UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA

CENTRO DE CIÊNCIAS AGRÁRIAS AMBIENTAIS E BIOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM SOLOS E QUALIDADE DE ECOSSISTEMAS

APRENDIZADO DE MÁQUINA E FLUORESCÊNCIA DE RAIO-X NO MAPEAMENTO DIGITAL DE ATRIBUTOS QUÍMICOS DOS SOLOS EM ÁREA DE COBERTURA PEDOLÓGICA COMPLEXA

ÍCARO BARRETO SOUZA

CRUZ DAS ALMAS
JUNHO - 2022

APRENDIZADO DE MÁQUINA E FLUORESCÊNCIA DE RAIO-X NO MAPEAMENTO DIGITAL DE ATRIBUTOS QUÍMICOS DOS SOLOS EM ÁREA DE COBERTURA PEDOLÓGICA COMPLEXA

ÍCARO BARRETO SOUZA

Engenheiro Agrônomo

Universidade Federal do Recôncavo da Bahia, 2019

Dissertação apresentada ao colegiado do Programa de Pós-Graduação em Solos e Qualidade de Ecossistemas da Universidade Federal do Recôncavo da Bahia, como requisito parcial para obtenção do título de Mestre em Solos e Qualidade de Ecossistemas.

Orientador: Prof. Dr. Francisco Alisson da Silva Xavier

CRUZ DAS ALMAS - BAHIA

JUNHO – 2022

FICHA CATALOGRÁFICA

S729a

Souza, Ícaro Barreto.

Aprendizado de máquina e fluorescência de raio-x no mapeamento digital de atributos químicos dos solos em área de cobertura pedológica complexa / Ícaro Barreto Souza._ Cruz das Almas, BA, 2022.

65f; il.

Dissertação (Mestrado) – Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias, Ambientais e Biológicas, Mestrado em Solos e Qualidade de Ecossistemas.

Orientador: Prof. Dr. Francisco Alisson da Silva Xavier.
Coorientador: Prof. Dr. Everton Luís Poelking.

1.Solos – Manejo. 2.Solos – Mapeamento da cobertura do solo. 3.Modelagem de dados – Análise. I.Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias, Ambientais e Biológicas. II.Título.

CDD: 631.4

Ficha elaborada pela Biblioteca Universitária de Cruz das Almas - UFRB. Responsável pela Elaboração – Antonio Marcos Sarmento das Chagas (Bibliotecário - CRB5 / 1615).

UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA

CENTRO DE CIÊNCIAS AGRÁRIAS AMBIENTAIS E BIOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM SOLOS E QUALIDADE DE ECOSSISTEMAS

APRENDIZADO DE MÁQUINA E FLUORESCÊNCIA DE RAIO-X NO MAPEAMENTO DIGITAL DE ATRIBUTOS QUÍMICOS DOS SOLOS EM ÁREA DE COBERTURA PEDOLÓGICA COMPLEXA

COMISSÃO EXAMINADORA DA DEFESA DE DISSERTAÇÃO DE ÍCARO BARRETO SOUZA

Dr. Francisco Alisson Da Silva Xavier

Embrapa Mandioca e Fruticultura

Prof. Dr. Everton Luís Poelking

Universidade Federal do Recôncavo da Bahia

Prof. Dr. Elpídio Inácio Fernandes Filho

Universidade Federal de Viçosa

Dissertação homologada pelo Colegiado do Curso de Mestrado em Solos e Qualidade de Ecossistemas em _____, conferindo o Grau de Mestre em Solos e Qualidade de Ecossistemas em_____.

DEDICO

A minha filha Vivá de Souza Barreto,

minha mãe Lídia Maria Barreto Souza,

meu pai Fábio Andrade Souza,

minha irmã Fabiane Barreto Souza,

à toda minha família, amigas e amigos...

AGRADECIMENTOS

Agradeço à minha família... minha Filha, minha Mãe, meu Pai (in memorian), Minha irmã, minha Avó, minha tia Meire, minhas tias Marluce e Lêda, meus primos Luís Carlos, Antônio Alfredo e Wesley por todo suporte, companheirismo, exemplos, incentivo e por serem a parte mais importante da minha vida.

Agradeço ao Professor Oldair Del'Arco Vinhas Costa, pela orientação, dedicação, compreensão, amizade, mentoria, transmissão de conhecimento, exemplo profissional e pessoal.

Agradeço ao Professor Luciano Silva Sousa, pelo acolhimento, palavras de incentivo, acompanhamento, orientação, exemplo profissional e pessoal.

Agradeço ao Professor Everton Luís Poelking, por todas as palavras de incentivo, acolhimento, amizade, mentoria, transmissão de conhecimento e exemplo profissional.

Agradeço ao Professor Thomas Vincent Gloaguen, a cooperação, a cessão dos dados que possibilitaram o desenvolvimento desse trabalho, a orientação profissional e dedicação a essa pesquisa.

Agradeço ao Professor Elpídio Inácio Fernandes Filho, pelo exemplo profissional, dedicação, pela oportunidade de conhecê-lo, transmissão de conhecimento, atenção, por ser um Professor no sentido mais amplo da palavra.

Agradeço ao colega, Engenheiro Felipe Torres Sampaio, pela amizade, companheirismo, pelas palavras de incentivo, pela dedicação, por tê-lo conhecido, meus agradecimentos à um grande amigo.

Agradeço ao Professor Júlio César de Azevedo Nóbrega, pelo incentivo, pelas palavras, pela orientação e exemplo profissional, pelo acolhimento e atenção.

Agradeço ao Professor José Maria Lima, apesar do breve contato, suas palavras de incentivo fizeram grande diferença em minha jornada profissional, busco exemplar-me em sua postura como pessoa incentivadora.

Agradeço às minhas colegas e meus colegas da CMP, por toda amizade verdadeira, pelo clima de trabalho extremamente saudável, é indescritível fazer parte dessa família. Meus agradecimentos a Jarbas (em especial), Deivisson, Lúcio, Clóvis, Carlos Roberto, Wilson, Alfredo, Bartolomeu, José Roberto, Marcelo, Jocélia, Manuella, Niúra, Nádia, Josy, Luan, Maurício... Sucesso em vossas jornadas sempre!!!

Agradeço aos meus amigos e amigas que sempre estiveram próximos, em especial ao amigo Osvaldo da Paz (Vadinho), um grande parceiro!!

APRENDIZADO DE MÁQUINA E FLUORESCÊNCIA DE RAIO-X NO MAPEAMENTO DIGITAL DE ATRIBUTOS QUÍMICOS DOS SOLOS EM ÁREA DE COBERTURA PEDOLÓGICA COMPLEXA

RESUMO

A associação do mapeamento digital de solos com a análise via sensores próximos apresenta-se como uma poderosa ferramenta para geração de modelos espaciais de atributos dos solos com alta resolução, elevada acurácia e baixo custo. Informações sobre a distribuição espacial de determinados elementos químicos presentes nos solos ao longo da paisagem, relacionados à composição geoquímica do material de origem e ao seu grau de intemperismo, podem potencializar os resultados de estudos pedológicos em áreas com configurações ambientais complexas, possibilitando a elaboração de mapas de solos com alto nível de detalhe. O objetivo desse trabalho foi mapear elementos químicos do solo relacionados à geoquímica e intemperismo dos diferentes materiais de origem, comparando a performance de sete algoritmos de machine learning na predição espacial e, estabelecer relações entre a variação espacial dos elementos mapeados e a distribuição de classes de solos. Para explicar a variação espacial desses elementos foi utilizado um conjunto de 42 covariáveis, obtidas a partir do modelo digital de elevação, de uma coleção de imagens do Landsat 8, além de variáveis de localização espacial. As variáveis alvo foram os teores total de Mg, Al, Si, K, Ca, Ti e Fe, representadas por um total de 546 amostras coletadas em 182 pontos aleatórios nas camadas de 0-5 cm, 5-20 cm e 20-40 cm e analisadas em aparelho de fluorescência de raio-x. A performance dos algoritmos foi avaliada através dos valores de R^2, MAE e RMSE.RF, GBM, SVMr, CUB e MARS foram os algoritmos que obtiveram os melhores ajustes, com performances muito próximas na maioria dos elementos. O RF destaca-se dos demais algoritmos por apresentar ajustes superiores com maior frequência. ANN, e principalmente o kNN apresentaram ajustes significativamente inferiores aos outros algoritmos para a maioria dos elementos estudados. Os elementos com melhor ajuste foram K e Mg, seguidos por Al, Fe e Ti. Os ajustes obtidos para Si e Ca foram inferiores aos apresentados pelos outros elementos. A análise dos mapas gerados permite afirmar que a distribuição espacial de Ca, Mg, K e Fe têm forte correlação positiva, e os resultado da análise PCA indicam uma relação significativa entre a distribuição espacial desses elementos e a ocorrência de Vertissolos. Ti e Al também apresentaram correlação positiva forte entre si, e, de acordo com os resultados da análise PCA a distribuição desses elementos está relacionada à ocorrência de Latossolos Amarelo e Argissolos Amarelos. Os mapas de Si apresentaram correlação negativa de média à forte com a maioria dos elementos, com os resultados obtidos na PCA indicando forte relação entre a distribuição espacial do elemento e a ocorrência de Neossolos Quartzarênicos. A integração entre as técnicas de machine learning e a análise de fluorescência de raio-X foi eficaz na predição espacial de atributos químicos do solo, com os modelos gerados apresentando valores de R^2 que variam entre 0.22 e 0.83 para os algoritmos com melhor desempenho nas variáveis estudadas. Os mapas gerados representam bem a variação espacial do fenômeno estudado, captando detalhes importantes para explicar a distribuição dos solos em uma paisagem que apresenta coberturas litológica e pedológica com alto nível de complexidade.

Palavras-Chave: Aprendizado de máquina, mapeamento digital de solos, fluorescência de raio-x

MACHINE LEARNING AND X-RAY FLUORESCENCE IN THE DIGITAL MAPPING OF CHEMICAL ATTRIBUTES OF SOILS IN A COMPLEX PEDOLOGICAL COVERAGE AREA

ABSTRACT

The association of digital soil mapping with analysis via proximal sensors presents itself as a powerful tool for generating spatial models of soil attributes with high resolution, high accuracy, and low cost. Information on the spatial distribution of certain chemical elements present in soils throughout the landscape, related to the geochemical composition of the source material and its degree of weathering, can enhance the results of pedological studies in areas with complex environmental configurations, enabling the elaboration of soil maps with high level of detail. The objective of this work was to map soil chemical elements related to geochemistry and weathering of different source materials, comparing the performance of seven machine learning algorithms in spatial prediction and to establish relationships between the spatial variation of the mapped elements and the distribution of classes of soils. To explain the spatial variation of these elements, a set of 42 covariates was used, obtained from the digital elevation model, from a collection of Landsat 8 images, in addition to spatial location variables. The target variables were the total Mg, Al, Si, K, Ca, Ti and Fe contents, represented by a total of 546 samples collected at 182 random points in the layers of 0-5 cm, 5-20 cm and 20-40 cm and analyzed in a fluorescence x-ray machine. The performance of the algorithms was evaluated through the values of R^2, MAE and RMSE.RF, GBM, SVMr, CUB and MARS were the algorithms that obtained the best adjustments, with very similar performances in most elements. The RF stands out from the other algorithms for presenting higher adjustments more frequently. ANN, and especially kNN, showed significantly lower adjustments than the other algorithms for most of the elements studied. The elements with the best fit were K and Mg, followed by Al, Fe and Ti. The adjustments obtained for Si and Ca were lower than those presented for the other elements. The analysis of the generated maps allows us to affirm that the spatial distribution of Ca, Mg, K and Fe have a strong positive correlation, and the results of the PCA analysis indicate a significant relationship between the spatial distribution of these elements and the occurrence of Vertisols. Ti and Al also showed a strong positive correlation with each other, and, according to the results of the PCA analysis, the distribution of these elements is related to the occurrence of Latosols and Ultisols. The Si maps showed a medium to strong negative correlation with most elements, with the results obtained in PCA indicating a strong relationship between the spatial distribution of the element and the occurrence of Neosols. The integration between machine learning techniques and X-ray fluorescence analysis was effective in the spatial prediction of soil chemical attributes, with the generated models presenting R^2 values ranging between 0.22 and 0.83 for the algorithms with the best performance in the variables studied. The generated maps represent well the spatial variation of the phenomenon studied, capturing important details to explain the distribution of soils in a landscape that presents lithological and pedological covers with a high level of complexity

Keywords: Machine learning, digital soil mapping, x-ray fluorescence

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Os mapas de solos são representações gráficas para transmissão de informação sobre a distribuição espacial de propriedades dos solos (YAALON, 1989), consistindo em ferramentas fundamentais no planejamento do uso da terra, porém os métodos de levantamento convencionais são onerosos e, os mapas já existentes possuem baixo nível de detalhe (COELHO et al., 2020). O mapeamento de atributos do solo em nível de detalhe adequado para uso eficaz requer uma grande quantidade de amostras (VASQUES et al., 2020) e, as análises de fertilidade convencionais têm execução complexa, alto custo financeiro, exigem tempo e geram resíduos contaminantes (SILVA et al., 2017; BENEDET et al., 2021).

Utilizar sensores próximos para quantificar atributos do solo reduz o tempo de análise, os custos financeiros e a geração de resíduos químicos, permitindo aumentar o esquema amostral e, consequentemente, o nível de detalhe nos resultados (RIBEIRO et al., 2017). A fluorescência de raio-x é uma técnica capaz de identificar e quantificar os elementos químicos que compõem materiais sólidos (WEINDORF et al., 2014), utilizada com sucesso em estudos na ciência do solo para determinação completa da composição elementar e estimativa de elementos químicos totais e atributos físicos do solo (RIBEIRO et al., 2017), caracteriza-se por ser uma análise de escaneamento *in-situ*, com rápida obtenção de resultados (KALNICKY & SINGHVI, 2001).

A análise XRF apresentou resultados promissores na predição e mapeamento de Ca^{2+}, Mg^{2+} e Al^{3+} trocáveis e P remanescente (SILVA et al., 2017; BENEDET et al., 2021); matéria orgânica, CTC efetiva, saturação por bases e pH (SILVA et al., 2017); Ca^{2+}, K^{+} e P (PELEGRINO et al., 2021) e elementos totais do solo (Al, Ca, Cr, Fe, K, P, Pb, S, Ti, V e Zr) (CAMPBELL et al., 2019). Comparado a outros sensores próximos, o pXRF apresentou melhor desempenho na predição e mapeamento de atributos químicos e físicos do solo, com performance superior na análise da densidade de partículas, teores de argila e matéria orgânica (VASQUES et al., 2020).

O mapeamento digital de solos utiliza técnicas computacionais para ajustar diferentes modelos (lineares generalizados, árvores de classificação e regressão, redes neurais, logica *fuzzy* e geoestatística), buscando encontrar padrões espaciais a partir de relações quantitativas entre variáveis ambientais e espaciais (MCBRATNEY et al., 2003; LI et al., 2020). O modelo SCORPAN (S - solo, C - clima, O - organismos, R - relevo, P - material de origem, A - tempo e N - posição espacial) baseia-se na equação proposta por Jenny (1941), utilizando os fatores de formação dos solos como variáveis independentes no processo de predição espacial de classes e atributos, possuindo capacidade de operar em configurações ambientais com alto nível complexidade (McBratney et al., 2003).

Técnicas de machine learning vêm sendo aplicadas de forma crescente no mapeamento digital de solos, entre os algoritmos mais utilizados estão as regressões lineares múltiplas (RLM), k-nearest neighbors (KNN), support vector regression (SVR), Cubist (CUB), random forest (RF) e redes neurais artificiais (ANN), desse modo, selecionar o algoritmo que melhor se ajusta ao conjunto de dados estudados é essencial para otimizar os resultados (KHALEDIAN et al., 2020). Em um trabalho sobre mapeamento da fertilidade do solo houve variação nos mapas resultantes de algoritmos diferentes, evidenciando a necessidade de testar múltiplos modelos preditores, selecionando aquele com melhor ajuste para uso em aplicações práticas (PELEGRINO et al., 2021). Apesar de testados em diferentes configurações ambientais, é incomum a avaliação de múltiplos modelos em uma mesma paisagem (BRUNGARD et al., 2015). A avaliação da performance de diferentes algoritmos de machine learning no mapeamento atributos químicos e físicos do solo é abordada nos

estudos de Beguin et al. (2017) RF, BRT, KKNN, Cubist, GAM, GLM; Jeong et al. (2017) GAM, RF, SVR; Campbell et al. (2019) RF, RIDGE, Cubist, PLS, PCR, FOBA, GBM, GLMBOOST; Zhang and Shi (2019) KNN, MLP RF, SVM, XGB; Zhou et al. (2019) BRT, SVM e RF; Keskin et al. (2019), CaRT, BaRT, BoRT, RF, SVM, PLSR; Zeraatpisheh et al. (2019) Cubist, RF, RT, MLR e Mendes et al. (2020) SVM, RF, Cubist.

Pesquisas no âmbito do mapeamento digital de solos vêm revelando informações úteis acerca dos processos pedogenéticos (MA et al., 2019), nesse contexto, houve incremento no uso do pXRF para fins pedológicos devido a facilidade para aquisição de dados (WEINDORF et al., 2014), e por ser uma ferramenta para análise rápida da geoquímica do solo (HSEU ET AL., 2016). A integração ente técnicas de análise espacial e análise pXRF para avaliação da variabilidade de elementos e propriedades do solo têm se mostrado eficaz (YANG et al., 2020), possibilitando aprofundar o entendimento da influência dos fatores de formação na pedogênese (TIGHE et al., 2018). Valores da concentração de Fe determinado via pXRF foram utilizados para identificar e mapear ocorrência de horizontes B com acumulação de argila, o índice de Ti e Zr totais refletiu o histórico de intemperismo do solo e Ca total indicou a ocorrência de solos originados de materiais calcáreos em área de produção vinícola (JANG et al., 2016). Teores de Si, K, Al Fe e Ti obtidos via pXRF foram utilizados para estudar processos litogênicos e pedológicos sítios arqueológicos, nesse mesmo trabalho teores totais de Ca sinalizaram a ocorrência de atividades antrópicas no solo (ARNOLDUSSEN & OS, 2014). A integração entre a análise de componentes principais e a krigagem bayesiana empírica permitiu estabelecer relações quantitativas entre a variabilidade espacial de elementos quantificados via pXRF e processos pedogenéticos em solos de golfo nos EUA (YANG et al., 2020)

Nesse sentido, a distribuição de determinados elementos químicos presentes nos solos ao longo da paisagem, que está relacionada ao grau de intemperismo e aos processos de perda por erosão e lixiviação sofridos, pode auxiliar nos estudos pedológicos em áreas com configurações ambientais complexas e, consequentemente, na elaboração de mapas de solos mais detalhados. Considerando a variedade de modelos preditores utilizados no mapeamento digital de solos, o uso potencial do sensoriamento próximo via fluorescência de raio-x como alternativa às análises laboratoriais convencionais em levantamento de solos e, a aplicação de mapas de atributos para aprofundar o entendimento dos processos pedogenéticos e da distribuição espacial pedológica, o presente trabalho objetivou mapear elementos químicos totais do solo (Mg, Al, Si, K, Ca, Ti e Fe) em área de bacia sedimentar tropical, utilizando análise XRF e técnicas de machine learning, avaliando sete algoritmos com operações diferentes na predição espacial; random forest (RF), gradient boost machine (GBM), suport vector machine radial (SVMr), Cubist (CUB), Earth (MARS), redes neurais artificiais (ANN) e K-nearest neighbors (KNN), identificando correlações entre os mapas de elementos totais gerados com processos pedogenéticos e distribuição espacial pedológica ocorrente na área.

2. REFERENCIAL TEÓRICO

2.1 Mapeamento digital de solos

O avanço nas tecnologias computacionais permitiu a exploração e manuseio de extensas bases de dados, utilizando ferramentas de mineração de dados e aprendizado de máquinas. Na ciência dos solos essas técnicas integram-se aos sistemas de informações geográficas (SIG), sistema de posicionamento global (GPS) e aos sensores remotos e próximos, para gerar informações espaciais com alto grau de complexidade, iniciando uma nova vertente para as pesquisas, principalmente no mapeamento de classes e atributos dos solos (MC BRATNEY et al., 2003).

O mapeamento digital de solos consiste na criação e alimentação de sistemas computacionais de informações espaciais dos solos, utilizando métodos numéricos para modelar as variações espaciais e temporais de classes e atributos dos solos, estabelecendo relações entre pontos de observação, conhecimento especialista e covariáveis ambientais (LAGACHERIE & MCBRATNEY, 2006). Geralmente há duas abordagens no mapeamento digital de solos, uma onde o mapeamento é automático, objetivo e quantitativo, faz uso de técnicas estatísticas, geoestatísticas, mineração de dados e aprendizado de máquina, e normalmente utiliza uma densa rede de pontos amostrais. A outra abordagem, mais próxima do levantamento de solos convencional, utiliza o conhecimento de campo do cientista de solos para ajustar os mapas, apoiando-se no uso de técnicas de engenharia para aquisição e representação do conhecimento, e inferência baseada em conhecimento especialista, normalmente desenvolvidas através de operadores *fuzzy* (SHI et al., 2009).

McBratney et al., 2003 apresentaram em seu artigo uma estrutura genérica para o mapeamento digital de solo, denominada SCORPAN-SSPFe (função de predição espacial de solos com erros espaciais auto correlacionados), cuja variável alvo é a classe ou atributo de solo a ser predito, e as variáveis explicativas são os fatores de formação do solo, propriedades do próprio solo e a posição espacial, que foram sintetizadas na seguinte equação:

$$S_c = f(S.C.O.R.P.A.N) \; ou \; S_a = f(S.C.O.R.P.A.N) \quad (1)$$

Onde S_c são classes de solo, S_a são atributos do solo, *S* são dados de solo (químicos, físicos, morfológicos, de sensoriamento remoto ou próximo), *C* são dados do clima, *O* são dados de organismos, *R* são dados de relevo, *P* são dados do material de origem, *A* é o fator tempo e *N* é a posição espacial.

O fluxo de trabalho do mapeamento digital consiste basicamente em combinar pilhas de rasters ambientais e espaciais com dados amostrais de solos obtidos via análises de laboratório, aplicando modelos empíricos para traduzir essas informações em um mapa digital de solos (PIKKI et al., 2021). Na ciência dos solos, o mapeamento digital pode ser o ponto de partida para formulação de novas hipóteses (WADOUX & MCBRATNEY, 2021).

A partir do final da década de 60 houve incremento no número de pesquisas para mapeamento digital de atributos do solo, utilizando abordagens “estritamente espaciais”, onde a predição leva em conta apenas a posição geográfica (CAMPBELL et al., 2019). O esquema amostral nessa abordagem deve ser denso o suficiente para garantir que o modelo encontre dependência espacial entre os pontos, caso contrário não é possível realizar predições (MENDES et al., 2020). Nos métodos estritamente espaciais os resultados são gerados por interpolação entre pontos de observação, resultando em modelos de superfície bastante simplificada, com representações artificiais dos fenômenos estudados (MC BRATNEY et al., 2003), não expressando na maioria das vezes a complexidade inerente das feições naturais.

