PRODUZIDO

A avaliação do modelo de predição de classes de solo foi realizada por meio de validação cruzada. Foi escolhida essa técnica devido à quantidade limitada de observações disponíveis para cada classe de solo no conjunto de dados do Maranhão e por tratar-se de variável-alvo categórica, classes de solos. Foram geradas as matrizes de confusão, obtendo-se os valores de acurácia global – AG e acurácia da classe - AC. A AG foi calculada para avaliar a proporção de píxeis corretamente preditos em relação ao número de píxeis totais nos mapas; e a AC foi utilizada para avaliar a proporção de píxeis corretos de cada classe de solo (Congalton, 1991). A avaliação da incerteza do modelo de predição foi realizada por meio da estatística do índice de confusão - IC (Burrough et al., 1997), o qual varia entre 0 e 1, onde 1 significa máxima confusão (mínima precisão) e 0 a mínima confusão (máxima precisão), conforme Tabela 4.

Tabela 4 - Acurácia por classe de solo (AC) e Acurácia global (AG) na validação cruzada.

Classe de Solo	Acurácia por classe (%)
Argissolo Amarelo - PA	56
Argissolo Vermelho - PV	55
Argissolo Vermelho-Amarelo - PVA	56
Espodossolo - E	78
Gleissolo Háplico GX	85
Latossolo Amarelo LA	81
Neossolo Litólico - RL	94
Neossolo Flúvico - RY	91
Neossolo Quartzarênico - RQ	67
Plintossolo Háplico – FX	58
Plintossolo Argilúvico – FT	67
Acurácia global – AG (%)	56

O mapa predito de classes de solos (Figura 10A) apresentou uma acurácia global (AG) de 56%. Em relação à acurácia de cada classe de solo, para os Neossolo Litólicos foi observada uma maior acurácia (94%), seguido dos Neossolos Flúvicos (91%), Gleissolos Háplicos (85%), Latossolos Amarelos (81%), Espodossolos (78%), Neossolos Quartzarênicos e Plitossolos Argilúvicos (67%), Plintossolos Háplicos (58%), Argissolos Amarelos e Argissolos Vermelho-Amarelos (56%) e Argissolos Vermelhos (55%). As acurácias mais baixas foram encontradas para as classes dos Argissolos Amarelos, Argissolos Vermelhos e Argissolos Vermelho-Amarelos. A menor AC observada para algumas classes de solo pode ser explicada pela baixa representatividade de perfis no conjunto de dados, somado-se à ocorrência de duas ou mais classes em ambientes semelhantes (mesmos intervalos de valores das covariáveis preditoras) na paisagem, por exemplo, Argissolos Amarelos e Argissolos Vermelho-Amarelos, Plitossolos Argilúvicos e Plintossolos Háplicos. Segundo Taghizadeh-Mehrjardi et al. (2015), a distribuição espacial e número de amostras representativas de cada classe de solo influencia a qualidade dos mapas digitais de classe de solos. Isso foi observado também por ten Caten et al. (2011) e Moura-Bueno et al. (2019) no MDS no Estado do Rio Grande do Sul, em que classes menos representativas tiveram sua predição comprometida.

No mapa de incertezas da predição de classes de solos (Figura 10B), pode-se observar os locais da paisagem onde o modelo de predição foi mais preciso (valores próximos de "zero") e menos preciso (valores próximos de "um"). As áreas com maior incerteza estão relacionadas à baixa representatividade de perfis de solo, com destaque para as classes de solos dos Argissolos Vermelhos (PV) e Argissolos Amarelos (PA), que apresentaram a menor acurácia da predição. Além disso, essas classes

ocorrem em posições semelhantes da paisagem, sendo de difícil discriminação por modelos matemáticos e até mesmo por pedólogos no campo, quando existe escassez de perfis de solo na base de dados. Nota-se que nas áreas onde predominam a classe dos Latossolos Amarelos (LA), Gleissolos Háplicos (GX) e Neossolos Litólicos (Figura 10A), as incertezas são menores (valores próximos de "um") no mapa da Figura 10B.

6. BOAS PRÁTICAS EM MAPEAMENTO DIGITAL DE SOLOS

A adoção do MDS tem se expandido em diferentes contextos — acadêmico, produtivo e institucional — devido à sua capacidade de gerar informações espaciais detalhadas sobre atributos do solo. No entanto, para que seus resultados sejam confiáveis, úteis e reprodutíveis, é necessário seguir um conjunto de boas práticas que abrangem desde o planejamento amostral até a divulgação dos produtos finais.