Nas duas últimas décadas houve um aumento significativo em pesquisas que utilizam as técnicas e conceitos do mapeamento digital para predição de atributos do solo (CHEN et al.,2022). Algoritmos de machine learning realizam a predição espacial de propriedades do solo a partir de um conjunto de variáveis explicativas, não havendo necessidade de dependência espacial entre os pontos para obtenção de resultados (MENDES et al., 2020).

Hartemink et al. (2013), cita uma discussão de Nefedov (1908) na Rússia, onde o cientista afirma que mapas de atributos do solo devem ser produzidos antes dos mapas de classes, para que as classes de solo sejam delineadas a partir de regiões que apresentem semelhança na configuração dos atributos. O *GobalSoilMap* é um projeto de nível global que surgiu no âmbito do 2º Workshop Global de Mapeamento

Digital de Solos, ocorrido em 2006 no Rio de Janeiro. O projeto tem por objetivo coordenar a produção de mapas de atributos do solo com alta resolução por toda superfície do globo terrestre. Nesse projeto serão mapeados doze atributos do solo relacionados à morfologia, química e física do solo (Chen et al., 2022).

2.2 Fluorescência de raio-X

A espectrometria por fluorescência de raio-X é uma técnica analítica para quantificação rápida de elementos químicos, uma das principais aplicações é a determinação de metais e solos em sedimentos. Essa é uma técnica não destrutiva que permite análises quantitativas e qualitativas de múltiplos elementos químicos na matriz amostral, com baixo custo, alta acurácia e rapidez na obtenção de resultados (WEINDORF et al., 2001).

Um raio-X que incide em átomo tem a capacidade de excitá-lo, forçando a transição de elétrons de uma camada mais externa para preencher o espaço criado na camada mais interna. A transição de elétrons emite fótons com energia na região do espectro eletromagnético do raio-X, que equivale a diferença de energia entre as duas camadas. Os diferentes elementos químicos possuem regiões específicas do espectro eletromagnético de raio-X onde a transição ocorre, por isso a detecção da energia desses fótons fluorescentes permite identificar o tipo de espectro e quantificar o elemento presente na amostra. O uso do XRF para quantificação de elementos requer a calibração do aparelho, usando padrões com composição química conhecida. O procedimento de calibração compara a intensidade do raio-X para elementos alvo com concentração conhecida, e então é construído um modelo de quantificação adequado para analisar o tipo de matriz amostral do padrão utilizado para calibração. O limite de detecção para análises de quantificação do aparelho XRF é a medida da menor concentração do analito que é detectável pelo equipamento com certo grau de confiança. Aumentar o tempo da análise melhora a performance do aparelho e diminui o limite de detecção(KALNICKY & SINGHVI, 2001).

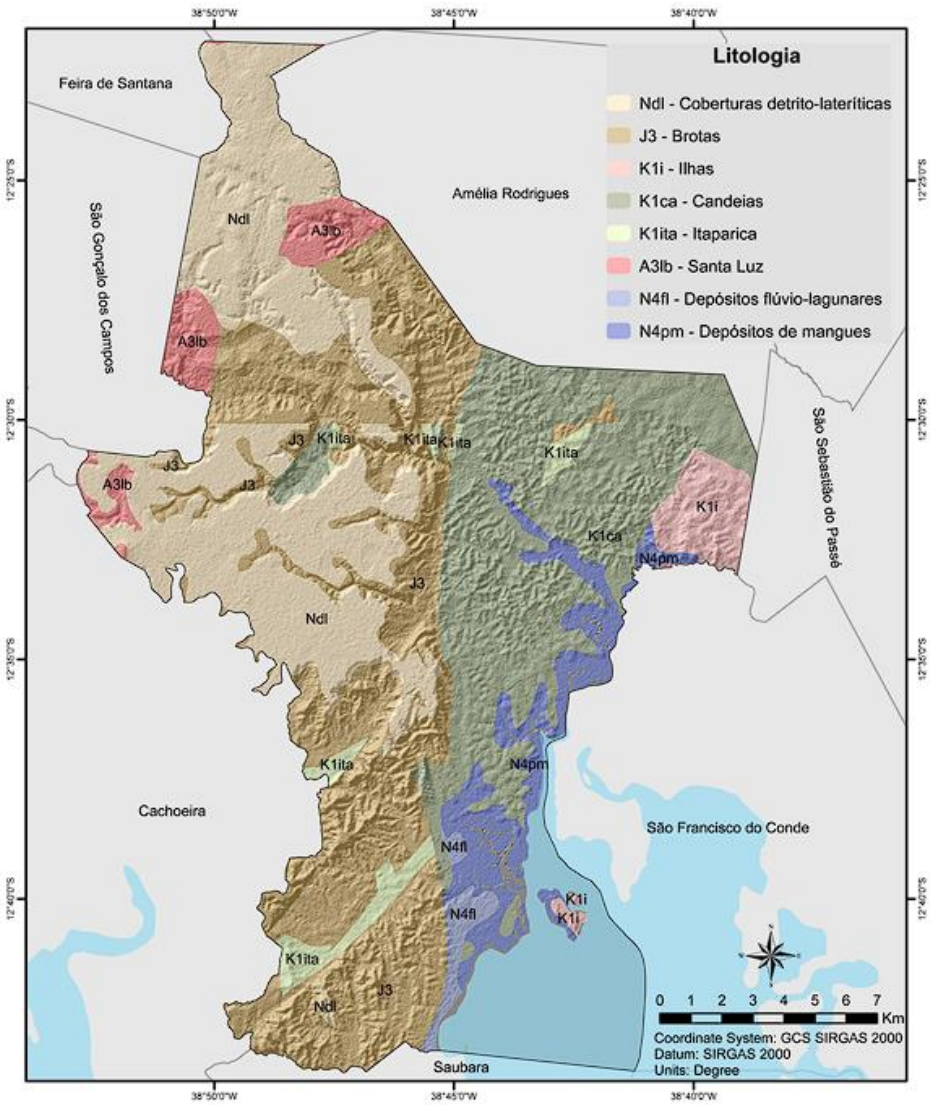
A análise por espectrometria de fluorescência de raio-X vem sendo utilizada com sucesso em trabalho e pesquisas nas áreas ambientais, agronômicas e pedológica, inclusive com resultados comparáveis a análises de laboratório convencionais. O método 6200 da Agência Americana de Proteção Ambiental dos Estados Unidos (US-EPA) reconhece o uso de aparelhos portáteis de XRF na determinação da concentração de elementos químicos em solos e sedimentos. Nos próximos 30 anos está previsto no Brasil a execução do PRONASSOLOS (Programa Nacional de Solos), cujo objetivo é mapear os solos do território brasileiro em nível de detalhe. Nesse sentido a análise de solos via XRF pode contribuir para a execução de análises potencialmente acuradas, sustentáveis, de baixo custo e com execução em campo, com fornecimento rápido de informações sobre atributos dos solos (RIBEIRO et al., 2017).

2.3 Litologia em Santo Amaro/BA

2.3.1 Litologia

A litologia em Santo Amaro (Figura 1) é marcada pela presença de rochas sedimentares (Grupos Santo Amaro, Brotas, Ilhas), rochas metamórficas (Complexo Santa Luz), sedimentos inconsolidados (Coberturas Detrito Lateríticas e Formação Barreiras) e depósitos sedimentares (Depósitos Flúvio-Lagunares, de Leques Aluviais e de Mangues) (CPRM, 2003,2008).

Figura 1. Mapa litológico do município de Santo Amaro/BA. Adaptado de CPRM (2003, 2008).



2.3.2 Coberturas detrito lateríticas

As Coberturas Detrito Lateríticas são formações terrígenas, originadas por sedimentação detrito-eluvionar, compostas por material arenoso e às vezes lateríticos, não consolidados. Sua formação ocorre em áreas planas, tabuleiros ou pediplanos, quando próximas ao litoral se confundem com os sedimentos Barreiras, com os quais

tem estreita correlação. A característica litológica é a presença de; areias, argilas, cascalhos e canga (SANTOS, 2015; BARBOSA; DOMINGUEZ,1996).

2.3.3 Grupo Brotas

O Grupo Brotas é a unidade inferior do Supergrupo Bahia, divide-se nas formações; Aliança localizada na base, e Sergi na parte superior (BRASIL, 1981). A Formação Aliança é composta por dois membros. O membro Boipeba tem litologia composta por arenitos com mineralogia variada, o mais representativo é um arcóseo de feldspato branco, sendo encontrados ainda quartzo-arenitos grosseiros e litoarenitos micáceos, predominam as cores marrom e vermelho, mas, encontram-se ainda arenitos cinzas que se assemelham aos da formação Sergi. O membro Capianga caracteriza-se por apresentar litologia composta predominantemente por folhelhos de cor vermelho-tijolo, lamitos e siltitos, e, na parte superior ocorrem arenitos variegados (BRASIL,1981). Os sedimentos que originam as rochas da Formação Aliança foram depositados por sistemas flúvio-lacustres sob clima árido. (CAIXETA et al. 1994). A formação Sergi é marcada pela presença de arenitos quartzosos cinza esverdeados e vermelhos, com diversas granulometrias e diferentes maturidades texturais. Ocorrem variações laterais por interdigitações e interestratificações de folhelhos, e siltitos vermelhos e verdes. Na parte superficial dessa formação encontra-se conglomerados finos a médios, e arenitos com seixos esparsos, representados principalmente pelos fenoclastos de sílex. Na base é comum encontrar bolas de argila em meio à massa de quartzo-arenitos mal selecionados. A Formação Sergi é a unidade superior do Grupo Brotas, e apresenta maior distribuição espacial na região do Recôncavo (BRASIL,1981). A deposição do material sedimentar que originou a Formação Sergi ocorreu por sistemas fluviais entrelaçados, com posterior retrabalhamento eólico. (CAIXETA et al. 1994).

2.3.4 Grupo Santo Amaro

O Grupo Santo Amaro subdivide-se nas formações: Itaparica, Água Grande, Candeias e Maracangalha (CAIXETA et al. 1994). De acordo com o Mapa Geológico da Região Metropolitana de Salvador publicado por (CPRM,2008) ocorrem em Santo Amaro as Formações Itaparica e Candeias. A litologia da Formação Itaparica caracteriza-se por dividir-se em seções que intercalam folhelhos e arenitos. Na parte inferior ocorrem folhelhos verdes com lâminas internas de calcita fibrosa; no meio encontra-se um arenito quartzoso fino a médio, de pequena espessura e bem selecionado, com seixos dispersos e estratificação cruzada; sobrepondo o arenito há outra camada de folhelho que se assemelha a anterior, a diferença fica por conta da existência de leitos de carbonatos impuros de cor creme, siltitos castanhos e vermelhos escuros, e bancos de folhelhos argilosos. A ocorrência desses folhelhos e siltitos torna possível identificar a presença da Formação Itaparica no Recôncavo, já que a Formação Candeias não possui na região rochas com as características supracitadas. É comum aos arenitos dessa formação, a ocorrência de conglomerados finos intraformacionais, e litoclastos dispersos de quartzo, sílex e argila (BRASIL,1981). A formação Candeias divide-se em dois membros: Tauá, que se localiza na base; Gomo, encontrado no estrato superior. Identifica-se a litologia do Membro Tauá pela presença de folhelhos, lamitos e siltitos finos, de cor cinzaescuro e preto; arenitos muitos finos calcíferos; e um leito carbonático escuro castanho situado na parte superior. O Membro Gomo é marcado pela presença de lamitos e siltitos calcíferos cinza-escuros e pretos; arenitos micáceos cinza-claros, distribuídos em leitos decimétricos, calcificados de forma intensa e irregular, comumente carbonatos arenosos. O estrato superficial, submetido ao intemperismo, é formado por folhelhos de cor cinza-esverdeado, laminados, com finas camadas de arenito cinza-claro. O carbonato de cálcio presente nas argilas encontra-se sob a forma de nódulos

espáticos (BRASIL,1981; CAIXETA et al. 1994). A deposição dos sedimentos que originaram as rochas da Formação Candeias ocorreu em ambiente lacustre, com rápida subsidência e aporte sedimentar intenso (CAIXETA et al. 1994).

2.3.5 Complexo Santa Luz

O Complexo Santa Luz é um domínio tectônico de idade mesoarqueana, composto por rochas com médio a alto grau de metamorfismo, sendo as principais: migmatitos, ortognaisses, granulitos, anfibolitos, e quartzo dioritos. Os afloramentos rochosos apresentam cores acinzentadas e avermelhadas, a composição mineralógica é caracterizada pela presença de quartzo, feldspatos, biotita e piroxênio. São observados a presença de veios de quartzo e pegmatitos, sendo estes derivados de processos finais de magmatismo. As rochas do Complexo Santa Luz sofreram intensos processos de metamorfismo, deformação e hidrotermalismo, resultando na alteração dos minerais por meio da saussuritização, na metamorfização em grau médio de fácies anfibolito e na obliteração do bandamento composicional de minerais félsicos e máficos. Devido a este último processo as rochas apresentam foliação gnáissica. Em alguns afloramentos, a foliação gnáissica já obliterada resulta em rochas com maior grau de homogeneidade (ROCHA et al. 2017).

2.3.6 Grupo Ilhas

Fazem parte do Grupo Ilhas as Formações; Pojuca, superior, e Marfim, inferior. A Formação Pojuca apresenta em sua composição siltitos cinza-claros, folhelhos cinza-esverdeados, calcários impuros castanhos e arenitos quartzosos cinzas muito finos a médios. É a formação superior do Grupo Ilhas. A Formação Marfim é caracterizada por apresentar como estrato superior uma faixa de calcário arenoso, as partes abaixo se dividem em Camadas Caruaçu e Membro Catu. A Camada Caruaçú são compostas por arenitos cinza-verde a amarelos, com corpos lenticulares, intercaladas por siltitos e folhelhos cinza-verde, calcíferos. O Membro Catu é formado de arenitos quartzosos cinzas, finos a médios, intercalados por siltitos e folhelhos (BRASIL,1981).

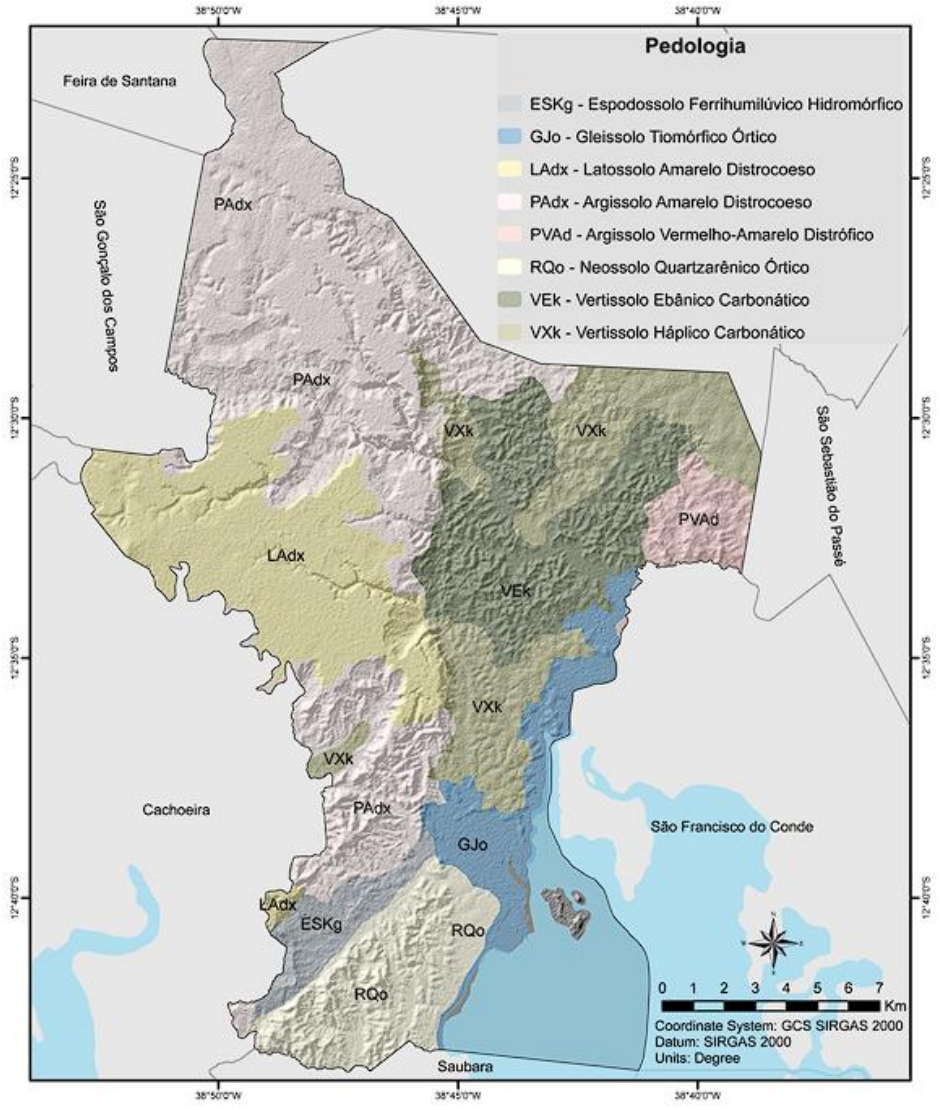
2.3.7 Sedimentos do Quaternário

Os depósitos de leques aluviais são encontrados no sopé de elevações ou distribuídos em encostas de vales, formando corpos irregulares e descontínuos. As partes mais altas desses depósitos variam de 15 a 20 metros acima do nível do mar. A mineralogia desse material é diversa, e depende da rocha matriz. As rochas do embasamento originam depósitos arcosianos, com quartzos angulares e feldspato alterado. O material rochoso do Mesozóico dá origem a areias quartzosas, com argilas, cascalhos de arenito e folhelhos. Quando os depósitos são originados a partir de material oriundo da Formação Barreiras, sua composição é baseada em areias quartzosas com frações variáveis de argila e quartzo arredondado (VILAS BOAS et al. 1985). Nos depósitos de mangue predominam materiais argilo-siltosos ricos em matéria orgânica. Distribuem-se espacialmente ao longo de margens de rios e riachos sujeitos a influência das marés, e no entorno das baías. Os sedimentos são indiferenciados e possuem pouco ou nenhum grau de consolidação (BARBOSA; DOMINGUEZ, 1996).

2.4 Pedologia em Santo Amaro/BA

De acordo com o arquivo vetorial pedológico do IBGE (2018) (Figura 2) oito classes de solos diferentes predominam em Santo Amaro/BA: LATOSSOLOS AMARELOS Distrocoesos, ARGISSOLOS AMARELOS Distrocoesos, ARGISSOLOS VERMELHO-AMARELO Distróficos, VERTISSOLOS EBÂNICOS Carbonáticos, VERTISSOLOS HÁPLICOS Carbonáticos, NEOSSOLOS QUARTZARÊNICOS Órticos, GLEISSOLOS TIOMÓRFICOS Órticos e ESPODOSSOLOS FERRI-HUMILÚVICO Hidromórficos.

Figura 2. Mapa pedológico do município de Santo Amaro/BA (IBGE,



2.4.1 Latossolos Amarelos Distrocoesos

Latossolos são solos altamente intemperizados, evoluídos e profundos, típicos de regiões equatoriais e tropicais, ocorrendo em área de relevo plano e suave ondulado. Geralmente há ausência de minerais primários, ou secundários de baixa resistência ao intemperismo na massa do solo, e os valores de Ki variam entre altos, nos solos de mineralogia caulinítica, a extremamente baixos em solos oxídicos.

Apresentam baixa CTC, baixa saturação por bases, são fortemente ácidos, distróficos ou alumínicos. Latossolos Amarelos Distrocoesos apresentam horizonte B latossólico imediatamente abaixo do A, matiz 7,5YR ou mais amarelo e saturação de bases menor que 50% na maior parte dos primeiros 100cm do horizonte B, e caráter coeso dentro de 150cm da superfície do solo (EMBRAPA, 2018).

A coesão é uma característica pedogenética que ocorre nos horizontes subsuperficiais de textura média, argilosa ou muito argilosa, geralmente entre 0,30 e 0,70m de profundidade, comum em solos formados a partir de sedimentos terciários da formação Barreiras, ou formações correlatas. A fração argila desses solos tem mineralogia caulinítica, e o predomínio de goethita entre os óxidos confere cores amareladas típicas. Uma característica comum nesses solos é o Ki elevado, normalmente com valores entre 1,7 e 2,0 (NETO et al., 2009).

1.4.2 Argissolos Amarelos Distrocoesos e Argissolos Vermelho-Amarelos Distróficos

Argissolos diferenciam-se por apresentar horizonte B com acumulação de argila de atividade baixa, ou atividade alta desde que associada a baixa saturação por bases ou caráter alumínico. A textura varia de arenosa a argilosa no horizonte A e de média a muito argilosa no horizonte Bt, sempre com aumento no teor de argila do A para o B. A acidez varia de forte a moderada, saturação por bases pode ser baixa ou alta, a mineralogia predominantemente caulinítica e os valores de Ki variam no intervalo de 1,0 a 3,3. Argissolos Amarelos Distrocoesos apresentam horizonte B textural imediatamente abaixo do A ou E, matiz 7,5YR ou mais amarelo e saturação de bases menor que 50% na maior parte dos primeiros 100cm do horizonte B, e caráter coeso dentro de 150cm da superfície do solo. Argissolos Vermelho-Amarelos Distróficos apresentam cores vermelho-amareladas que não se enquadram nas outras subordens, e saturação por bases menor que 50% na maior parte dos primeiros 100cm do horizonte B (EMBRAPA, 2018).

1.4.3 Vertissolos Ebânicos Carbonáticos e Vertissolos Háplicos Carbonáticos

A classe dos Vertissolos compreende solos minerais com horizonte vértico e variação textural ao longo do perfil insuficiente para caracterizar um horizonte B textural. O aumento do teor de água no solo causa mudanças consideráveis no volume, com movimentação da massa do solo que ocasiona contração e fendilhamento em períodos seco, e expansão quando úmido. A movimentação do solo é evidenciada pela formação das superfícies de fricção e do microrelevo gilgai. Vertissolos apresentam alta CTC, saturação de bases superior a 50%, teores elevados de Ca e Mg, relação Ki maior que 2,0, e pH frequentemente neutro ou alcalino (EMBRAPA, 2018).

Os VERTISSOLOS EBÂNICOS Carbonáticos devem apresentar caráter ebânico na maior parte dos horizontes B e/ou C, e caráter carbonático ou camadas com horizonte cálcico, dentro de 100cm a partir da superfície. Os VERTISSOLOS HÁPLICOS Carbonáticos devem apresentar caráter carbonático ou camadas com horizonte cálcico, dentro de 100cm a partir da superfície, e não se enquadrar nas subordens hidromórficos e ebânicos. O caráter ebânico diz respeito à dominância de cores escuras, quase pretas, na maior parte do horizonte diagnóstico. O caráter carbonático refere-se à ocorrência de $CaCO_3$ em concentração superior à 150 g/kg de solo, em forma que não satisfaça requisitos para caracterizar um horizonte cálcico (EMBRAPA, 2018).

1.4.4 Neossolos Quartzarênicos Órticos

Neossolos são solos que não apresentam alterações expressivas em relação ao material de origem devido à baixa atuação dos processos pedogenéticos, têm perfil pouco evoluído com ausência de horizonte B diagnóstico de qualquer tipo. Os Neossolos Quartzarênicos Órticos não apresentam contato lítico ou fragmentário em 50 cm a partir da superfície, têm sequência de horizontes A-C, mineralogia fundamentalmente quartzosa e não apresentam no perfil sinais da presença de lençol freático durante período prolongado (EMBRAPA, 2018).

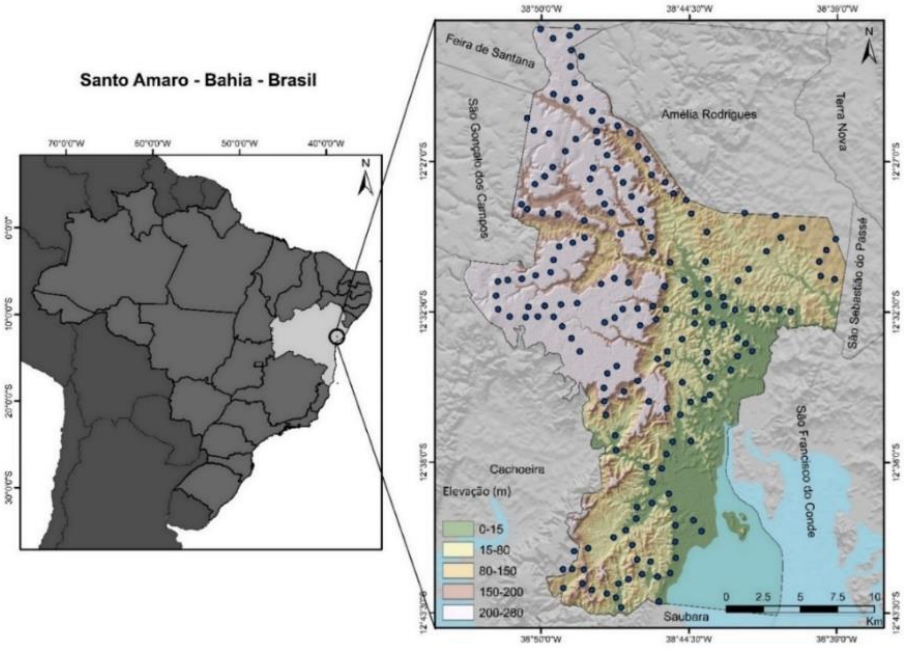
1.4.5 Gleissolos Tiomórficos Órticos

Gleissolos são solos minerais, hidromórficos, com horizonte glei na profundidade de até 50cm a partir da superfície. Esses solos permanecem saturados por água permanente ou periodicamente, gerando um ambiente redutor que implica no processo de redução e solubilização do ferro, fazendo com que o perfil do solo apresente cores acinzentadas, azuladas ou esverdeadas. Sedimentos recentes são os principais materiais de origem desses solos, que predominam em áreas planas de terraços fluviais, lacustres ou marinhos e em abaciados, depressões ou adjacentes às linhas de drenagem. Os Gleissolos Tiomórficos Órticos apresentam material ou horizonte sulfídrico dentro de 100cm a partir da superfície e não possuem características para serem enquadrados no grande grupo hístico (EMBRAPA, 2018).

1.4.6 Espodossolos Ferri-Humilúvicos

Espodossolos são solos minerais que apresentam horizonte B espódico, formado a partir da eluviação de matéria orgânica humificada e alumínio. São solos de baixa fertilidade, acidez moderada a forte, baixa saturação de bases, altos teores de alumínio. Sua pedogênese está ligada a materiais arenoquartzosos em condições de elevada umidade, distribuem-se em regiões de clima tropical com relevo plano a suave ondulado, principalmente em áreas de surgente, abaciados e depressões. Espodossolos Ferri-Humilúvicos Hidromórficos apresentam horizonte B espódico em profundidade igual ou menor a 200cm e sinais de saturação por água dentro de 100cm a partir da superfície (EMBRAPA, 2018).

Figura 3. Localização do município de Santo Amaro/BA,dos pontos amostrais e hipsometria da área.



3. MATERIAL E MÉTODOS

3.1 Área de estudo

O presente trabalho foi desenvolvido em uma área de 492,91 km², com coordenadas centrais 12° 34' 23'' S e 38° 42' 53'' W, correspondente aos limites do município de Santo Amaro, localizado na região do Recôncavo, Bahia (SEI-BA, 2018). O clima predominante é o tropical sem estação seca, Af no sistema de classificação de Koeppen (ALVARES et al., 2014), com precipitação e temperatura média anual de 1900mm e 33°C, respectivamente (CPRM, 2011). A Figura 3 apresenta a localização da área de estudo e dos pontos amostrais de solos utilizados neste estudo.

3.2 Dados pedológicos

Foram coletadas por tradagem 546 amostras de solo nas camadas de 0-5cm, 5-20cm e 20-40cm, em esquema de malha aleatória, correspondente a 182 pontos georreferenciados. As amostras foram secas ao ar, destorroadas e peneiradas em malha com 2mm de diâmetro, uma alíquota de 2g foi separada para análise na fluorescência de raio-x. As análises foram executadas em laboratório com dois espectrômetros Bruker Titan 600 (Billerica, MA) utilizando o método 6200 para análises com pXRF (USEPA, 2007). A calibração do equipamento foi realizada com amostras referência de solos certificados Montana, CS-M2, San Joaquim e IPT-32, que foram analisados 10 vezes cada. A partir dos dados da calibração foram calculados coeficientes de variação (Tabela 1). Curvas de calibração foram geradas por regressão linear em uma planilha de cálculo utilizando os valores medidos para as amostras referência e os valores certificados que são fornecidos pelo fabricante do equipamento pXRF. As equações de regressão linear foram utilizadas para corrigir as concentrações das análises das amostras.

O software Geochem (BRUKER DALTONICS INC., BILLERICA, MA, USA) e a aplicação Dual Soil, que são fornecidos pelo fabricante do equipamento foram utilizados para especificar o material a ser analisado e a unidade de concentração dos elementos. A concentração elementar de cada amostra foi avaliada três vezes e a duração de cada avaliação foi de 120 segundos. O tempo de leitura foi escolhido para atingir as concentrações mínimas detectáveis pelo aparelho (LOD) e as mesmo tempo proporcionar agilidade ao processo de análise, de acordo com metodologia sugerida por Kalnicky e Singhvi (2001).

Tabela 1. Precisão da leitura dos elementos nos solos referência para XRF 1 e XRF 2 e homogeneidade das triplicatas amostrais para as três profundidades estudadas.

	CV %							
		Mg	K	Ca	Al	Si	Ti	Fe	XRF
Padrão	Montana	16.2	0.2	0.4	1.5	0.8	1.4	0.3	1
	Montana	16.4	0.5	0.4	1.6	0.9	0.5	0.2	2
	CS-M2	13.2	0.2	0.5	1.0	0.3	0.7	0.4	1
	CS-M2	20.1	0.4	0.4	2.3	0.5	1.1	0.3	2
	San Joaquim	9.2	0.5	0.4	1.0	0.6	0.9	0.5	1
	San Joaquim	7.9	0.3	0.3	1.6	0.2	0.6	0.3	2
	IPT-32	-	0.7	1.3	1.0	0.4	0.4	0.6	1
	IPT-32	-	0.9	1.1	1.2	0.7	0.8	0.4	2
Amostras	0 a 5 cm	4.9	2.5	3.7	2.0	1.2	1.5	1.5	-
	5 a 20 cm	7.1	5.6	8.5	4.2	3.0	3.8	3.9	-
	20 a 40 cm	5.4	3.6	4.6	3.0	2.0	2.9	2.5	-

3.3 Covariáveis

A composição do conjunto inicial de covariáveis foi fundamentada no conceito de desenvolvimento de solo (SCORPAN) proposto por McBratney et al., (2003) e no conhecimento de campo construído durante as campanhas de campo para coleta das amostras. O conjunto inicial consistia em 73 covariáveis correspondentes aos seguintes fatores de formação do solo, relevo, material de origem, clima e organismos, além de covariáveis representantes de propriedades dos solos e de localização espacial.

Litologia e pedologia são conhecidamente fatores com alto grau de importância para explicar a variação espacial na concentração de elementos na área, porém esses dados não foram utilizados nesse estudo por dois motivos; inconsistências no delineamento, e escalas pouco detalhadas das bases cartográficas já existentes. Além disso pretende-se utilizar os mapas gerados neste trabalho como variáveis explicativas em futuros trabalhos de mapeamento do material de origem e de classes de solos em Santo Amaro/BA, desse modo utilizar bases pedológicas e litológicas para mapear os elementos, utilizando esses dados para novamente mapear solos e rochas incorreria em redundância, mascarando os resultados e a performance dos algoritmos em trabalhos posteriores.

3.3.1. Sensoriamento Remoto

Para geração das covariáveis de sensoriamento remoto foi processada na plataforma Google Earth Engine® uma coleção de imagens do Landsat 8 – Surface Reflectance Tier 2 no período entre 01/2016 e 07/2021. As cenas foram submetidas à um filtro para extração de nuvens, e foi gerada uma imagem sintética com pixels contendo os valores médios no período selecionado, as bandas dessa imagem foram utilizadas como covariáveis. Na calculadora raster do software ArcGIS 10.2 (ESRI, 2013) as bandas da imagem do Landsat 8 foram utilizadas para calcular índices de solo, vegetação e minerais, os quais foram também utilizados como covariáveis.

3.3.2. Relevo

Para a confecção do MDE, curvas de nível correspondentes a um mapa topográfico em escala 1:25.000 gerado por aerofotogrametria foram acessadas na base de dados digitais do Exército Brasileiro (DSG, 2019) e interpoladas por meio do algoritmo *Topo to raster* no ArcGIS 10.2 (ESRI, 2013), a partir do qual foram derivadas 40 variáveis morfométricas usando um script desenvolvido no software R (R CORE TEAM, 2022) que acessa as funções de análise de terreno do software SAGA-GIS (CONRAD et al., 2015), através do pacote Rsaga (BRENNING, 2008). Para harmonização entre as bases de dados o MDE foi gerado com 30m de resolução espacial.

3.3.3. Localização

Covariáveis representantes da localização espacial (N) foram geradas a partir do MDE utilizando uma rotina desenvolvida no software R para extrair a latitude e longitude em cada célula, a partir de onde foram criados dois rasters, X (long), e Y (lat). Através da ferramenta *Map álgebra* no ArcGIS 10.2 (ESRI, 2013) foi gerado um raster de localização de segunda ordem, calculado através da equação descrita por (COSTA et al., 2020).

$$XY = \frac{(X^2+Y^2+X*Y)}{10^6} \quad (1)$$

3.3.4. Clima

A covariável Radiação Solar foi gerada no ArcGIS 10.2 (ESRI, 2013) utilizando o MDE como dado de entrada no algoritmo *Area Solar Radiation*, que foi configurado para calcular a radiação incidente na área de estudo, em WH/m², com intervalo diário, durante o período do ano de 2021.

3.3.5 Seleção de covariáveis

O conjunto inicial com 72 covariáveis foi submetido à um processo de eliminação dividido em três etapas, exclusão por baixa variância (I), por alta correlação (II) e seleção de covariáveis por importância (III). A seleção de covariáveis é um processo que tem por objetivo determinar um subconjunto ótimo que consiga equilibrar o número de covariáveis e a acurácia do modelo (CHEN et al., 2018). Entre os principais motivos para selecionar subconjuntos de covariáveis pode-se citar: rapidez na calibração dos modelos, redução da complexidade, aumento da acurácia na predição, reduzir a possibilidade de ocorrência de multicolineariedade e prevenir a ocorrência de over-fitting (WADOUX et al., 2020).

A primeira etapa do processo foi de exclusão da covariáveis com baixa variância, executada através da aplicação da função *nearZeroVar* disponível no pacote *Caret* (KUNH, 2020). Covariáveis com variância zero, ou próxima a zero, carregam pouca informação e praticamente não têm efeito nos cálculos de modelagem (KUHN & JOHNSON, 2013). Na segunda etapa foram excluídas covariáveis que apresentavam alta correlação entre si, utilizando a função *findcorrelation* disponível no pacote *Caret* (KUNH, 2020). A presença de covariáveis altamente correlatas em um conjunto causa redundância, preditor aumenta o tempo de processamento e a complexidade do modelo, reduz a acurácia e, dificulta o entendimento das relações entre variáveis preditoras e alvo (BEHRENS et al., 2010; TEN CATEN et al., 2013; CAMPOS et al., 2018). Após finalização dos procedimentos I e II houve uma redução de 72 para 42 covariáveis, que corresponde a uma redução de 42% no número de covariáveis.

Na terceira etapa as covariáveis foram selecionadas pela importância, esse procedimento foi executado durante o processo de modelagem, após a separação dos dados em treino e teste, e foi executada no conjunto de treino, em loop, antes do ajuste do modelo final, com objetivo de selecionar um subconjunto de covariáveis que melhor se ajustavam a cada um dos algoritmos na predição das variáveis alvo. O RFE – *Recursive feature elimination* (GUYON et al., 2002; KUHN & JOHNSON, 2013) disponível no pacote *Caret* (KUNH, 2020) foi utilizado nessa etapa como ferramenta para seleção das covariáveis. O RFE é um método do tipo “wrapper” que seleciona as covariáveis pela recalibração iterativa de modelos de machine learning. Ao executar repetidas vezes o processo torna-se possível excluir covariáveis que apresentam baixa relevância para ajuste do modelo, com pouco ou nenhum decréscimo na acurácia da predição (WADOUX et al., 2020). O algoritmo identifica subconjuntos de covariáveis que apresentam melhor ajuste a partir da construção de um modelo de classificação com o conjunto completo de covariáveis, ranqueando-as por relevância, eliminando as variáveis com menor importância (KUHN & JOHNSON, 2013).

No presente trabalho foram definidos subconjuntos de 5, 10, 15, 20 ,25, 30 e 40 covariáveis para ajuste do RFE. Os modelos gerados foram avaliados pelo método de validação cruzada repetida (repeatedcv), com 10 folds e 3 repetições, o erro médio absoluto (MAE) foi a métrica utilizada para avaliar a acurácia e selecionar o subconjunto de covariáveis com melhor ajuste, o qual era posteriormente utilizado no processo de predição da variável alvo. A função rfe no *Caret* pode ser aplicada na seleção de covariáveis para diferentes modelos preditivos, desse modo é necessário

utilizar “funções de bases” específicas para cada algoritmo avaliado (KUNH & JHONSON, 2013). A “função de base” “*rfFuncs”* foi utilizada no algoritmo RF, para os outros algoritmos foi utilizada a função *“caretFuncs”*.

O uso do RFE como ferramenta para seleção de covariáveis no mapeamento digital de solos é recorrente na literatura, tendo sido aplicado em estudos para predição espacial de classes de solos (BRUNGARD et al., 2015; MEIER et al., 2018; TAGHIZADEH-MEHRJARDI et al., 2019; JEUNE et al., 2018), espessura do solo (CHEN et al., 2021), carbono (JEONG et al.,2017), matéria orgânica do solo (LUO et a., 2022), nitrogênio e fósforo (JEONG et al., 2017).

Figura 4. Exemplos de covariáveis utilizadas no processo de modelagem: (**a**) MDE (m), (**b**) Coordenadas oblíquas (m), (**c**) Radiação solar (WH/m²), (**d**) Altitude normalizada (adimensional), (**e**) Declividade (°), (**f**) Aspecto (°), (**g**) Índice de composição do solo (adimensional), (**h**) Índice de argila (adimensional) e (**i**) NDVI (adimensional).

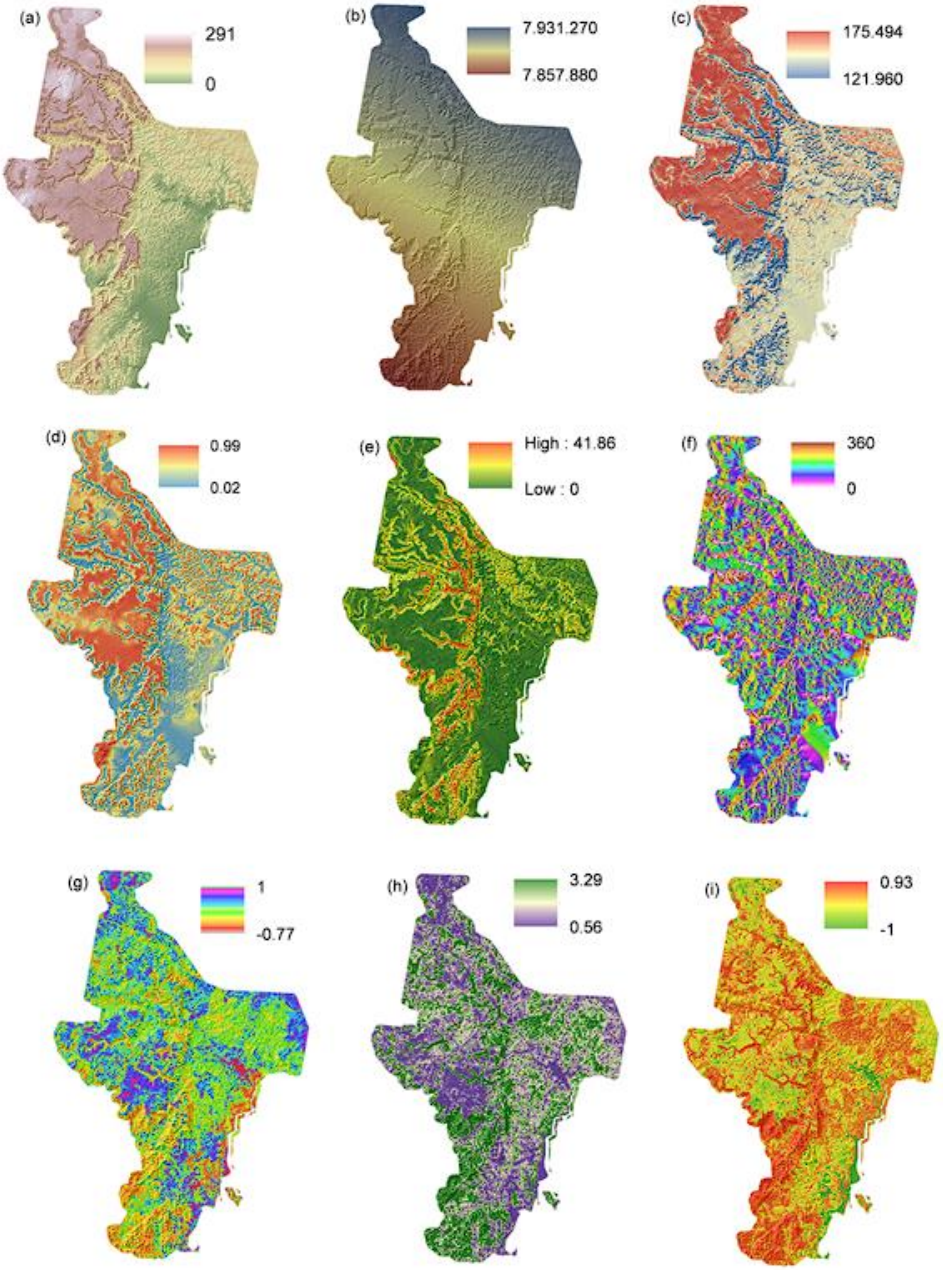


Tabela 2. Covariáveis preditoras utilizadas no mapeamento de elementos totais em Santo Amaro/BA.

Covariáveis	Descrição	Fonte	SCORPAN
Sensoriamento Remoto			
Bandas: 1 - 3 - 5 - 6 - 8 - 10	-	Landsat 8	**-**
Normalized Difference Vegetation Index	NDVI $\frac{NIR - RED}{NIR + RED}$	-	**O**
Iron Normalized Ratio	INR $\frac{RED - SWIR\ II}{RED + SWIR\ II}$	USGS	
Clay Normalized Ratio	CNR $\frac{SWIR\ I - SWIR\ II}{SWIR\ I + SWIR\ II}$	USGS	
Simple Ratio Clay Minerals	SRCM $\frac{SWIR\ I}{SWIR\ II}$	IDB	**P**
Ferrous² Mineral Ratio	FMR^{2+} $\frac{SWIR\ II}{NIR} + \frac{GREEN}{RED}$	(Rowan, 2003)	
Simple Ratio Iron Oxide	SRIO $\frac{RED}{BLUE}$	(Hewson, 2001)	
Soil Composite Index	SCI $\frac{SWIR\ I - NIR}{SWIR\ I + NIR}$	(Al-Khaier,2003)	
Bare Soil Index	$BSI = \frac{(SWIR\ II + RED) - (NIR + B)}{(SWIR\ II + RED) + (NIR + B)}$	(Diek, 2017)	**S**
Normalized Difference Bare Land Index	NBLI = $\frac{R - TIR}{R + TIR}$	(Li, 2017)	
Localização			
XY	XY $\frac{(X^2 + Y^2 + X * Y)}{10^6}$	(Costa et al., 2019)	**N**
Clima			
Radiação Solar		(ESRI., CA)	**C**
Topográficas			
Aspect			
Convergence index			
Curvature flow line			
Curvature minimal			
Curvature profile			
Curvature total			
Curvature general			
Curvature maximal			
Curvature plan			
Curvature tangencial			
Curvature classification			
Diurnal anisotropic heat			
Hill			
Hill index			
Landforms tpi based		(DSG, 2019)	
MDE			
Mid slope position			**R**
Morphometric protection index			
MRRTF			
MRVBF			
Normalized height			
Real surface area			
Saga wetness index			
Slope height			
Slope index			
Surface specific points			
Terrain surface classification iwahashi			
Terrain surface convexity			
Terrain surface texture			
Valley			
Valley index			
Valley depth			

3.4 Algoritmos

Sete algoritmos de machine learning foram avaliados no mapeamento de elementos totais neste estudo, random forest (RF), support vector machine com kernel radial (SVMRadial), cubist (CUB), multivariate adaptive regression splines (MARS), gradient boosting machine (GBM), k- nearest neighboor (KNN) e monotone multilayer preceptron (mMLP). Esses algoritmos foram selecionados por representarem as diferentes famílias de preditores de machine learning (árvores de decisão, support vector machines, redes neurais, regressão linear, rule-based), mais utilizadas atualmente no âmbito do mapeamento digital de solo. Uma descrição resumida sobre cada um dos algoritmos é exposta nos subtópicos a seguir.

3.4.1 Random forest

Random forest são algoritmos de árvores de decisão que demonstram alta performance no mapeamento digital de solos (LAGACHERIE et al., 2022; MELO et al., 2022). O random forest constrói um conjunto de árvores de decisão baseado em particionamento binário recursivo, onde cada nó particiona-se a partir de divisões binárias de subconjuntos de variáveis preditoras selecionadas aleatoriamente (WERE et al., 2014). A seleção das observações aleatórias é executada a partir de um método de bootstrap chamado de “bagging”, e o processo é repetido diversas vezes para construir múltiplas árvores independentes umas das outras. Para avaliar o erro do modelo o algoritmo usa uma estratégia chamada out-of-bag, que consiste em separar dois terços das observações selecionadas para treino, e o restante das observações são utilizadas no teste (BREINMAN, 2001; KHALEDIAN & MILLER, 2020). Os valores preditos são calculados pela média dos valores obtidos a partir do número de árvores criadas, que é um paramâmetro definido pelo usuário (LIAW & WIENER, 2002).