Uma boa prática fundamental é garantir que a variável-alvo esteja claramente definida e seja pedologicamente significativa. A escolha de uma propriedade ou função do solo para a modelagem deve estar ancorada na compreensão da sua relação com os fatores ambientais. Além disso, a natureza da variável (contínua ou categórica) determinará os métodos mais adequados de modelagem, validação e representação da incerteza.

No que se refere aos dados de solo, é essencial assegurar que sejam obtidos por métodos laboratoriais reconhecidos, com georreferenciamento preciso e metadados bem documentados. O uso de dados legados pode ser vantajoso, desde que haja cuidados com a harmonização e a rastreabilidade. A densidade e a distribuição espacial das amostras devem ser planejadas visando capturar a variabilidade ambiental, e não simplesmente atingir um número arbitrário de pontos.

As covariáveis ambientais utilizadas devem ter relação plausível com os processos pedogenéticos e serem escolhidas com base em critérios teóricos, além de análises exploratórias. Covariáveis redundantes, fortemente correlacionadas ou irrelevantes podem comprometer a eficiência do modelo e aumentar o risco de overfitting. A seleção pode ser orientada por conhecimento especializado, testes estatísticos, análise de importância de variáveis ou técnicas automatizadas de redução de dimensionalidade.

Na etapa de modelagem, recomenda-se iniciar com abordagens simples e compreensíveis, avançando para algoritmos mais complexos, à medida que se compreendem os dados e suas relações. A transparência no uso dos algoritmos, incluindo seus parâmetros e processos de ajuste, é essencial para garantir a reprodutibilidade do estudo. Sempre que possível, a estrutura dos roteiros (scripts) e a documentação completa do fluxo de trabalho devem ser disponibilizadas em repositórios acessíveis.

A validação do modelo não deve ser tratada como etapa opcional. O uso de validação externa, quando viável, oferece maior robustez à avaliação. Quando se opta pela validação cruzada, é importante relatar claramente as limitações e os possíveis vieses. Além das métricas de desempenho, a inclusão de mapas de incerteza amplia a compreensão dos resultados e permite seu uso mais cauteloso.

Por fim, os produtos gerados — como mapas, gráficos, relatórios e roteirosscripts — devem ser comunicados de forma clara, com legenda, escala, unidades e metadados completos. A explicitação das limitações do estudo, como áreas com alta incerteza ou baixa densidade amostral, é um sinal de rigor científico e fortalece a credibilidade dos resultados. A adoção de princípios de ciência aberta, como o compartilhamento de dados e métodos, contribui para o avanço coletivo da ciência do solo e favorece a replicação e a melhoria contínua das abordagens.

7. CONSIDERAÇÕES FINAIS

O MDS representa uma abordagem moderna e eficiente para transformar observações pontuais em informações espaciais contínuas, com potencial de aplicação em diversas escalas e contextos e com economia de tempo e recursos. Sua força reside na combinação entre conhecimento pedológico, dados ambientais e métodos quantitativos, permitindo compreender e representar a variabilidade dos solos de forma mais transparente e reprodutível.

Ao aplicar esses princípios à predição da granulometria e classe de solos do Maranhão, foi evidenciado o potencial do MDS como ferramenta de apoio ao planejamento territorial, manejo agrícola e conservação dos recursos naturais. A adoção de boas práticas, como o uso de transformações adequadas, validação cuidadosa, análise da incerteza e fundamentação pedológica das decisões, é o que garante a utilidade e a credibilidade dos mapas gerados. Mais do que um produto final, o MDS deve ser compreendido como um processo contínuo de aprendizado sobre o solo na paisagem, que pode ser melhorado e aperfeiçoado com a disponibilidade de mais informações de solos para se ter uma maior cobertura e distribuição espacial dos pontos amostrados.

8. ACESSO AOS DADOS USADOS NAS MODELAGENS

Os dados utilizados neste capítulo estão disponíveis publicamente em diferentes plataformas. Todos os dados de treinamento encontram-se no SoilData, no Identificador de Objeto Digital (DOI): https://doi.org/10.60502/SoilData/HOIDT7 com conjunto denominado "Mapeamento Digital de Solos do Maranhão: fundamentos, boas práticas e exemplos de mapeamento de classes e atributos". Ele reúne, no arquivo "granulometria_amostras" 489 amostras com teores de areia, silte e argila, além

das variáveis transformadas por log-ratio (log_clay_sand e log_silt_sand). Os scripts correspondentes, escritos em linguagem JavaScript para execução no Google Earth Engine (GEE), estão organizados no repositório GitHub solos-maranhao (https://github.com/taciaraz/solos-maranhao), nas pastas covariate-module, modeling e predicao-espacial. As covariáveis necessárias para a execução destes scripts são carregadas diretamente do repositório MapBiomas Workspace no próprio GEE, dispensando download prévio.