3.4.2 Support vector machines

Support vector machines (SVM) são algoritmos de machine learning que apresentam bons resultados em pesquisas de mapeamento digital de solos (ESTÉVEZ et al., 2022; ARAUJO-CARRILLO et al., 2021). O SVM mapeia os dados utilizando uma linha (hiperplano) que separa os pontos, a distância entre o hiperplano e o ponto (vetor de suporte) mais próximo de cada classe é chamada de margem (ASSAMI & HAMDI-AÏSSA, 2019). O algoritmo consegue particionar dados complexos, que apresentam relações não lineares, a partir da construção de modelos lineares baseados em funções kernel radial ou polinomial em um espaço hiperdimensional (HEUNG et al., 2016). No novo hiperespaço o SVM constrói um hiperplano ideal que se ajusta aos dados e previsões com risco empírico mínimo e complexidade da função de modelagem (WERE et al., 2014). Os modelos lineares simples construídos no hiperespaço correspondem a relações não lineares existentes nos dados de entrada originais (KARATZOGLOU et al., 2006; SMOLA & SCHÖLKOPF, 2004).

3.4.3 Multivariate adaptive regression splines

MARS é um algoritmo de regressão multivariada não paramétrica capaz de mapear relações não lineares entre variáveis de entrada e saída (MEHDIZADEH et al., 2017), aplicado com sucesso em trabalhos de mapeamento digital de solos (KHANIFAR, 2022; DE BENEDETTO et al., 2022). O método consiste em dividir o conjunto total de dados de treinamento em segmentos lineares menores (splines) com diferentes gradientes (slope). As splines se conectam aditiva ou interativamente, construindo um modelo flexível capaz de emular comportamento lineares e não lineares entre variáveis alvo e preditoras. Os pontos de conexão entre as curvas são

chamados de nós e sinalizam regiões onde há mudança na interação entre as variáveis. O algoritmo é executado em duas fases, na primeira (forward) nós candidatos são colocados em posições aleatórias dentro do intervalo de valores cada variável preditora, buscando definir um par de splines. O processo de adição de splines é repetido até que um número máximo seja obtido, que normalmente resulta em um modelo complexo e com overfitting. Na segunda fase (backward) o modelo exclui splines redundantes, com baixa contribuição para predição (ZHANG & GOH, 2016; EVERINGHAM, 2011).

3.4.4 Cubist

Cubist é um algoritmo de árvores de regressão, não paramétrico, capaz de predizer relações não lineares entre variáveis preditoras e alvo (KUNH, 2012), que apresenta alta performance no mapeamento digital de solos (MENDES et al., 2020; HOUNKPATIN et al., 2022). A diferença entre o Cubist e outros algoritmos regressores de árvores de decisão é que um modelo de regressão linear é ajustado em cada nó da árvore (MILLER et al., 2015; POULADI et al., 2019). A estrutura das árvores é criada com base no conjunto de covariáveis fornecidas, a partir de onde são geradas regras baseadas no método boosting, que consiste em um treinamento de reforço que converte preditores de baixo em alto potencial, ao aplicar maior peso aos preditores de maior potencial (QUINLAN, 1992; KHALEDIAN & MILLER, 2020). O algoritmo aplica um conjunto regras condicionais para particionar os dados de treinamento, levando em consideração a associação destes com as variáveis preditoras. A variável alvo é predita utilizando modelos lineares multivariados que combinam os valores das variáveis preditoras para encontrar as regras que apresentam o melhor ajuste. (LACOSTE et al., 2014; APPELHANS et al., 2015)

3.4.5 Gradient boosting machine

Gradient boosting machine (GBM) é um algoritmo baseado em árvores de decisão que apresenta alta acurácia no mapeamento digital de solos (ESTÉVEZ et al., 2022; SAHIN, 2020). O algoritmo utiliza um método de aprendizado conhecido como “boosting”, que consiste em realizar ajustes sequenciais onde cada modelo subsequente busca corrigir os erros do modelo anterior (NAWAR & MOUAZEN, 2017). A cada iteração o algoritmo adiciona uma árvore de decisão mal modelada para que o modelo assimile e corrija a erro quadrado médio (no caso de regressão), então uma nova árvore é ajustada ao resíduo atual e adicionada ao modelo anterior para atualização resíduo. Ao ajustar as árvores de decisão ao resíduo o modelo consegue melhorar acurácia em regiões onde não performou bem (TOUZANI et at., 2017).

3.4.6 Redes neurais artificiais

Redes neurais artificiais são algoritmos com bons resultados em trabalhos de mapeamento digital de solos (BODAGHABADI et al., 2015; KALAMBUKATTU et al., 2018). Funcionam simulando redes neurais biológicas a partir da construção de um conjunto com múltiplas camadas compostas por centenas de neurônios, conectados através de coeficientes por onde as informações são transmitidas (WERE et al., 2015). O processo de predição das redes neurais funciona em dois estágios, no primeiro a rede é treinada para entender as condições de ocorrência da variável alvo através de padrões encontrados nas variáveis preditoras, que na arquitetura estão representadas pelos neurônios, os quais conectam-se por coeficientes. O ajuste desses coeficientes consiste no processo de aprendizado propriamente dito. Na segunda etapa os coeficientes ajustados são utilizados para predizer valores em células onde não há valores de treinamento (KHALEDIAN & MILLER, 2020; BEHRENS et al., 2005). No

presente estudo foi utilizado o algoritmo Monotone Multi-Layer Perceptron (mMLP) (LANG, 2005).

3.4.7 *k*-nearest neigbohrs

K-nearest neighbors (KNN) é um algoritmo cuja performance foi analisada em trabalhos de mapeamento digital de solos (MANSUY et al., 2014; CAHYANA et al., 2021). É um método não paramétrico, multivariado (KHALEDIAN & MILLER, 2020), baseado no conceito da Primeira Lei da Geografia de Tobler que estabelece relações entre variáveis com base na proximidade, onde valor do pixel é predito de acordo com a distância até os pontos amostrais mais próximos (HEUNG et al., 2016). O conjunto de variáveis preditoras é explorado de modo a determinar qual configuração se aproxima mais daquelas encontradas nos locais onde há dados observados. A similaridade entre o pixel predito e aquele com dado observado é medido em termos de distância euclidiana, (TAGHIZADEH-MEHRJARDI et al., 2015).

3.5 Modelagem preditiva espacial e avalição da performance dos modelos

Os 182 pontos amostrais com os dados dos elementos totais do solo determinados via pXRF, nas profundidades de 0-5cm, 5-20cm e 20-40cm, foram utilizados como dados de treinamento e teste dos modelos, sendo estes particionados em 75% para treino e 25% para teste. O primeiro conjunto de dados (75%) foi utilizado para calibrar os parâmetros do modelo através do método de validação cruzada 10-fold com 5 repetições. O segundo conjunto de dados foi utilizado para estimar a performance dos modelos.

Três indicadores foram utilizados para mensurar a performance dos modelos, (1) raiz do erro quadrático médio (RMSE), (2) erro médio absoluto (MAE) e (3) coeficiente de determinação (R^2). RMSE e MAE descrevem a acurácia do modelo medindo a diferença entre os valores reais e preditos. O MAE é calculado pela média dos valores absolutos do resíduo, e é menos sensível a valores discrepantes (outliers) que o erro médio quadrático. O RMSE é calculado como a raiz quadrada da soma dos quadrados do resíduo e penaliza a ocorrência de grandes diferenças entre valores reais e preditos, demonstrando o impacto de outliers na predição. O R^2 mede o percentual da variância que é explicado pelo modelo (KHALEDIAN & MILLER, 2020). Para avaliação global da performance dos algoritmos as três métricas foram analisados em conjunto, os modelos de maior acurácia foram definidos como aqueles que apresentaram maior R^2 e menores MAE e RMSE simultaneamente.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}[\hat{y}_i - y_i]^2}{n}} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i-1}^{n}|y_i - y_j| \quad (3)$$

$$R^2 = 1 - \frac{\sqrt{\sum_{i-1}^{n}(\hat{y}_i - y_i)^2}}{\sqrt{\sum_{i-1}^{n}(\hat{y}_i - \overline{y}_i)^2}} \quad (4)$$

Onde $\hat{y}_i$, y_i , $\overline{y_i}$ e *n* são respectivamente os valores preditos, valores observados, médias dos valores preditos e tamanho da amostragem.

Os algoritmos foram executados no software R versão (R CORE TEAM, 2022) usando o pacote Caret - Classification and Regression Training (KUNH, 2022), para automatização e ajuste dos hiperparâmetros. Para gerar resultados robustos o procedimento foi executado 50 vezes para todas as variáveis, um loop foi construído, dentro da qual eram executadas a seleção de covariáveis através do RFE, partição dos dados, treinamento e validação dos modelos e predição espacial (mapas). O mapa final era gerado a partir da média das 50 repetições. A metodologia de múltiplas repetições é importante para controlar a variabilidade dos modelos preditos, já que durante o processo de partição dos dados em treino e teste diferentes conjuntos são gerados, resultando na obtenção de resultados distintos em cada repetição (KUNH & JHONSON, 2013). No mapeamento digital de solos essa prática é adotada em diversas pesquisas, porém a quantidade de repetições varia entre estudos.

3.6 Distribuição espacial dos elementos e relações com a Pedologia

Para avaliar as relações entre a distribuição espacial da cobertura pedológica e os teores dos elementos, foi executada uma análise PCA em uma matriz composta pelos valores da concentração dos elementos totais extraídos a partir dos mapas preditos e das informações de classes de solo extraídas a partir do mapa pedológico acessado no IBGE (2018). Inicialmente foi gerado um conjunto de 11.375 pontos a partir da transformação do raster de pedologia, com resolução de 200m, em um shapefile de pontos, onde cada ponto representava a classe de solo original do mapa pedológico do IBGE (2018). Em ambiente R, o conjunto de rasters dos teores de elementos foi empilhado utilizando a função *rast*, do pacote *terra*. Os pontos com informações das classes de solos foram carregados e utilizados para extrair os valores dos teores dos elementos na pilha dos rasters, gerando um dataframe que continha os dados das classes de solos e os valores de concentração dos elementos mapeados em cada ponto. Em seguida, os dataframes gerados para cada camada foram submetidos a análises de componentes principais (PCA), utilizando a função PCA, do pacote FactoMineR. Os resultados foram plotados em gráfico de formato biplot, que representa variáveis e indivíduos, permitindo analisar a ocorrência de padrões entre variáveis (elementos) e indivíduos (classes de solo).

Para analisar o valor da concentração média dos elementos nas manchas de solos do mapa pedológico (IBGE, 2018) foi utilizada a ferramenta *Zonal Statistics as Table*, no ArcGIS 10.2. Essa é ferramenta que executa operações estatísticas em valores das células nos rasters (teores de elementos, no caso do presente estudo), utilizando zonas definidas por outra camada (classes de solo) e, entrega os resultados em forma de tabela. Cabe ressaltar que essas estratégias descritas acima foram adotadas estritamente como forma de avaliar a existência de padrões entre os mapas de elementos e de classes de solos, consistindo em análise de relação entre os mapas, e, não, em um estudo da concentração real dos elementos para cada classes de solo. A adoção baseia-se no fato de que não houve levantamentos suficientes de pontos amostrais observados para todas as classes de solo ocorrentes na área, portanto não havia maneira de analisar a ocorrência de relações classes de solo x teores de elementos com dados observados.

4.RESULTADOS E DISCUSSÃO

A tabela 3 apresenta os valores para as estatísticas descritivas, nas duas profundidades, das variáveis mapeadas nesse estudo. Al, Fe, Ti e K foram os elementos que apresentam valores de assimetria mais próximos a zero, indicando distribuição dos dados próxima ao normal. O Ti foi o único elemento a apresentar assimetria negativa, ocorrendo nas camadas de 5-20cm e 20-40cm. O Ca foi o elemento que apresentou maior assimetria, 7.76 na camada superior e 5.83 na camada inferior, indicando que os dados possuem distribuição assimétrica positiva forte, com alta concentração de valores menores que a média. Os elementos K, Fe e Ti apresentaram curtose negativa nas duas camadas mapeadas e Al na camada de 20-40cm, com todos os outros elementos apresentando valores de curtose levemente positiva, à exceção do Ca, que apresentou curtose positiva muito alta indicando a ocorrência de valores extremos no conjunto de dados.

Tabela 3. Estatística descritiva dos atributos químicos do solo para as profundidades de 0-5cm, 5-20cm e 20-40cm, em Santo Amaro-BA.

	Média	Desvio Padrão	Mediana	Mínimo	Máximo	Range	Skew	Curtose
				0-5cm				
				g/kg				
Mg	9,41	5,17	6,78	3,20	24,05	20,85	1,16	0,06
Al	133,41	61,52	119,39	6,81	304,53	297,72	0,54	-0,18
Si	589,66	138,59	558,49	376,39	1016,12	639,73	0,84	0,08
K	7,68	9,10	2,01	0,27	31,30	31,03	1,05	-0,22
Ca	5,21	8,59	1,87	0,03	84,47	84,44	5,13	39,70
Ti	8,06	3,64	7,83	0,59	20,09	19,50	0,11	-0,16
Fe	38,14	29,26	28,09	1,34	150,82	149,48	0,82	-0,18
				5-20 cm				
				g/kg				
Mg	9,57	5,67	6,49	3,00	27,67	24,67	1,29	0,51
Al	113,02	46,87	108,72	0,97	268,17	267,20	0,27	0,03
Si	582,96	136,30	548,95	354,16	1025,88	671,72	1,12	0,85
K	7,95	9,54	1,88	0,04	31,62	31,58	0,93	-0,61
Ca	4,52	12,51	0,44	0,02	141,18	141,16	7,76	77,03
Ti	8,54	3,61	8,92	0,83	18,11	17,28	-0,12	-0,38
Fe	41,48	30,73	29,78	0,55	140,03	139,48	0,67	-0,78
				20-40 cm				
				g/kg				
Mg	9,60	5,24	6,65	4,10	26,14	22,04	1,23	0,23
Al	138,94	57,08	119,58	1,60	310,85	309,25	0,58	-0,16
Si	521,93	122,07	485,18	328,53	1025,08	696,55	1,36	1,81
K	7,36	8,47	2,61	0,05	27,36	27,31	0,84	-0,79
Ca	5,10	14,03	0,43	0,03	132,22	132,19	5,83	41,81
Ti	8,81	3,68	9,09	0,88	18,39	17,51	-0,23	-0,64
Fe	43,56	30,11	34,19	0,82	143,47	142,65	0,57	-0,82

Assimetria e curtose são medidas da distribuição de frequência dos dados. Assimetrias positivas indicam que há concentração de dados em valores mais baixos que a média, distribuições assimétricas negativas concentram dados com valores mais altos que a média, valores de assimetria próximos à zero indicam distribuição normal

dos dados. O valor da curtose para distribuição normal (mesocúrtica) é zero, distribuições leptocúrticas têm valor de curtose positivo e geralmente indicam contaminação ou presença de valores extremos nos dados. Curtoses negativas correspondem a distribuições platicúrticas, apresentando curvas mais achatadas (CAIN et al.,2017).

4.1 Comparação dos algoritmos

Avaliando os resultados das métricas (R^2, RMSE e MAE) observa-se que houve variação na performance dos algoritmos para diferentes variáveis, e para a mesma variável em diferentes camadas. O desempenho dos algoritmos de machine learning depende do conjunto de dados, e a performance de diferentes modelos pode variar no mesmo conjunto de dados (ROSSEL & BEHRENS, 2010). Desse modo avaliar a performance de múltiplos algoritmos para o mesmo conjunto de dados é uma prática adequada para gerar modelos com a melhor acurácia possível (HEUNG et al., 2016; PELEGRINO et al., 2021).

Na tabela 4 estão representados os resultados médios dos 50 modelos para o MAE, RMSE e R^2 calculados através da validação cruzada com 10 repetições e 5-folds para cada algoritmo avaliado, nas três profundidades. RF, GBM, CUB, SVMr e MARS foram os algoritmos que apresentaram os melhores ajustes, com performances muito próximas na maioria das variáveis. Apesar de apresentar boas performances em algumas variáveis, o ANN obteve ajustes consideravelmente inferiores aos outros algoritmos em boa parte dos elementos estudados. O kNN foi o algoritmo que apresentou pior performance em termos gerais, com ajustes significativamente inferiores na maioria dos elementos quando comparado aos outros algoritmos. A exceção no kNN fica por conta do elemento Si, onde o algoritmo apresentou bons ajustes em termos de métricas, porém, quando avaliado a qualidade da predição espacial os mapas gerados apresentaram baixa qualidade.

O RF foi o algoritmo que apresentou resultados superiores com maior frequência, com ajustes superiores, ou estatisticamente iguais aos modelos com maior média, na predição de quase todos os elementos nas três profundidades. Apenas para MAE no Ca, na profundidade de 5-20 cm, o RF foi inferior aos algoritmos com melhor ajuste. Random forest é um dos algoritmos mais estudados no mapeamento digital de solos, na literatura é possível encontrar diversos trabalhos que apresentam resultados semelhantes aos obtidos aqui. Zhang & Shi, (2019) reportaram performance superior do RF quando comparado a outros algoritmos (SVM, MLP, kNN e XGB) na predição espacial de classes texturais de solos. Brungard et al., (2015) testaram oito algoritmos de machine learning, entre eles RF, SVMr, kNN, MLP, na predição de classes de solos, e observaram a obtenção dos melhores resultados com o RF, utilizando o RFE como ferramenta para seleção de covariáveis. Campbell et al., (2019) testando sete algoritmos diferentes na predição de elementos totais no solo reportaram o RF como algoritmo com melhor performance geral, obtendo resultados superiores de predição de três (Fe, Mn e V), entre dez elementos na profundidade de 0-10cm, e sete (Al, Ca, Fe, K, P, Pb, Zr), entre onze elementos, na profundidade de 10-30cm.

Minasny et al., (2018) reportaram performances superiores do RF e CUB em relação ao GBM, SVMr, kNN e ANN no mapeamento de estoque de carbono em turfeiras na Indonésia. Contudo, os autores sinalizam que a seleção de covariáveis pelo método wrapper (ex: RFE), utilizando o RF como algoritmo de base, pode ter otimizado a seleção covariáveis para uso em modelos de árvore de decisão. Para evitar esse efeito e possibilitar uma comparação eficaz entre os algoritmos, no presente trabalho a seleção de covariáveis foi realizada com o mesmo algoritmo que seria utilizado no ajuste da predição. Assim, o algoritmo utilizado no ajuste do modelo

para seleção de covariáveis no RFE era o mesmo que executaria o processo de predição final.

GBM e CUB foram os algoritmos que apresentaram os melhores resultados após o RF, com performances muitos similares entre si. Em termos de R^2 o GBM obteve ajustes superiores, ou estatisticamente iguais aos algoritmos com melhor média, em 6 elementos nas três camadas (Al, Ca, K, Mg, Si e Ti). No MAE o GBM obteve alta performance em 6 elementos na primeira camada (Al, Ca, K, Mg, Si e Ti), 5 elementos na segunda camada (Al, K, Mg, Si e Ti) e quatro elementos na terceira camada (Al, K, Si e Ti). Avaliando o RMSE o GBM obteve alta performance em 6 elementos na primeira e terceira camada (Al, Ca, K, Mg, Si e Ti) e em todos os elementos na segunda camada. Campbell et al., (2019) também constataram o GBM como algoritmo com alta performance no mapeamento digital de solos, com ajustes superiores a outros modelos na predição de K na profundidade de 0-10cm, e de Ti e V na profundidade de 10-30cm. Estévez et al., (2022) obtiveram bons ajustes no mapeamento de sulfato ácido em solos utilizando o GBM e RF, indicando ainda que o GBM produziu modelos com acurácia de 5 a 6% superior ao RF.

O CUB obteve ajustes superiores para o R^2, ou estatisticamente iguais aos algoritmos de melhor média, na predição de 5 elementos na primeira camada (Fe, K, Mg, Si e Ti) e 6 elementos na segunda camada (Al, Fe, K, Mg, Si e Ti) e 6 elementos na terceira camada (Ca, Fe, K, Mg, Si e Ti). Para o MAE o algoritmo apresentou alta performance na predição de 3 elementos na primeira camada (Ca, Fe e K), em todos os elementos na segunda camada, e 6 elementos na terceira camada (Ca, Fe, K, Mg, Si e Ti). Avaliando o RMSE, CUB apresentou bons resultados na predição de 5 elementos na primeira camada (Ca, Fe, K, Mg e Ti), e em todos os elementos na segunda e terceira camadas. Na literatura foram encontrados resultados semelhantes obtidos por outros autores, para o referido algoritmo. Campbell et al., (2019) reportaram ajustes superiores do CUB na predição de dois elementos (Pb e Ti) na profundidade de 0-10 cm. Mendes et al., (2020) reportaram o CUB como melhor algoritmo na predição espacial da Saturação de bases (V%) na camada superficial.

O MARS foi o quarto algoritmo em termos de performance geral, apresentando resultados superiores, ou estaticamente iguais aos algoritmos com maior média no R^2 em, 3 elementos na primeira camada (Fe, K e Ti) e 6 elementos na segunda e terceira camadas (Al, Fe, K, Mg, Si e Ti). Analisando o MAE, o algoritmo apresentou bons resultados em 1 elemento na primeira camada (Fe), 3 elementos na segunda camada (Al, Si e Ti) e 4 elementos na terceira camada (Al, Fe, Si e Ti). Para o RMSE, o MARS obteve bons resultados na predição de 3 elementos na primeira camada (Fe, K e Ti), 5 elementos na segunda camada (Al, Fe, Mg, Si e Ti) e em todos os elementos na terceira camada. Outros estudos encontrados na literatura confirmam a boa performance do MARS em trabalhos de mapeamento digital de solos. Adler (2022) reporta o MARS como algoritmo com melhor performance na predição espacial de Zn e Cd, quando comparado ao RF e MLR. Nawar et al., (2014) observaram ajustes superiores do MARS, em relação ao PLSR, no mapeamento da salinidade em solos de uma região semiárida.

O SVMr está entre os algoritmos com melhores ajustes em grande parte dos elementos mapeados no presente estudo, apesar de obter performance superior na predição de apenas um elemento, Ca na profundidade de 5-20cm, os valores das métricas obtidos pelo SVMr para os outros elementos é sempre próximo ao dos algoritmos com performance superior. Este algoritmo é amplamente utilizado em trabalhos de mapeamento digital de solos, com bons resultados reportados na literatura. Were et al., (2014) reportaram o SVMr como algoritmo com melhor performance, quando comparado ao ANN e RF, na predição de SOC. Heung et al.,

(2016) observaram ajustes superiores do SVMr, quando comparado a outros algoritmos (RF, ANN, MLR, CART),na predição espacial de classes de solos.

ANN apresentou bons ajustes para R^2 na predição de 2 elementos na primeira camada (Mg e Ti), 2 elementos na segunda camada (Si e Ti) e 2 elementos na terceira camada (Mg e Ti). Para o MAE o algoritmo obteve bons ajustes em 1 elemento na primeira camada (Mg), dois elementos na segunda camada (Mg e Ti) e 1 elemento na terceira camada (Mg). Para o RMSE ANN conseguiu bons ajustes em 1 elemento na primeira camada, 1 elemento na segunda camada (Si) e 2 elementos na terceira camada (Ca e Si). Khaledian and Miller (2020) afirmam que a maior limitação no uso do ANN é a sensibilidade dos algoritmos dessa família ao tamanho do conjunto de amostra de treinamento, sendo necessário uma grande quantidade de amostras para que os algoritmos dessa família possam predizer com alta performance. Apesar disso em alguns trabalhos é possível observar alta performance das ANNs com um baixo volume de amostras, à exemplo do mapeamento do conteúdo de umidade do solo executado por Mahmoudabadi et al., (2017) onde os autores conseguiram um valor de 0.85 para o R^2 utilizando um conjunto com 137 pontos amostrais. Taghizadeh-Mehrjardi et al., (2016) reportaram o ANN como algoritmo superior ao RF, SVMr e KNN na predição de SOC, utilizando um conjunto de 188 pontos amostrais.

KNN foi o algoritmo que apresentou os piores ajustes de maneira geral, obtendo bons ajustes apenas para o Si, onde foi superior ou estatisticamente igual aos algoritmos com maior média, além disso os mapas gerados pelo KNN eram em sua maioria altamente generalistas, com limites artificiais, e com efeito dominante da covariável de localização espacial (XY). Heung et al., (2016) em seu trabalho de mapeamento digital de classes de solos sinalizam que, apesar de obter boa performance nas métricas, os mapas gerados pelo KNN apresentavam alto nível de ruído e eram de difícil interpretação, o autor atribui isso ao efeito do overfitting. Em termos de performance, Taghizadeh-Mehrjardi et al., (2019) reportaram o KNN como o algoritmo com os piores ajustes na predição espacial de classes de solos, quando comparado a outros quatro algoritmos (RF, XGB, C5.0 e SVM).

A figura 5 mostra os gráficos de violino com a distribuição dos valores R^2 obtidos nas 50 repetições, para cada elemento e algoritmo nas três camadas. Além de retratar os intervalos interquartis, a mediana, a ocorrência de outliers, valores máximos e mínimos, os gráficos de violino conseguem representar toda a distribuição dos dados. A variabilidade dos valores de R^2 obtidos em cada repetição individual e, observada nos gráficos, deixa evidente a necessidade de executar repetidas vezes o processo de modelagem, utilizando o valor médio dos pixels de todas as repetições para sintetizar o mapa final, sendo possível desse modo reduzir o efeito da variabilidade observada nos mapas gerados por predições individuais. Kunh and Johnson, (2013) afirmam que essa é uma prática importante para controlar a variabilidade inerente ao processo de formação do conjunto de dados durante a etapa de partição em treino e teste.

Na literatura ainda não foi possível encontrar estudos que avaliassem o efeito do uso de diferentes números de repetições durante o processo de modelagem, na qualidade dos resultados obtidos em mapeamento digital de solos. Porém, entre os trabalhos revisados observou-se uma tendência no uso de 50 ou 100 repetições. Jeong et al., (2015), no mapeamento de nutrientes em solos na Coréia do Sul, e Siqueira et al., (2021) no mapeamento de geoformas na Antártica, ambos utilizando abordagem de machine learning, estão entre alguns exemplos de autores que utilizaram 100 repetições durante o processo de modelagem. Em nosso trabalho, devido à quantidade de variáveis e camadas estudadas e, levando em consideração o custo computacional das análises, optamos pelo uso de 50 repetições. Entre os

autores que também adotaram 50 repetições em seus trabalhos estão, Gomes et al., (2019) no mapeamento de estoques de carbono no Brasil e Chen et al., (2021), no mapeamento da espessura de solos originados de sedimentos *loess* na França.

Tabela 4. Tabela com a média do R^2 para os algoritmos avaliados no mapeamento de elementos totais em solos nas camadas de 0-5cm, 5-20cm e 20-40cm. Santo Amaro/BA. Médias na mesma linha que não compartilham letras iguais apresentaram diferença estatisticamente significativa.

		R^2						
	ELEMENTO	CUB	GBM	KNN	MARS	ANN	RF	SVMr
0-5 cm	Al	0.46 b	0.56 a	0.31 c	0.48 b	0.44 b	0.56 a	0.47 b
	Ca	0.24 b	0.33 a	0.12 c	0.19 b	0.09 c	0.38 a	0.31 ab
	Fe	0.56 ab	0.48 b	0.18 d	0.57 ab	0.36 c	0.59 a	0.38 c
	K	0.83 a	0.83 a	0.33 c	0.80 a	0.73 b	0.82 a	0.72 b
	Mg	0.71 ab	0.73 ab	0.23 d	0.69 bc	0.70 ab	0.75 a	0.64 c
	Si	0.28 ab	0.29 ab	0.35 a	0.26 b	0.25 b	0.34 a	0.29 ab
	Ti	0.55 ab	0.56 a	0.48c	0.56 ab	0.55 ab	0.60 a	0.51 bc
5-20cm	Al	0.48 ab	0.50 a	0.37 c	0.52 a	0.43 bc	0.51 a	0.51 a
	Ca	0.26 b	0.23 a	0.07 c	0.17 b	0.08 c	0.25 a	0.31 ab
	Fe	0.58 ab	0.50 bcd	0.16 e	0.53 abc	0.45 cd	0.59 a	0.42 d
	K	0.81 a	0.83 a	0.31 c	0.79 ab	0.73 b	0.81 a	0.75 b
	Mg	0.67 abc	0.71 ab	0.24 d	0.67 abc	0.64 bc	0.71 a	0.63 c
	Si	0.29 a	0.31 a	0.35 a	0.30 a	0.28 a	0.32 a	0.27 a
	Ti	0.54 ab	0.57 a	0.45 c	0.52 ab	0.54 ab	0.56 a	0.50 bc
20-40cm	Al	0.54 bc	0.59 ab	0.30 e	0.56 abc	0.47 d	0.61 a	0.50 cd
	Ca	0.16 ab	0.21 a	0.06 c	0.11 b	0.07 c	0.22 a	0.20 a
	Fe	0.52 ab	0.44 bc	0.14 e	0.50 ab	0.33 d	0.54 a	0.36 cd
	K	0.80 a	0.82 a	0.32 c	0.79 a	0.71 b	0.81 a	0.73 b
	Mg	0.73 ab	0.73 ab	0.33 c	0.73 a	0.73 a	0.73 a	0.68 b
	Si	0.36 ab	0.38 ab	0.39 ab	0.35 ab	0.31 b	0.42 a	0.37 ab
	Ti	0.59 a	0.60 a	0.42 b	0.57 a	0.57 a	0.60 a	0.55 a

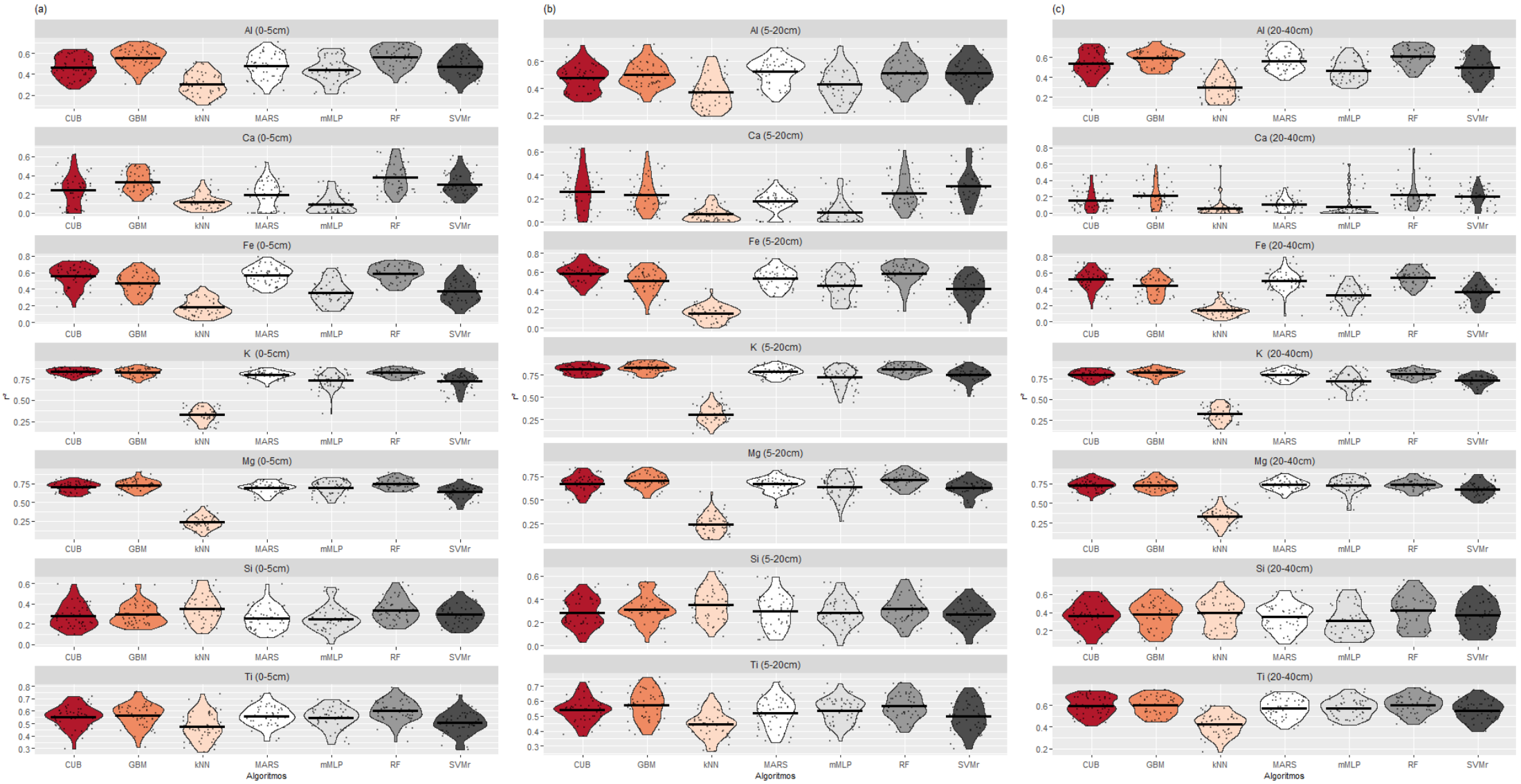
Tabela 5. Tabela com a média do RMSE para os algoritmos avaliados no mapeamento de elementos totais em solos nas camadas de 0-5cm, 5-20cm e 20-40cm. Santo Amaro/BA. Médias na mesma linha que não compartilham letras iguais apresentaram diferença estatisticamente significativa.

		RMSE						
	ELEMENTO	CUB	GBM	kNN	MARS	ANN	RF	SVMr
0-5 cm	Al	45.90 c	41.48 ab	52.04 d	45.54 bc	47.28 c	41.10 a	45.04 bc
	Ca	8.72 abc	6.78 ab	8.04 bcd	11.61 cd	13.68 d	6.63 a	7.02 ab
	Fe	19.99 ab	21.53 bc	27.34 e	19.49 ab	24.40 d	19.11 a	23.88 cd
	K	3.79 a	3.85 a	7.54 c	4.13 a	4.85 b	3.88 a	4.85 b
	Mg	2.83 ab	2.69 ab	4.56 d	2.88 bc	2.88 bc	2.58 a	3.13 c
	Si	120.97 bc	118.39 abc	112.55 a	127.66 c	123.49 c	113.38 a	117.11 abc
	Ti	2.45 ab	2.43 ab	2.65 c	2.43 ab	2.47 ab	2.30 a	2.57 bc
5-20cm	Al	34.58 ab	33.57 a	37.53 b	32.88 a	37.42 b	33.24 a	33.02 a
	Ca	10.04 ab	9.52 ab	10.86 b	12.24 b	26.68 c	9.87 ab	8.93 a
	Fe	20.09 a	21.78 a	28.84 c	21.25 a	23.16 b	19.83 a	23.88 b
	K	4.13 ab	3.92 a	8.01 d	4.39 bc	5.08 c	4.12 a	4.76 c
	Mg	3.29 ab	3.09 a	5.00 c	3.33 ab	3.54 b	3.06 a	3.47 b
	Si	118.25 a	114.77 a	110.89 a	117.52 a	119.17 a	114.27 a	118.67 a
	Ti	2.50 ab	2.41 a	2.73 c	2.55 ab	2.51 ab	2.43 a	2.61 bc
20-40cm	Al	39.80 abc	37.58 ab	49.05 d	38.70 ab	43.07 c	36.38 a	41.17 bc
	Ca	13.59 a	12.57 a	13.91 a	16.60 ab	28.09 b	12.90 a	12.64 a
	Fe	21.40 ab	22.80 bc	28.76 e	21.47 ab	25.42 d	20.62 a	24.69 cd
	K	3.91 a	3.66 a	7.08 c	3.89 a	4.67 b	3.80 a	4.49 b
	Mg	2.81 ab	2.78 ab	4.43 c	2.75 a	2.77 a	2.75 ab	3.07 b
	Si	100.53 ab	98.05 ab	97.25 ab	101.33 ab	106.53 b	94.37 a	98.82 ab
	Ti	2.37 a	2.35 a	2.82 b	2.44 a	2.43 a	2.34 a	2.49 a

Tabela 6. Tabela com a média do MAE para os algoritmos avaliados no mapeamento de elementos totais em solos nas camadas de 0-5cm, 5-20cm e 20-40cm. Santo Amaro/BA. Médias na mesma linha que não compartilham letras iguais apresentaram diferença estatisticamente significativa.

		MAE						
	ELEMENTO	CUB	GBM	kNN	MARS	ANN	RF	SVMr
0-5 cm	Al	35.29 b	31.09 a	39.16 c	35.12 b	36.32 bc	30.90 a	35.01 b
	Ca	3.86 a	3.80 a	4.67 b	5.11 b	6.12 b	3.53 a	3.63 a
	Fe	13.99 ab	15.30 bc	21.00 e	14.76 ab	16.60 cd	13.49 a	17.24 d
	K	2.47 a	2.57 a	5.39 c	3.06 b	3.16 b	2.53 a	3.36 b
	Mg	1.90 bc	1.81 ab	3.31 e	2.07 cd	1.83 ab	1.70 a	2.13 d
	Si	89.82 bc	88.96 abc	83.49 a	91.53 c	93.13 c	85.09 ab	88.20 abc
	Ti	1.93 b	1.87 ab	2.10 d	1.90 b	1.95 bc	1.77 a	2.01 cd
5-20cm	Al	26.02 ab	24.93 a	28.45 c	24.52 a	27.67 bc	25.05 a	24.79 a
	Ca	3.7 ab2	4.73 cd	5.35 de	5.17 cd	9.14 e	4.39 bc	3.61 a
	Fe	14.57 a	16.33 b	22.85 c	16.11 b	16.53 b	14.56 a	17.68 b
	K	2.62 a	2.57 a	5.72 c	3.07 b	3.23 b	2.59 a	3.23 b
	Mg	2.08 ab	2.06 ab	3.48 d	2.31 c	2.15 abc	1.99 a	2.27 bc
	Si	86.08 ab	85.64 ab	81.26 a	84.50 ab	88.72 b	84.84 ab	87.87 b
	Ti	1.94 ab	1.87 a	2.15 c	1.94 ab	1.96 ab	1.85 a	2.03 bc
20-40cm	Al	28.53 b	27.49 ab	34.78 d	28.22 ab	31.08 c	25.96 a	29.88 abc
	Ca	5.05 ab	5.66 bc	6.19 cd	7.01 cd	10.20 d	5.27 ab	4.62 a
	Fe	15.69 ab	16.95 bc	22.67 d	16.45 ab	17.96 c	15.18 a	18.13 c
	K	2.55 ab	2.47 a	5.12 e	2.73 bc	3.02 cd	2.52 ab	3.16 d
	Mg	1.80 ab	1.88 bc	3.15 d	1.91 bc	1.69 a	1.77 ab	2.06 c
	Si	68.06 a	68.04 a	67.46 a	69.50 ab	75.01 b	65.65 a	68.00 a
	Ti	1.85 abc	1.79 ab	2.21 d	1.86 abc	1.90 bc	1.77 a	1.96 c

Figura 5. Gráficos de violino mostrando a dispersão para os valores de R^2 obtidos nas 50 repetições em cada um dos algoritmos testados na predição de elementos totais em solos nas camadas de 0-5cm (a), 5-20cm (b) e 20-40cm (c) em Santo Amaro/BA.



4.2 Predição espacial dos elementos

Foram observadas variações na configuração espacial dos mapas de concentração dos elementos gerados por diferentes modelos, reforçando a importância de utilizar múltiplos algoritmos em estudos de MDS, avaliando a performance, e, selecionando os modelos de melhor ajuste para uso prático. Esses resultados foram relatados em outros estudos de comparação de algoritmos em MDS. Heung et al., (2016) observaram variações drásticas entre mapas de classes de solo gerados por diferentes algoritmos de machine learning, sugerindo que as pesquisas de DSM não se restrinjam a utilizar um único ou pequeno conjunto de algoritmos. Os autores ainda recomendam o uso pacote *caret* no R como uma ferramenta eficiente para comparação de múltiplos algoritmos. Pelegrino et al., (2021) observaram que os mapas de classes de fertilidades gerados em seu estudo variaram de acordo com o algoritmo utilizado, demonstrando a importância de utilizar múltiplos algoritmos para o mesmo conjunto de dados, avaliando a performance de cada um e, selecionando os modelos com melhor acurácia para uso prático e tomada de decisão.

As covariáveis representantes dos fatores relevo e posição espacial foram selecionadas com maior frequência para explicar a variação espacial dos elementos mapeados em Santo Amaro. MDE e XY apresentaram maior frequência de seleção entre todas as variáveis, seguidas por Normalized height, Slope heigh e Solar radiation, esta última, apesar de representar o fator clima, foi gerada à partir do MDE, apresentando características similares às covariáveis de relevo, desse modo em trabalhos posteriores recomenda-se o uso de covariáveis climáticas oriundas de bases de dados oficiais, não secundárias, para que seja possível avaliar com eficácia o efeito do clima nos objetos de estudo.

A alta frequência na seleção do MDE, e de covariáveis representantes do relevo, na modelagem espacial das variações dos teores dos elementos em Santo Amaro, está relacionada a litoestratigrafia da região, representada pela posição dos diferentes materiais de origem na topossequência da paisagem, e captada pelos algoritmos no MDE através dos intervalos de altitude. As rochas sedimentares que ocorrem em Santo Amaro apresentam características geoquímicas bastante distintas e, a posição desses diferentes materiais na paisagem tem grande influência sobre a variação espacial nos teores dos elementos químicos no município.

A litoestratigrafia é o estudo do empilhamento, ou sucessão estratigráfica vertical das unidades litológicas e da continuidade lateral dessas unidades. No caso da Bacia do Recôncavo esse empilhamento ocorreu em duas grandes sequências. Na sequência do continente, os sedimentos se depositaram durante a fase pré-rifte, envolvendo as formações: Aliança, Sergi, Itaparica e Água Grande. Na sequência do lago os sedimentos foram depositados durante a fase rifte, e envolvem as formações Candeias, Marfim, Pojuca, Taquipe, São Sebastião e Salvador (CAIXETA et al., 1994; GUZMAN et al., 2015).

A litologia em Santo Amaro apresenta padrões de distribuição em sentido longitudinais e latitudinais, fato que explica a importância da covariável XY para explicar a variação espacial dos teores de elemento. Além disso, ocorrências localizadas como o afloramento de rochas do Complexo Santa Luz e áreas de deposição dos sedimentos quaternários puderam ser rastreadas com o uso dos dados espaciais. Behrens et al., (2018) reportam em seu trabalho que o uso de dados de posição espacial (latitude, longitude, distância euclidiana etc.) permite aos algoritmos preditores identificar variações locais e modelar condições não estacionárias, incrementando a performance dos algoritmos de machine learning.

4.2.1 Alumínio

A Figura 6(a) apresenta os mapas de teor de alumínio gerados pelos algoritmos com melhor ajuste. O alumínio apresentou ajuste satisfatório no processo de modelagem espacial no presente estudo, com valores de R^2 variando entre 0.31 (kNN) e 0 .56 (RF e GBM) na camada de 0-5cm, 0.37 (kNN) e 0.52 (MARS) na camada de 5-20cm, e 0.30 (kNN) e 0.61 (RF) na camada de 20-40cm. Outros algoritmos com bons ajustes para o elemento em questão foram, SVMr (0.51) e GBM (0.50) na camada de 5-20cm. Campbell et al., (2019) observaram o GLMBOOST e RF como algoritmos com melhor ajuste para Al na camada de 0-10 cm, com valores de R^2 iguais a 0.40 e 0.39, respectivamente, já na camada de 10-30cm, RF e GBM apresentaram os melhores ajustes para o elemento em questão, com valores de R^2 iguais a 0.37 e 0.36 respectivamente.

A Figura 6(b) mostra os gráficos da frequência de seleção de covariáveis dos algoritmos utilizados para gerar os mapas finais, nas três camadas. MDE e XY foram as covariáveis com maior frequência de seleção, sendo aplicadas em todas as repetições nas diferentes camadas. Normalized height, Solar radiation e Slope height foram outras covariáveis com alta frequência de seleção para modelar a distribuição de alumínio na área. A Figura 6(c) mostra os gráficos de valores preditos e observados para o teor de Al, a Tabela 5 mostra os valores preditos e observados, máximos e mínimos.

Tabela 7. Valores observados e preditos, máximos e mínimos para Al em solos do município de Santo Amaro/BA

Al	Observados (g/kg)		Preditos (g/kg)	
	Máximo	Mínimo	Máximo	Mínimo
0-5 cm	304,53	6,81	267,77	32,31
5-20 cm	268,17	0,97	167,53	36,74
20-40 cm	310,85	1,60	239,51	53,29

Os maiores valores de Al foram mapeados nas regiões mais altas e planas da área de estudo, locais de predomínio de Latossolos e Argissolos Amarelos, refletindo a maior concentração de minerais aluminossilicatados e alumínio trocável na composição desses solos. Argissolos e Latossolos são solos bastante intemperizados que têm em sua composição um fração significativa de argilominerais de alumínio hidratado e possuem baixa saturação por bases. A Gibbsita é óxido de alumínio em maior concentração nesses solos, oriunda da alteração da Caulinita, que por sua vez, é formada pelo intemperismo intenso de rochas ricas no mineral primário feldspato (EMBRAPA, 2018). As manchas de concentração mais elevada de Al na área ocorrem no tabuleiro central, onde, de acordo com o mapa pedológico do IBGE (2018), predominam Latossolos Amarelos. Nos tabuleiros localizados ao norte há um predomínio de Argissolos Amarelos, região onde também foram mapeados teores altos de alumínio, porém sensivelmente mais baixos que aqueles do tabuleiro central. Nos modelos de distribuição espacial de alumínio é possível observar um gradiente espacial decrescente no teor do elemento nos tabuleiros, no sentido centro-norte, evidenciando a variação pedológica que ocorre nessa área. Os menores valores de Al foram mapeados nas regiões mais baixas da paisagem, em regiões de domínio dos Neossolos e Gleissolos. Reforçando as informações encontradas no presente estudo, Bonfim (2014) encontrou valores muitos baixos para a saturação de alumínio no complexo de troca em perfis de Gleissolos Tiomórficos distribuídos na bacia do Rio Subaé.