Para o mapeamento de classes de solo, o mesmo conjunto do SoilData, o arquivo "classe_de_solo_amostras", disponibiliza 364 observações com as classes taxonômicas de solos no 2º nível categórico do SiBCS. O processamento é realizado por meio de um script em linguagem R, disponível na pasta classe-solo do mesmo repositório GitHub, que deve ser executado em ambiente R configurado com os pacotes necessários. Nesse caso, as covariáveis utilizadas precisam ser previamente baixadas a partir do endereço https://drive.google.com/drive/folders/1sGOP6-b7p9ERx5tidED47mUN7Z_MD4DC.

Informações detalhadas sobre a organização dos scripts, a execução dos modelos e a lista completa das covariáveis empregadas podem ser consultadas no arquivo README do repositório GitHub disponível em https://github.com/taciaraz/solos-maranhao.

REFERENCIAS

AITCHISON, J. The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological), v. 44, p. 139-177, 1982.

ALVARES, C. A.; STAPE, J. L.; SENTELHAS, P. C.; MORAES GON, J. L. de; SPAROVEK, G. Köppen's climate classification map for Brazil. Meteorologische Zeitschrift, v. 22, p. 711–728, 2013. DOI: https://doi.org/10.1127/0941-2948/2013/0507.

AMATULLI, G. et al. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Scientific Data, v. 5, 180040, 2018. DOI: https://doi.org/10.1038/sdata.2018.40.

AMATULLI, G. et al. Geomorpho90m: empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. Scientific Data, v. 7, 162, 2020. DOI: https://doi.org/10.1038/s41597-020-0479-6.

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

CANCIAN, L. C.; DALMOLIN, R. S. D.; CATEN, A. T. Bibliometric analysis for pattern exploration in worldwide digital soil mapping publications. Anais da Academia Brasileira de Ciências, v. 90, n. 4, p. 3911-3923, 2018.

COELHO, M. L. et al. Systematic review on digital soil mapping in Brazil: current status and future perspectives. Revista Brasileira de Ciência do Solo, v. 45, e0200115, 2021.

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

DOKUCHAEV, V. V. Russian Chernozem. Trad. para o inglês por N. Kaner. Jerusalem: Israel Program for Scientific Translations Ltd. (para USDA-NSF), 1967. Originalmente publicado em 1883.

EMBRAPA. Mapa de Solos do Maranhão. Rio de Janeiro: Embrapa Solos, 1986. Disponível em: https://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/336095.

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

GIASSON, E. et al. Mapeamento digital de solos para a área de estudo do projeto "Usinas da Paz" no Estado do Rio Grande do Sul. Porto Alegre: UFRGS, 2006.

HARTEMINK, A. E.; MCBRATNEY, A.; MENDONÇA-SANTOS, M. de L. (ed.). Digital soil mapping with limited data. Berlim: Springer, 2008.

HENDRIKS, S. et al. Exploring the potential of legacy soil data for digital soil mapping in Brazil. In: International Conference on Digital Soil Mapping, 4., 2019, Wageningen. Wageningen: Wageningen University & Research, 2019.

HENGL, T. et al. SoilGrids250m: global gridded soil information based on machine learning. PLOS One, v. 12, e0169748, 2017. DOI: https://doi.org/10.1371/journal.pone.0169748.

HORST, T. Z. Variação de parâmetros dendrométricos de Pinus taeda L. e a distribuição espacial de atributos do solo por técnicas de mapeamento digital de solos. 2017. 122 f. Dissertação (Mestrado em Ciência do Solo) – Universidade Federal de Santa Maria, Santa Maria, 2017.

HORST, T. Z. et al. Mapeamento digital de solos do Maranhão: fundamentos, boas práticas e exemplos de mapeamento de classes e atributos. SoilData, 2025. DOI: https://doi.org/10.60502/SoilData/HOIDT7. Acesso em: 8 ago. 2025.

IBGE. Manual Técnico da Vegetação Brasileira. 2. ed. rev. e ampl. Rio de Janeiro: IBGE, 2012.

IBGE. Biomas e sistema costeiro-marinho do Brasil: compatível com a escala 1:250 000. Relatórios metodológicos. Rio de Janeiro: IBGE, 2019.

IBGE. Banco de Dados e Informações Ambientais (BDiA): Mapeamento de Recursos Naturais (MRN): Escala 1:250 000. Rio de Janeiro: IBGE, 2023.

JENNY, H. Factors of soil formation: a system of quantitative pedology. New York: McGraw-Hill, 1941.

LAGACHERIE, P. et al. Digital soil mapping: an introductory note for the Geoderma special issue on digital soil mapping. Geoderma, v. 136, n. 1, p. 1-2, 2006.