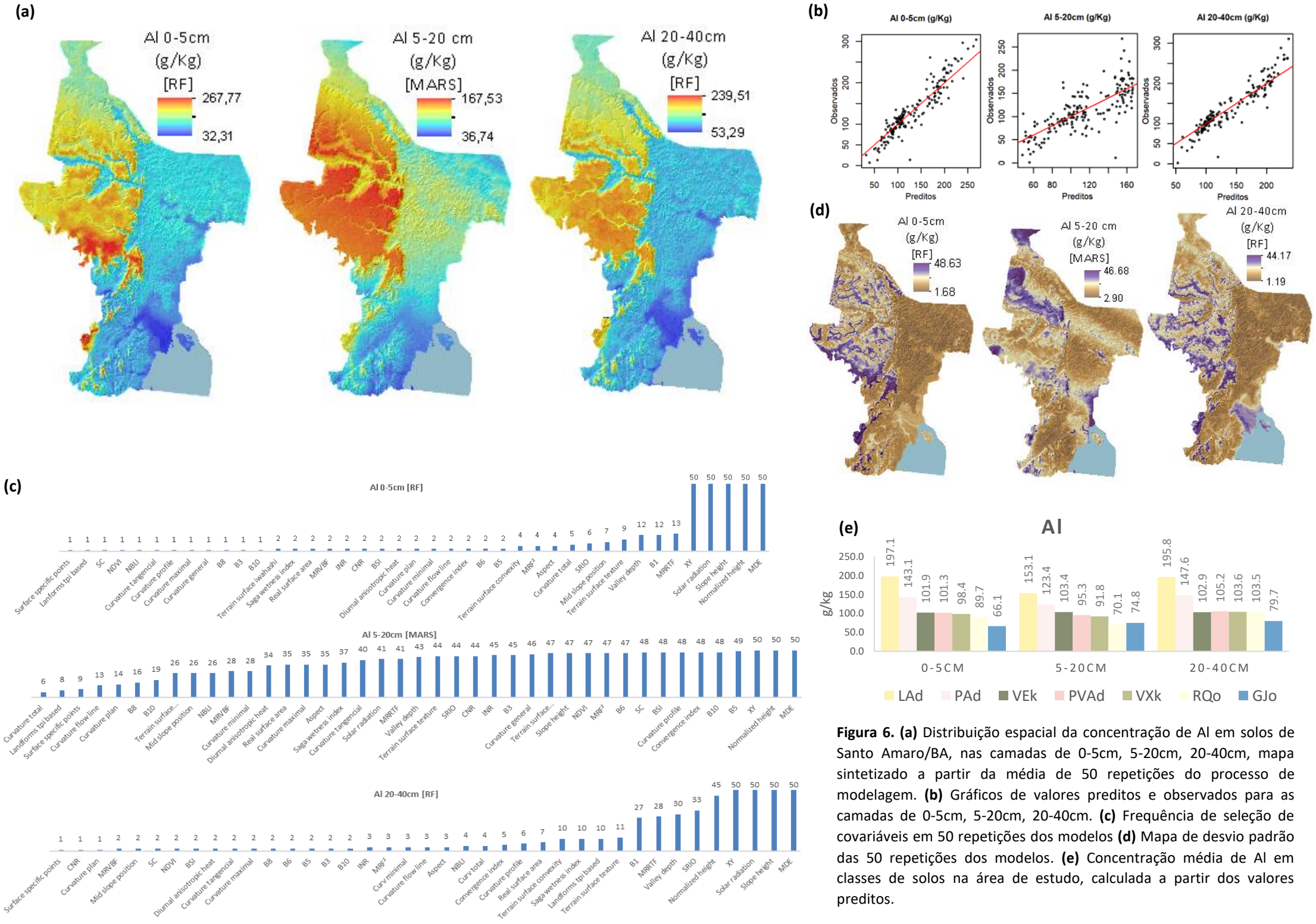


Figura 6. (a) Distribuição espacial da concentração de Al em solos de Santo Amaro/BA, nas camadas de 0-5cm, 5-20cm, 20-40cm, mapa sintetizado a partir da média de 50 repetições do processo de modelagem. **(b)** Gráficos de valores preditos e observados para as camadas de 0-5cm, 5-20cm, 20-40cm. **(c)** Frequência de seleção de covariáveis em 50 repetições dos modelos **(d)** Mapa de desvio padrão das 50 repetições dos modelos. **(e)** Concentração média de Al em classes de solos na área de estudo, calculada a partir dos valores preditos.

4.2.2 Cálcio

A variabilidade espacial do teor de Ca ao longo da área está representada nos mapas da Figura 7(a) .Dentre os elementos estudados o Ca foi o que apresentou pior ajuste, com valores de R^2 variando entre 0.09 (ANN) e 0 .38 (RF) na camada de 0-5cm, 0.07 (kNN) e 0.31 (RF) na camada de 5-20cm, e 0.06 (kNN) e 0.22 (RF) na camada de 20-40cm. RF, SVMr, GBM e CUB foram, nessa ordem, os algoritmos com melhor ajuste para o elemento. Resultados similares foram observados por Campbell et al., (2019) no mapeamento de Ca utilizando pXRF, com valores de R^2 que variaram entre 0.24 e 0.32 na camada de 10-30cm, sendo o RF o melhor algoritmo para predição espacial do elemento. Os autores afirmam em seu estudo que não obtiveram resultados satisfatórios no mapeamento de Ca na camada de 0-10cm.

Na camada de 20-40cm os valores das métricas para o Ca apresentaram um comportamento incomum. O algoritmo que obteve o melhor valor de R^2 (RF), não apresentou paralelamente os menores valores nas outras métricas, para o MAE o menor valor foi observado no SVMr, e para o RMSE o menor valor foi observado no GBM. A Figura 7(b) mostra o gráfico da frequência de seleção de covariáveis utilizadas na predição espacial de Ca. MDE e XY foram as covariáveis com maior frequência de seleção, sendo utilizadas em todas as repetições para as três camadas. Na camada de 0-5cm a covariável Solar radiation foi também selecionada em todas as 50 repetições. Na camada de 5-20cm um conjunto maior de covariáveis foi frequentemente selecionado para explicar a variação espacial do elemento, com variáveis oriundas de sensoriamento estando entre as mais utilizadas, à exemplo das bandas 1,5 e 3 do Landsat 8 e do índice CNR. A Figura 7(c) mostra o gráfico de valores preditos e observados para o teor de Ca, a Tabela 6 mostra os valores preditos e observados, máximos e mínimos.

Tabela 8. Valores observados e preditos, máximos e mínimos para Ca em solos do município de Santo Amaro/BA

Ca	Observados (g/kg)		Preditos (g/kg)	
	Máximo	Mínimo	Máximo	Mínimo
0-5 cm	84,47	0,03	37,54	0,54
5-20 cm	141,18	0,02	10,17	0,00
20-40 cm	132,22	0,03	57,48	0,10

Em relação à distribuição espacial do Ca, os mapas gerados e os resultados das estatísticas zonais permitem verificar que teores mais elevados do elemento foram mapeados em locais da paisagem onde predominam os Vertissolos, em regiões de altitude intermediaria a baixa, localizadas na porção centro-leste da área de estudo. Em Santo Amaro/BA os Vertissolos se formam a partir da alteração de rochas da Formação Candeias, que apresentam alta concentração de Ca devido a ocorrência de lentes carbonáticas, originadas pelo processo de precipitação de sais marinho durante a litogênese dos folhelhos (ASEVEDO, 2012). Os teores mais baixos para o Ca na área de estudo, foram mapeados nas áreas de maior altitude onde predominam Latossolos Amarelos e Argissolos Amarelos e, em regiões de baixa altitude onde predominam Neossolos Quartzarênicos. Jang et al., (2016) utilizaram o pXRF no mapeamento de teores de Ca em uma vinícola localizada na Austrália. Com os dados de concentração de Ca os autores puderam identificar solos originados a partir de rochas calcáreas que se distribuem em locais específicos na área, argumentando que o uso da fluorescência de raio-x no mapeamento de solos mostrou-se poderosa ferramenta para auxiliar na tomada de decisão das operações de manejo.

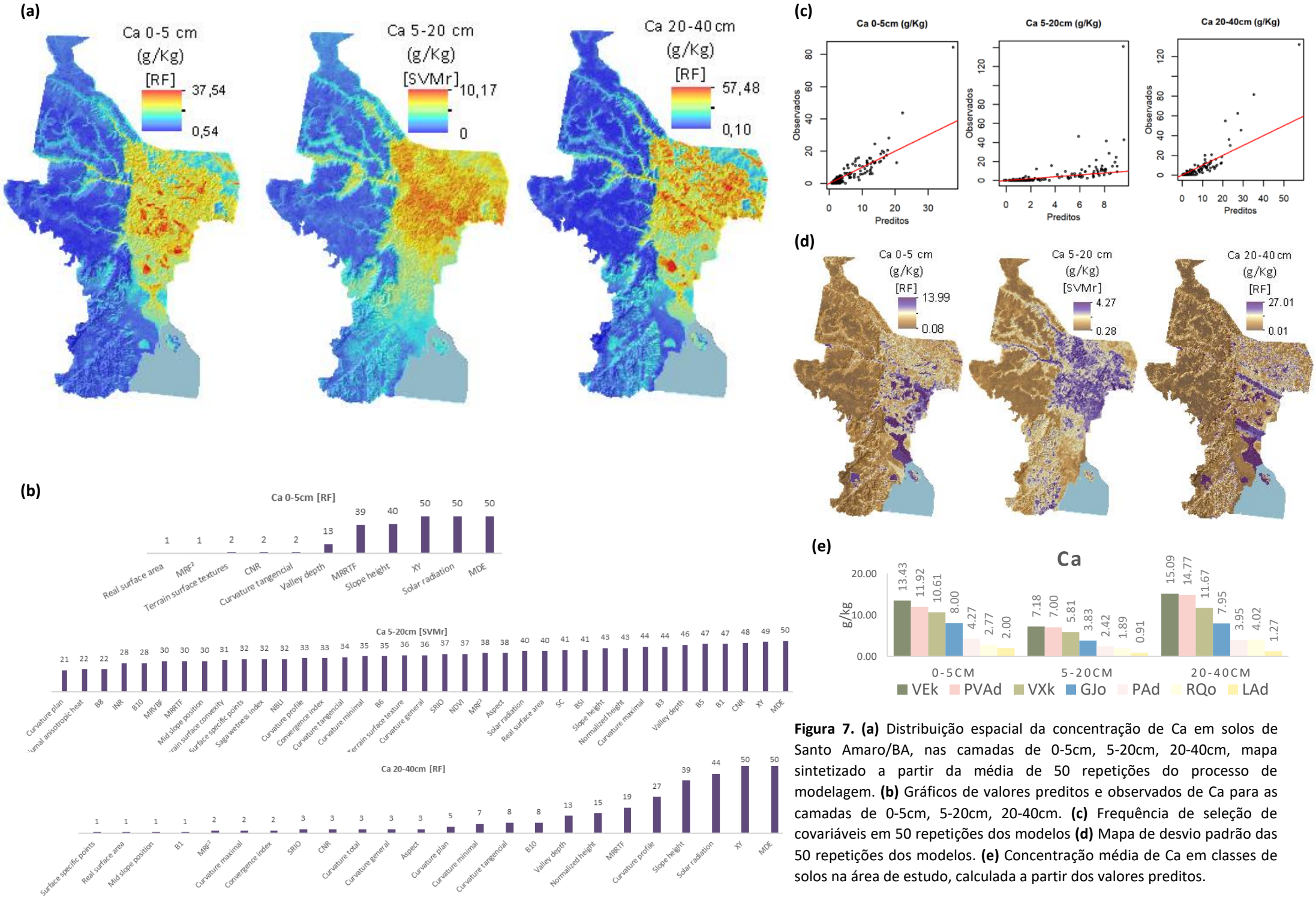


Figura 7. (a) Distribuição espacial da concentração de Ca em solos de Santo Amaro/BA, nas camadas de 0-5cm, 5-20cm, 20-40cm, mapa sintetizado a partir da média de 50 repetições do processo de modelagem. **(b)** Gráficos de valores preditos e observados de Ca para as camadas de 0-5cm, 5-20cm, 20-40cm. **(c)** Frequência de seleção de covariáveis em 50 repetições dos modelos **(d)** Mapa de desvio padrão das 50 repetições dos modelos. **(e)** Concentração média de Ca em classes de solos na área de estudo, calculada a partir dos valores preditos.

4.2.3 Ferro

A variabilidade espacial do teor de Fe na área de estudo está representada nos mapas da Figura 8(a). O ferro apresentou obteve bons ajustes no processo de modelagem, com valores de R^2 que variam entre 0.18 (kNN) e 0.59 (RF) na camada de 0-5cm, 0.16 (kNN) e 0.59 (RF) na camada de 5-20cm, 0.14 (kNN) e 0.54 (RF) na camada de 20-40cm. Além do RF, outros algoritmos que apresentaram bons ajustes na predição espacial de Fe no presente estudo foram: MARS (0.57) e CUB (0.56) na camada de 0-5cm, CUB (0.58) na camada de 5-20cm e CUB (0.52) na camada de 20-40cm. O kNN apresentou performance destoante e significativamente inferior aos outros algoritmos na predição de K. Em seu estudo, Campbell et al., (2019) obtiveram ajustes sensivelmente inferiores no mapeamento de Fe utilizando o pXRF, com valores de R^2 variando entre 0.40 (PLS) a 0.49 (RF) ma camada de 0-10cm, e 0.33 (GLMBOOST, PCR, PLS e RIDGE) a 0.43 (RF) na camada de 10-30cm.

A Figura 8(b) mostra o gráfico da frequência de seleção de covariáveis utilizadas para explicar a variação espacial de Fe na área de estudo. MDE, XY e Solar radiation foram as únicas covariáveis selecionadas em todas as repetições, nas três camadas. Além destas Slope height, Normalized height e Valley depth foram selecionadas com alta frequência nas repetições. A Figura 8(c) mostra o gráfico de valores preditos e observados para o teor de Fe, a Tabela 7 mostra os valores preditos e observados, máximos e mínimos.

Tabela 9. Valores observados e preditos, máximos e mínimos para Fe em solos do município de Santo Amaro/BA

Fe	Observados (g/kg)		Preditos (g/kg)	
	Máximo	Mínimo	Máximo	Mínimo
0-5 cm	150,82	1,34	95,97	7,69
5-20 cm	140,03	0,55	93,92	9,17
20-40 cm	143,47	0,82	92,27	9,83

Teores elevados de Fe foram mapeados nas regiões de predomínio dos Vertissolos, em regiões de baixa e média altitude localizadas no centro e leste do município. Os altos teores do elemento podem ser explicados tanto pelas argilas de alta atividade presentes nessa classe de solos, conferindo a estes alta capacidade de retenção de metais, quanto pela alta proporção de óxidos de Fe na composição da matriz desses solos (MONTE NERO, 2020). Nos mapas de Fe é possível notar manchas de elevada concentração do elemento que ocorrem nos pontos com as cotas altimétricas mais altas do município. Essas manchas localizam-se no extremo norte, norte, e oeste do município e assemelham-se a “ilhas”, já que ocorrem em locais de predomínio de baixos teores de Fe. Nesses locais estão localizadas elevações formadas por rochas metamórficas do Complexo Santa Luz. O complexo peridotítico Santa-Luz tem em sua estrutura uma associação de rochas máficas-ultramáficas, formadas por percolação de magma basáltico em peridotitos do manto. Em relação à geoquímica, parte das rochas que compõem o complexo são abundantes em Fe e Ti (OLIVEIRA et al., 2007). As classes de solo que ocorrem nos locais de afloramento das rochas do Complexo Santa-Luz diferenciam-se daquelas que estão delineadas no mapa pedológico do IBGE (2018), de acordo com perfis completos e observações de campo, nessas área formam-se Latossolos Vermelhos e Argissolos Vermelhos. Desse modo, com os modelos de distribuição da concentração de Fe em Santo Amaro será possível identificar e mapear as manchas das classes citadas, podendo ser utilizada também para ajustar o mapa litológico atual.

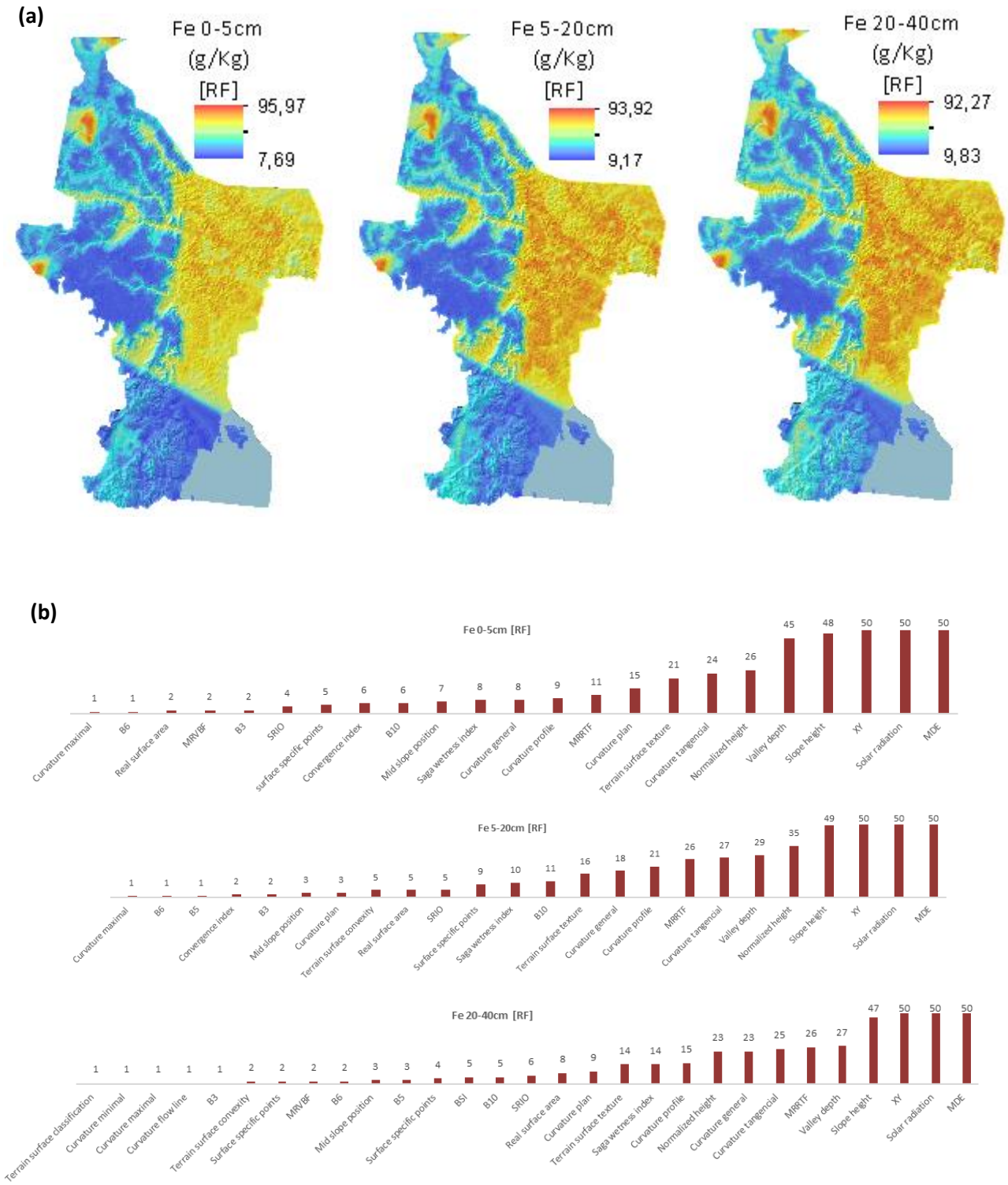


Figura 8. (a) Distribuição espacial da concentração de Fe em solos de Santo Amaro/BA , nas camadas de 0-5cm, 5-20cm, 20-40cm, mapa sintetizado a partir da média de 50 repetições do processo de modelagem. **(b)** Gráficos de valores preditos e observados para as camadas de 0-5cm, 5-20cm, 20-40cm. **(c)** Frequência de seleção de covariáveis em 50 repetições dos modelos. **(d)** Mapa de desvio padrão das 50 repetições dos modelos. **(e)** Concentração média de Fe em classes de solos na área de estudo, calculada a partir dos valores preditos.

4.2.4 Potássio

A variabilidade espacial do teor de K na área de estudo está representada nos mapas da figura 9(a). O potássio foi o elemento que apresentou o melhor ajuste no processo de modelagem, com valores de R^2 que variam entre 0.31 (kNN) e 0.83 (CUB e GBM) na camada de 0-5cm, 0.31 (kNN) e 0.83 (GBM) na camada de 5-20cm, 0.32 (kNN) e 0.82 (GBM) na camada de 20-40cm. Além do CUB e GBM, outros algoritmos que apresentaram bons ajustes na predição espacial de K no presente estudo foram: MARS (0.80) e RF (0.82) na camada de 0-5cm, RF (0.81) na camada de 5-20cm e RF (0.81) na camada de 20-40cm. O kNN apresentou performance destoante e significativamente inferior aos outros algoritmos na predição de K.

Os resultados observados por Campbell et al., (2019) em um estudo onde utilizaram o pXRF no mapeamento de K divergem dos encontrados no presente trabalho. Os autores reportaram valores de R^2 inferiores aos obtidos aqui, com resultados que variam entre 0.23 (PCR) a 0.41 (GBM) na camada de 0-10cm, e 0.32 (CUB) a 0.38 (RF).

A figura 9(b) mostra o gráfico da frequência de seleção de covariáveis utilizadas para explicar a variação espacial de K na área de estudo. MDE, XY foram as únicas covariáveis selecionadas em todas as repetições, nas três camadas. Além destas Slope height, Normalized height e Valley depth foram selecionadas com alta frequência nas repetições. A figura 9(c) mostra o gráfico de valores preditos e observados para o teor de K, a Tabela 8 mostra os valores preditos e observados, máximos e mínimos.