LAGACHERIE, P.; MCBRATNEY, A. B.; VOLTZ, M. (ed.). Digital soil mapping: an introductory perspective. Amsterdã: Elsevier, 2007.

MAPBIOMAS. Mapeamento anual do estoque de carbono orgânico do solo no Brasil 1985-2021 (coleção beta). Documento de base teórica do algoritmo e resultados. 2023. DOI: https://doi.org/10.58053/MapBiomas/3KXXVV.

MAPBIOMAS PROJECT. MapBiomas General "Handbook" - Algorithm Theoretical Basis Document (ATBD) - Collection 9. 2025. DOI: https://doi.org/10.58053/MapBiomas/ICCL5B.

MCBRATNEY, A. B.; MENDONÇA-SANTOS, M. L.; MINASNY, B. On digital soil mapping. Geoderma, v. 117, n. 3-4, p. 297-324, 2003.

MENDONÇA-SANTOS, M. de L. et al. Aplicação de técnicas de mapeamento digital de solos no âmbito do zoneamento ecológico-econômico do bioma Amazônia no Maranhão. 2020. Disponível em: https://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/1129349.

MENDONÇA-SANTOS, M. L. et al. Digital soil mapping of the state of Maranhão, Brazil. Rio de Janeiro: Embrapa Solos, 2020.

MENDONÇA-SANTOS, M. L. et al. Mapeamento digital do carbono orgânico do solo no estado do Rio de Janeiro. Rio de Janeiro: Embrapa Solos, 2007.

MENDONÇA-SANTOS, M. L.; SANTOS, H. G. The state of the art of Brazilian soil mapping and prospects for digital soil mapping. In: LAGACHERIE, P.; MCBRATNEY, A. B.; VOLTZ, M. (ed.). Digital soil mapping: an introductory perspective. Amsterdã: Elsevier, 2007. p. 487-502.

MENDONÇA-SANTOS, M. de L. et al. Mapeamento digital de solos no estado do Rio de Janeiro, Brasil: dados, modelagem e predição. In: HARTEMINK, A. E.; MCBRATNEY, A.; MENDONÇA-SANTOS, M. de L. (ed.). Digital soil mapping with limited data. Berlim: Springer, 2006. p. 381-396.

MOURA-BUENO, J. M. de et al. Digital soil mapping of soil classes in the state of Rio Grande do Sul, Brazil, using legacy data and Random Forest. Revista Brasileira de Ciência do Solo, v. 43, e0180211, 2019.

MOURA-BUENO, J. M. et al. Prediction of soil classes in a complex landscape in Southern Brazil. Pesquisa Agropecuária Brasileira, v. 54, e00420, 2019.

R CORE TEAM. R: a language and environment for statistical computing. Viena: R Foundation for Statistical Computing, 2021. Disponível em: https://www.R-project.org/. Acesso em: 5 ago. 2025.

SAMUEL-ROSA, A. et al. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma, v. 243, p. 214-227, 2015.

SAMUEL-ROSA, A. et al. SoilData: a free Brazilian repository of soil data. In: Congresso Mundial de Ciência do Solo, 21., 2018, Rio de Janeiro. Rio de Janeiro: Sociedade Brasileira de Ciência do Solo, 2018.

SAMUEL-ROSA, A. Funções de predição espacial de propriedades do solo. 2012. 201 f. Dissertação (Mestrado em Ciência do Solo) – Universidade Federal de Santa Maria, 2012.

SANTOS, H. G. dos et al. Sistema Brasileiro de Classificação de Solos. 5. ed. Brasília: Embrapa, 2018.

STUMPF, R. et al. A novel approach for mapping soil classes using legacy data in the Brazilian Amazon. In: International Conference on Digital Soil Mapping, 2., 2016, Piracicaba. Piracicaba: ESALQ/USP, 2016.

TAGHIZADEH-MEHRJARDI, R. et al. Digital mapping of soil properties in a small catchment in northern Iran. Geoderma, v. 257, p. 167-175, 2015.

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

TEN CATEN, A. et al. Mapeamento digital de classes de solos: características da abordagem brasileira. Ciência Rural, v. 42, p. 1989-1997, 2012.

VASQUES, G. M. et al. Digital soil organic carbon map of Brazil for the Global Soil Organic Carbon Map (GSOCmap). In: International Workshop on Global Soil Organic Carbon Map, 2., 2017, Roma. Roma: FAO, 2017.

YAMAZAKI, D. et al. A high-accuracy map of global terrain elevations. Geophysical Research Letters, v. 44, p. 5844–5853, 2017. DOI: https://doi.org/10.1002/2017GL072874.