Tabela 10. Valores observados e preditos, máximos e mínimos para K em solos do município de Santo Amaro/BA

K	Observados (g/kg)		Preditos (g/kg)	
	Máximo	Mínimo	Máximo	Mínimo
0-5 cm	31,30	0,27	29,61	0,29
5-20 cm	31,62	0,04	26,98	0,00
20-40 cm	27,36	0,05	22,97	0,00

Em relação à distribuição espacial de K na área, os maiores teores foram mapeados em locais de predomínio de Vertissolos, teores intermediários foram mapeados em áreas de ocorrência de Neossolos, e os teores mais baixos estão localizados nas partes mais altas do município onde predominam Latossolos e Argissolos. Os Vertissolos que ocorrem em Santo Amaro têm uma grande proporção de Ilita, um mineral de argila rico em K, o que explica os teores elevados do elemento encontrados nesses solos (ASEVEDO et al., 2012), além disso, a alta CTC dos Vertissolos conferem a estes elevada capacidade de retenção de cátions básicos. Outro fator que explica a alta concentração de cátions básicos nos Vertissolos em Santo Amaro é a baixa altitude do relevo nos locais de sua ocorrência, o que dificulta o processo de lixiviação da porção trocável. As Coberturas-Detrito Lateríticas são formações terrígenas de material não consolidado, com litologia composta por areias, argilas, cascalhos e cangas (SANTOS, 2015), material pobre em cátions básicos. Em Santo Amaro o produto do intemperismo desse material são os Latossolos Amarelos e Argissolos Amarelos, que ocorre nos topos dos tabuleiros, locais onde foram mapeados os menores teores de K.

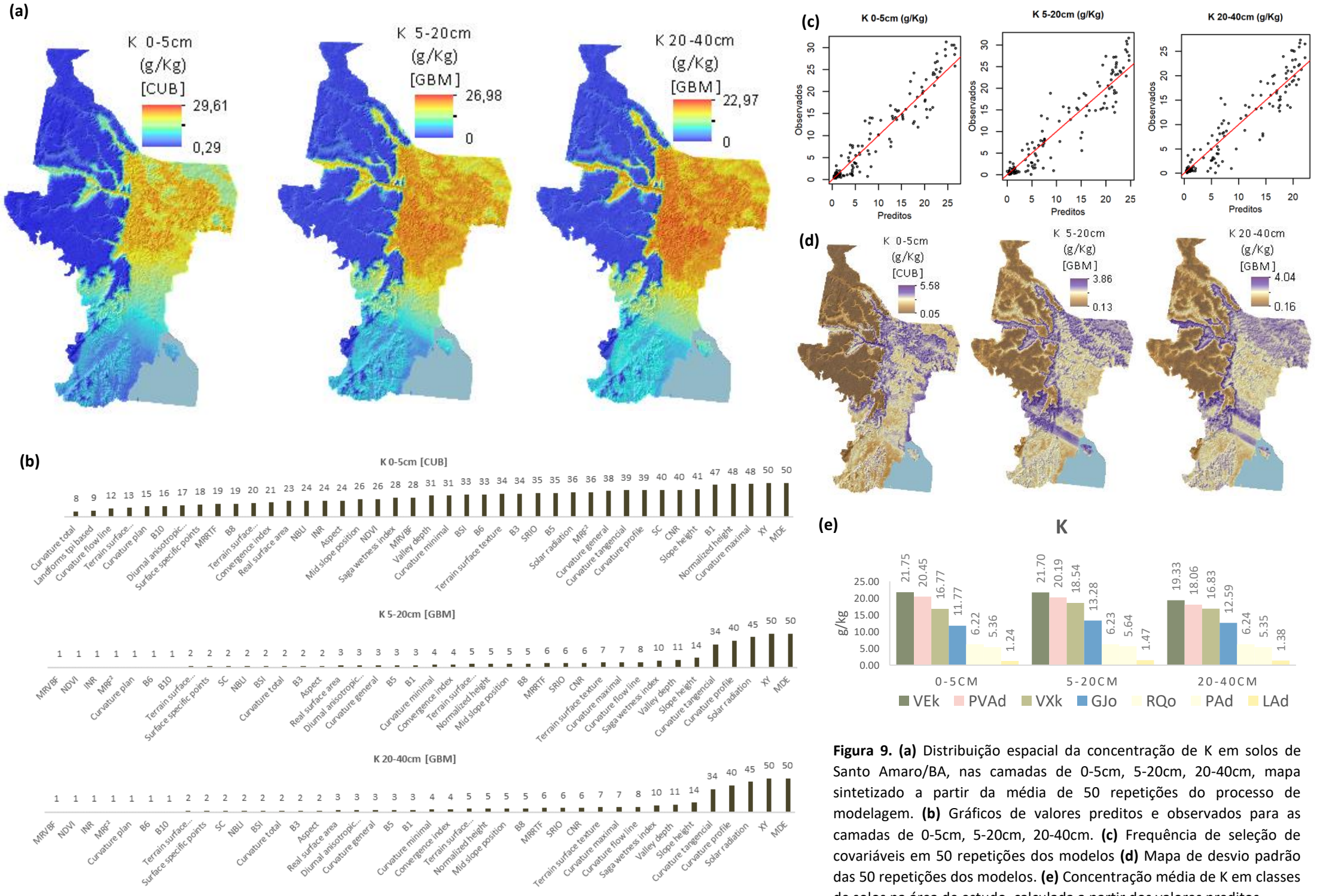


Figura 9. (a) Distribuição espacial da concentração de K em solos de Santo Amaro/BA, nas camadas de 0-5cm, 5-20cm, 20-40cm, mapa sintetizado a partir da média de 50 repetições do processo de modelagem. **(b)** Gráficos de valores preditos e observados para as camadas de 0-5cm, 5-20cm, 20-40cm. **(c)** Frequência de seleção de covariáveis em 50 repetições dos modelos **(d)** Mapa de desvio padrão das 50 repetições dos modelos. **(e)** Concentração média de K em classes de solos na área de estudo, calculada a partir dos valores preditos.

4.2.5 Magnésio

A variabilidade espacial do teor de Mg na área de estudo está representada nos mapas da figura 10(a). O magnésio foi o elemento que apresentou o segundo melhor ajuste, ficando atrás apenas do K. Os valores de R^2 para o elemento variaram entre 0.23 (kNN) e 0.75 (RF) na camada de 0-5cm, 0.24 (kNN) e 0.71 (RF e GBM) na camada de 5-20cm, 0.33 (kNN) e 0.73 (ANN,GBM, CUB, MARS e RF) na camada de 20-40cm. Outros algoritmos que apresentaram bons ajustes na predição espacial de K no presente estudo foram: GBM (0.73), CUB (0.71) e ANN (0.70) na camada de 0-5cm e GBM (0.71) na camada de 5-20cm. O kNN apresentou performance destoante e significativamente inferior aos outros algoritmos na predição de Mg.

Chama atenção na camada 20-40cm o fato de que o com o ANN foi o algoritmo que apresentou o menor MAE (1.69), porém, RF e MARS apresentaram o menor valor de RMSE, mostrando assincronia entre os valores das métricas para o elemento e camada em questão. Como, em termos de unidade de medida, o valor do MAE obtido pelo ANN foi sensivelmente inferior ao dos outros algoritmos, o modelo gerado pelo método em questão foi utilizado no mapa final de Mg para a camada de 20-40cm. Apesar disto é possível observar que o mapa gerado apresenta generalizações significativas na distribuição espacial do elemento, não representando os gradientes de transição entre valores extremos, tampouco indicando a ocorrência de valores intermediários. Além disso, é possível observar uma maior suavização nos valores máximos e mínimos do que as verificadas nos mapas das outras camadas, que foram gerados pelo RF.

A figura 10(b) mostra o gráfico da frequência de seleção de covariáveis utilizadas para explicar a variação espacial de Mg. MDE e XY foram novamente as únicas covariáveis selecionadas em todas as repetições, nas três camadas. Outras covariáveis com alta frequência de seleção foram Normalized height, Solar radiation, Slope height e Valley depth. A figura 10(c) mostra o gráfico de valores preditos e observados para o teor de Mg, a Tabela 9 mostra os valores preditos e observados, máximos e mínimos.

Tabela 11. Valores observados e preditos, máximos e mínimos para Mg em solos do município de Santo Amaro/BA

Mg	Observados (g/kg)		Preditos (g/kg)	
	Máximo	Mínimo	Máximo	Mínimo
0-5 cm	24,05	3,20	21,03	4,58
5-20 cm	27,67	3,00	21,95	4,76
20-40 cm	26,14	4,10	16,93	6,40

Em relação à distribuição espacial dos teores de Mg, os maiores teores foram mapeados em área de baixa a média altitude onde predominam Vertissolos, teores intermediários ocorrem principalmente na zona de transição entre o domínio dos Vertissolos e dos Neossolos, localizada ao Sul. Os menores teores de Mg foram mapeados nas partes mais altas da paisagem, no topo dos tabuleiros onde predominam Latossolos e Argissolos. Os Vertissolos em Santo Amaro formam-se a partir do intemperismo das rochas da formação Candeias, principalmente de folhelhos esverdeados que possuem alta concentração de Mg, devido a dolomita presente em sua composição (ASEVEDO, 2012), o que explica os altos teores do elemento mapeados nas áreas de domínio desses solos. Os teores de Mg mais baixos mapeados na área encontram-se nos domínios de solos com caráter caulinítico, Latossolos e Argissolos, que possuem baixa saturação por bases.

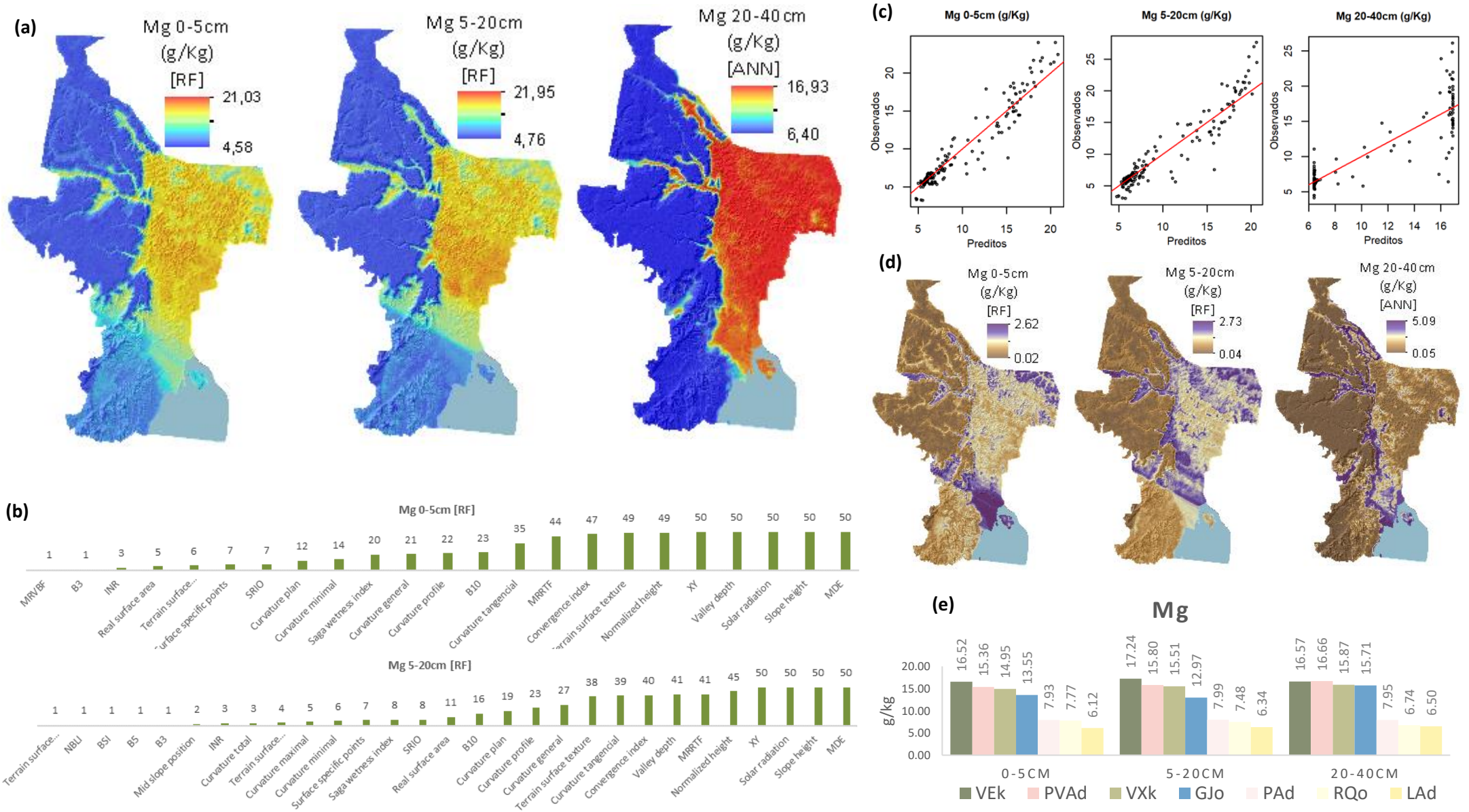


Figura 10. (a) Distribuição espacial da concentração de Mg em solos de Santo Amaro/BA, nas camadas de 0-5cm, 5-20cm, 20-40cm, mapa sintetizado a partir da média de 50 repetições do processo de modelagem. **(b)** Gráficos de valores preditos e observados para as camadas de 0-5cm, 5-20cm, 20-40cm. **(c)** Frequência de seleção de covariáveis em 50 repetições dos modelos **(d)** Mapa de desvio padrão das 50 repetições dos modelos. **(e)** Concentração média de Mg em classes de solos na área de estudo, calculada a partir dos valores preditos.

4.2.6 Silício

A distribuição espacial de Si na área de estudo está representada nos mapas da Figura 11(a). O silício foi um dos elementos com pior ajuste dentre os mapeados nesse trabalho, com valores de R^2 que variaram entre 0.26 (MARS) a 0.35 (kNN) na camada de 0-5cm, 0.27 (SVMr) a 0.35 (kNN) na camada de 5-20cm e 0.31 (ANN) a 0.42 (RF) na camada de 20-40cm. Dentre os algoritmos que apresentaram boa performance para o elemento podem-se citar, RF (0.34) na camada de 0-5cm, RF (0.32) e GBM (0.31) na camada de 5-20cm, kNN (0.39), GBM (0.38) e SVMr (0.37) na camada de 20-40cm.

Apesar de apresentar boas performances na predição de Si, sendo inclusive o algoritmo com melhores resultados nas duas primeiras camadas para o elemento, o kNN gerou mapas com baixa qualidade espacial, com delineamentos artificiais e efeito dominante da variável XY, não representando a geometria fractal característica das feições naturais. Por esse motivo os mapas finais de Si foram gerados a partir dos algoritmos que apresentaram melhores performances após o kNN, correspondendo ao RF nas camadas de 0-5cm e 5-20cm. Em seu trabalho de mapeamento de classes de solos Heung et al., 2016 observaram resultados semelhantes para o kNN, com o algoritmo apresentando performance superior a outros métodos, porém, gerando mapas "pontilhados", de difícil interpretação. O autor atribui o efeito observado em seu trabalho à ocorrência de "overffiting" durante o processo de modelagem com o kNN. A Figura 11(b) mostra o gráfico da frequência de seleção de covariáveis utilizadas para explicar a variação espacial de Si. MDE e XY foram as únicas variáveis selecionadas em todas as repetições nas três camadas.A Figura 11(c) mostra o gráfico de valores preditos e observados para o teor de Si, a Tabela 10 mostra os valores preditos e observados, máximos e mínimos.

Tabela 12. Valores observados e preditos, máximos e mínimos para Si em solos do município de Santo Amaro/BA

Si	Observados (g/kg)		Preditos (g/kg)	
	Máximo	Mínimo	Máximo	Mínimo
0-5 cm	1.016,12	376,39	860,87	450,14
5-20 cm	1.025,88	354,16	854,16	434,55
20-40 cm	1025,08	328,53	813,80	391,45

Nos mapas de Si gerados neste estudo foi possível identificar manchas com elevados teores do elemento localizadas no extremo sul da área de estudo, concordando com os locais de ocorrência de Neossolos formados por sedimentos arenosos do quaternário. Ao sul da área predomina a unidade litológica dos Depósitos Flúvio-Lagunares, formada por sedimentos costeiros do quaternário ricos em quartzo (SiO2), material de origem de Neossolos Flúvicos e Quartzarênicos (PASSE, 2015; GLOAGUEN & PASSE, 2017). Gloaguen & Passe (2017) encontraram baixas concentrações de metais nos Neossolos formados por sedimentos do Quaternário. Nos tabuleiros localizados ao norte, locais de predomínio de Argissolos, foram mapeados teores sensivelmente maiores que aqueles encontrados nos tabuleiros centrais, locais de predomínio de Latossolos e, onde distribuem-se os menores teores de Si em toda a área. A diferença nos teores de SiO_2 entre as camadas superiores de Argissolos e Latossolos deriva de concentrações sensivelmente maiores de areia no horizonte superficial dos primeiros, uma das características que diferencia essas classes (PEDRON et al.,2012).

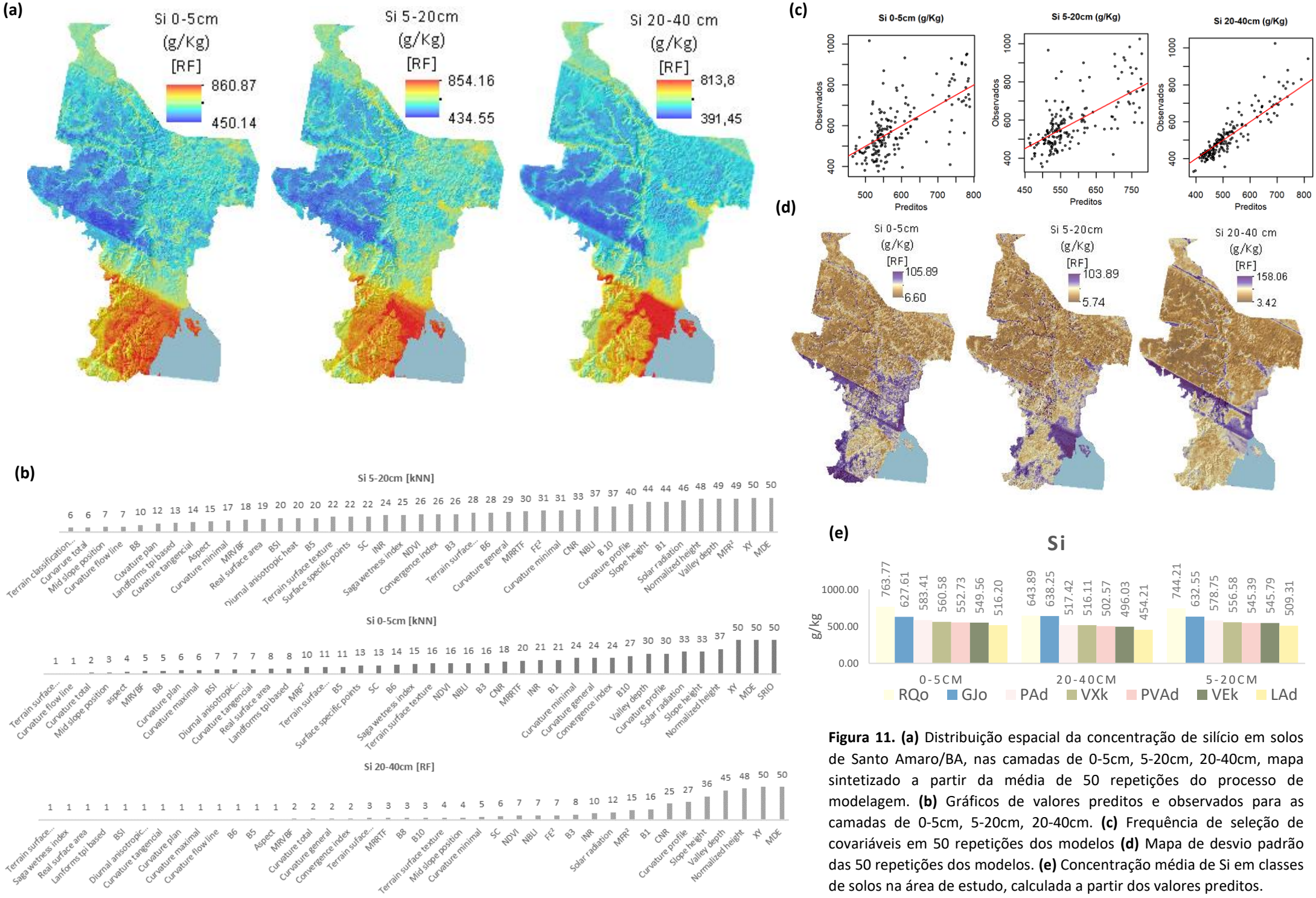


Figura 11. (a) Distribuição espacial da concentração de silício em solos de Santo Amaro/BA, nas camadas de 0-5cm, 5-20cm, 20-40cm, mapa sintetizado a partir da média de 50 repetições do processo de modelagem. **(b)** Gráficos de valores preditos e observados para as camadas de 0-5cm, 5-20cm, 20-40cm. **(c)** Frequência de seleção de covariáveis em 50 repetições dos modelos **(d)** Mapa de desvio padrão das 50 repetições dos modelos. **(e)** Concentração média de Si em classes de solos na área de estudo, calculada a partir dos valores preditos.

4.2.7 Titânio

A distribuição espacial de Ti na área de estudo está representada nos mapas da Figura 12(a). O titânio foi um dos elementos com melhor ajuste dentre os mapeados nesse trabalho, com valores de R^2 que variaram entre 0.48 (kNN) a 0.60 (RF) na camada de 0-5cm, 0.45 (kNN) a 0.57 (GBM) na camada de 5-20cm e 0.42 (kNN) a 0.60 (RF e GBM) na camada de 20-40cm. Dentre os algoritmos que apresentaram boa performance para o elemento podem-se citar, GBM e MARS (0.56) na camada de 0-5cm, RF (0.54) na camada de 5-20cm e CUB (0.59) na camada de 20-40cm. Campbell et al., (2019) obtiveram ajustes similares ao nossos para o Ti, com valores de R^2 um pouco mais baixos, variando entre 0.48 (GLMBOOST e RF) a 0.50 (CUBIST, PLS, PCR, FOBA e RIDGE) na camada de 0-10cm e 0.42 (RIDGE) a 0.56 (GBM) na camada de 10-30cm.

A Figura 12(b) mostra o gráfico da frequência de seleção de covariáveis utilizadas para explicar a variação espacial de Si. MDE e Solar radiation foram as únicas covariáveis selecionadas em todas as repetições nas três camadas, XY não foi utilizada em apenas duas repetições na camada de 20-40cm. Dentre as variáveis oriundas de sensoriamento remoto, a banda 1 do Landsat 8 apresentou alta frequência de seleção nas três camadas. No caso das variáveis derivadas do MDE, Slope height e Normalized height foram frequentemente selecionadas, como ocorreu em outros elementos, a diferença para o elemento Ti está na alta frequência de seleção para as Curvaturas total e de perfil, que ocorreu nas três camadas estudadas. A Figura 12(c) mostra o gráfico de valores preditos e observados para o teor de Ti, a Tabela 11 mostra os valores preditos e observados, máximos e mínimos.

Tabela 13. Valores observados e preditos, máximos e mínimos para Ti em solos do município de Santo Amaro/BA

Ti	Observados (g/kg)		Preditos (g/kg)	
	Máximo	Mínimo	Máximo	Mínimo
0-5 cm	20,09	0,59	16,18	2,24
5-20 cm	18,11	0,83	13,31	2,38
20-40 cm	18,39	0,88	15,38	2,88

Os mapas de teores de Ti apontam a ocorrência de valores mais baixos do elemento nas zonas de menor altitude, principalmente nas áreas localizadas ao Sul do município, onde predominam Neossolos Quartzarênicos e Gleissolos Tiomórficos. Os teores mais altos do elemento foram mapeados nas regiões mais altas do município, nos tabuleiros onde predominam Latossolos Amarelos e Argissolos Amarelos. No extremo norte do município foi mapeada uma mancha com os teores mais elevados do elemento, local onde, de acordo com observações de campo, ocorrem Latossolos Vermelhos originados a partir da alteração das rochas metamórficas do Complexo Santa Luz. Os minerais de Ti são muito resistentes ao intemperismo, conhecidos pela alta estabilidade em solos, ocorrendo nestes na forma dos minerais secundários Rutilo, Anatase e Brookita. Teores altos de Ti em solos estão normalmente associados a intemperismo intenso, localização em regiões tropicais ou formação a partir de materiais de origem ricos em Ti (KABATA-PENDIAS, 2000). Latossolos e Argissolos são solos que passaram por longo e intenso processo de intemperismo, fato que explica os altos teores do elemento mapeados nos domínios dessa classe em Santo Amaro. Além disso, as manchas de maior concentração, localizadas no extremo norte, estão ligadas à ocorrência de material de origem rico em Ti do Complexo Santa-Luz. De acordo com Oliveira et al., 2007, as rochas máficas e ultramáficas que compõe o Complexo Santa-Luz são abundantes em óxidos de Ti.

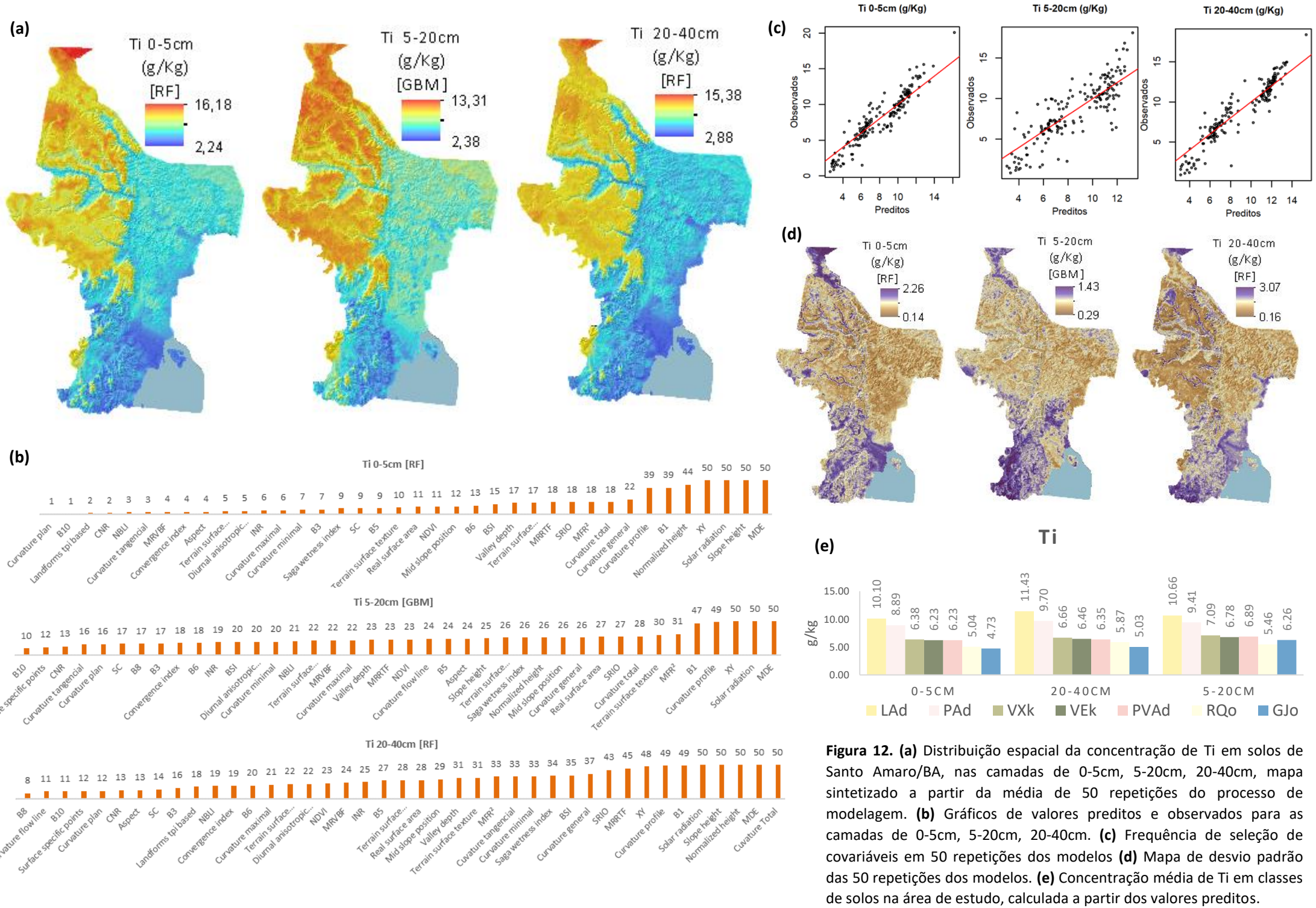
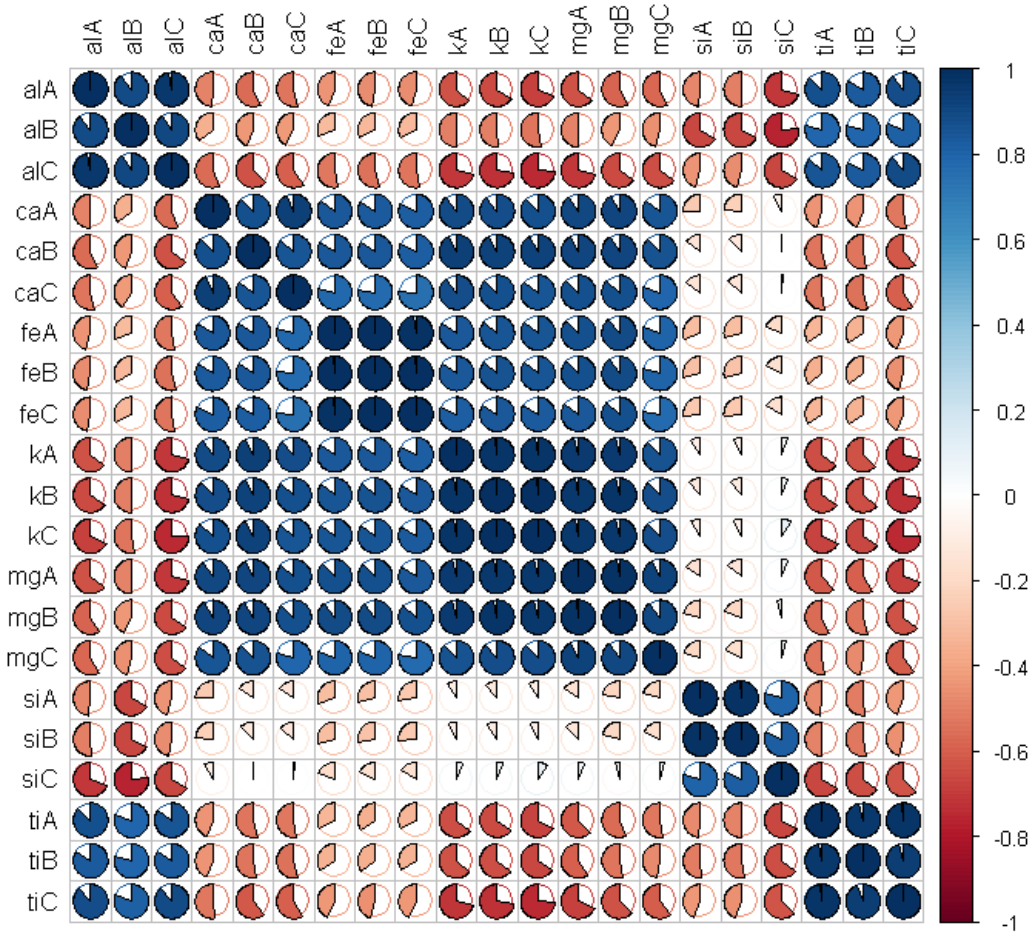


Figura 12. (a) Distribuição espacial da concentração de Ti em solos de Santo Amaro/BA, nas camadas de 0-5cm, 5-20cm, 20-40cm, mapa sintetizado a partir da média de 50 repetições do processo de modelagem. **(b)** Gráficos de valores preditos e observados para as camadas de 0-5cm, 5-20cm, 20-40cm. **(c)** Frequência de seleção de covariáveis em 50 repetições dos modelos **(d)** Mapa de desvio padrão das 50 repetições dos modelos. **(e)** Concentração média de Ti em classes de solos na área de estudo, calculada a partir dos valores preditos.

4.3 Correlação espacial entre elementos e classes de solos

A figura 13 apresenta o gráfico de correlação de Pearson para os mapas dos elementos gerados. O coeficiente de correlação de Pearson mede a força de associação linear entre duas variáveis, utilizando teste não-paramétrico para calcular a significância. O coeficiente é medido em escala adimensional, com valores que variam de -1 a +1. Valores iguais a 0 indicam inexistência de correlação entre as variáveis, valores menores que 0 indicam ocorrência de correlação negativa, e para valores maiores que 0 a correlação é positiva (SEDGWICK, 2012).

Figura 13. Gráfico de correlação de Pearson para os mapas dos teores de elementos em solos, nas camadas de 0-5cm (A), 5-20cm (B) e 20-40cm (C) em Santo Amaro/BA.



As Figuras 14 , 15 e 16 apresentam os resultados da análise de componentes principais PCA entre os mapas de elementos (variáveis) e o mapa pedológico (indivíduos), para as três camadas estudadas. Quando aplicada à ciência do solo a PCA permite identificar a existência de correlação entre variáveis (setas) e agrupar indivíduos (pontos) pelo seu grau de similaridade com essas variáveis, facilitando o entendimento de interações entre variáveis e indivíduos e demonstrando o potencial das informações para explicar variações no solo (OLIVEIRA et al., 2015). As setas representam as variáveis (elementos), e os pontos os indivíduos (classes de solos). Setas próximas e na mesma direção indicam correlação positiva entre as variáveis, setas em direção opostas indicam correlação negativa entre as variáveis, setas ortogonais indicam ausência de correlação entre as variáveis. Indivíduos que se

agrupam na direção de determinada seta apresentam alta correlação com a variável representada por essa seta (KASSAMBARA, 2017).

A soma dos componentes 1 e 2 explica 93.1% da variabilidade dos dados na camada de 0-5cm, 89,8% na camada de 5-20cm e 90,2% na camada de 20-40cm. K e Mg foram os elementos com maior importância no componente 1 em todas as camadas, seguido por Ca, Fe, Al e Ti. No componente 2 o Si foi elemento com maior importância nas três camadas, seguido por Al e Ti.

A análise do gráfico de variáveis da PCA em conjunto com o gráfico de Pearson, permite concluir que há alta correlação entre os elementos K, Mg, Ca e Fe, demonstrando uma relação diretamente proporcional na distribuição espacial dos teores desses elementos. No gráfico biplot da PCA é possível observar um agrupamento de pontos referentes à classe do Vertissolos na direção das setas desses elementos, indicando uma estreita relação entre a ocorrência desses solos e a distribuição espacial de K, Mg, Ca e Fe. Vertissolos normalmente desenvolvem-se em ambientes de bacias sedimentares, a partir de material com granulometria fina e com altos teores de cátions básicos, ou diretamente de rochas básicas com altos de teores de Ca e Mg. Em relação ao relevo, distribuem-se em área planas ou suave onduladas, mais raramente em áreas mais movimentadas (MOUSTAKAS, 2012; EMBRAPA, 2018). Em Santo Amaro os Vertissolos predominam em cotas de média a baixa altitude e relevo suave ondulado que se estendem do centro ao leste do município. O material de origem desses solos são as rochas sedimentares da formação Candeias, composta por folhelhos, lamitos, siltitos e lâminas de arenitos (BRASIL, 1981; CAIXETA 1994). Essas rochas são ricas em Ca e Mg, devido a ocorrência de lentes carbonáticas e dolomita, e, em K, devido à presença das Ilitas, argilominerais secundários com alta capacidade de fixação de potássio (AZEVEDO, 2012). Os altos teores de Fe em Vertissolos podem ser explicados tanto pelas argilas de alta atividade presentes nestes solos, conferindo a estes alta capacidade de retenção de metais, quanto pela alta proporção de óxidos de Fe na composição da matriz desses solos (MONTE NERO, 2020).

De acordo com o gráfico de correlação, Ti e Al apresentaram forte correlação positiva entre si, e negativa com todos os outros elementos. A análise PCA confirma essa informação e aponta existência de correlação entre esses elementos, com as respectivas setas convergindo na mesma direção. No gráfico biplot há agrupamento de pontos das classes dos Argissolos Amarelos e Latossolos Amarelos na direção das setas do Ti e Al, indicando relação entre a ocorrência desses solos e a distribuição espacial desses elementos. Argissolos e Latossolos são solos bastante intemperizados que possuem baixa saturação por bases e alta concentração de argilominerais de alumínio, principalmente gibbsita e caulinita (EMBRAPA, 2018). Solos intensamente intemperizados localizados em regiões tropicais apresentam teores elevados de Ti, explicados pela resistência dos minerais de Ti ao intemperismo e pela alta estabilidade destes nos solos (KABATA-PENDIAS, 2000). Em Santo Amaro esses solos distribuem-se nos tabuleiros planos e são produtos do intemperismo dos materiais das Coberturas Detrito-Lateríticas e do Grupo Brotas. As Coberturas Detrito-Lateríticas são formadas por sedimentos não consolidadados, com litologia marcada pela presença de areias, argilas, cascalhos e cangas (BARBOSA & DOMINGUEZ, 1996). O Grupo Brotas divide-se nas formações: Aliança, na base, com litologia marcada pela presença de arenitos variegados e folhelhos vermelhos e; Sergi, no topo, com presença de arenitos quartzosos e conglomerados finos a médios (BRASIL, 1981).

O Si apresentou correlação negativa com quase todos os elementos, com exceção do K na camada de 20-40cm onde os elementos apresentaram correlação positiva muito fraca. Os maiores níveis de correlação negativa para o Si foram observados para Al e Ti. Os resultados obtidos na PCA apontam o Si como elemento praticamente isolado no segundo componente, não apresentando correlação com nenhuma das outras variáveis. Houve um agrupamento de pontos de Neossolos Quartzarênicos na direção da seta que representa o Si, indicando uma estreita relação entre a ocorrência dessa classe de solos e a distribuição espacial do elemento. Por meio dos mapas gerados foi possível identificar manchas com elevados teores de Si localizadas no extremo sul da área de estudo, concordando com locais de ocorrência de Neossolos Quartzarênicos. A unidade litológica dos Depósitos Flúvio-Lagunares é formada por sedimentos costeiros do quaternário ricos em quartzo (SiO2), material de origem de Neossolos Flúvicos e Quartzarênicos (PASSE, 2015; GLOAGUEN & PASSE, 2017). Gloaguen & Passe (2017) encontraram baixas concentrações de metais nos Neossolos formados por sedimentos do Quaternário. A quase ausência de argila, e a CTC baixa são fatores que explicam baixos teores de metais nesses solos.

Não foi possível obter informações conclusivas nas análises PCA a respeito das classes dos Gleissolos Tiomórficos Órticos e dos Argissolos Vermelho-Amarelo Distróficos, devido principalmente à falta ou baixa ocorrência de pontos amostrais nas áreas de predomínio desses solos. Porém, uma observação visual dos mapas, reforçadas pelos resultados das estatísticas zonais, permite observar semelhanças no delineamento das manchas de baixos teores de Ti e Al com o delineamento dos domínios dos Gleissolos Tiomórficos Órticos em Santo Amaro/BA.

Figura 14. Resultados da análise de componentes principais dos mapas de teores de elementos em solos para a camada de 0-5cm, e do mapa pedológico IBGE (2018) em Santo Amaro/Ba. (a) Percentual de variância explicada por dimensão. (b) Percentual de importância das variáveis por dimensão. (c) Gráfico de dispersão dos indivíduos (classes de solo) e variáveis (elementos).

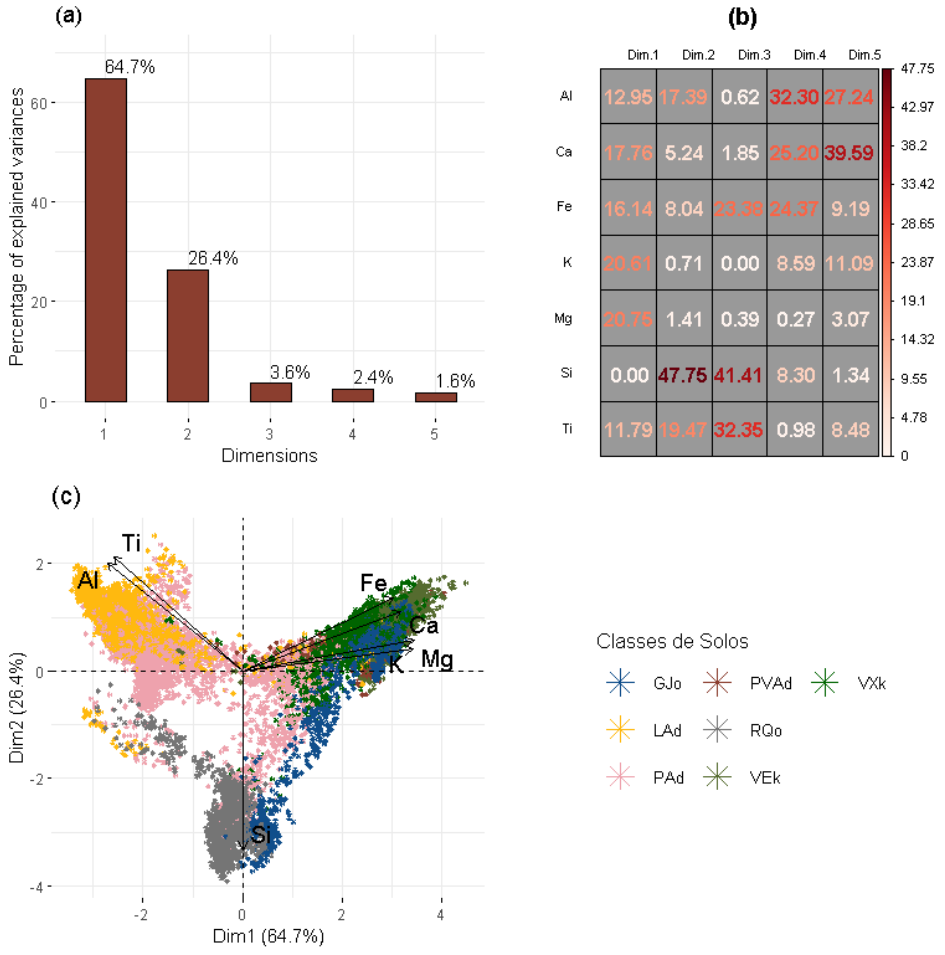


Figura 15. Resultados da análise de componentes principais dos mapas de teores de elementos em solos para a camada de 5-20cm, e do mapa pedológico IBGE (2018) em Santo Amaro/Ba. (a) Percentual de variância explicada por dimensão. (b) Percentual de importância das variáveis por dimensão. (c) Gráfico de dispersão dos indivíduos (classes de solo) e variáveis (elementos).

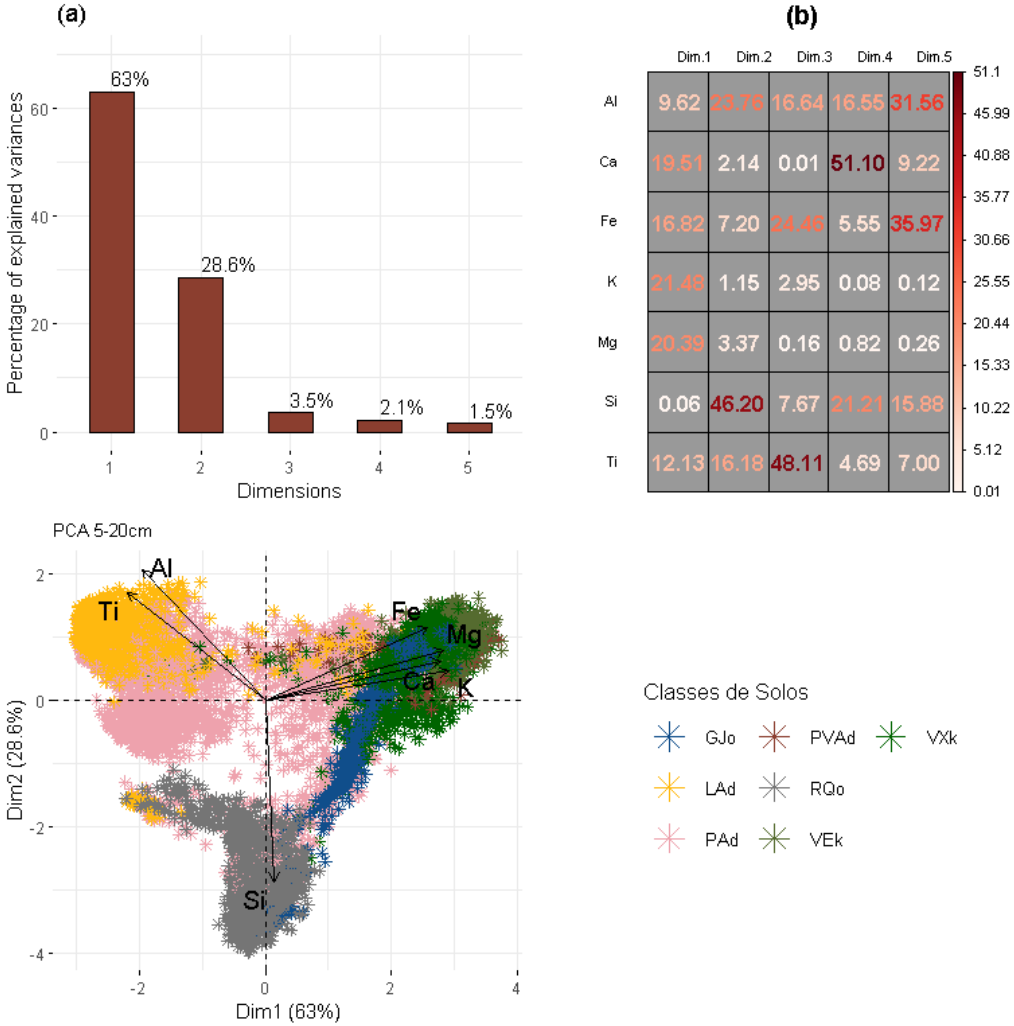
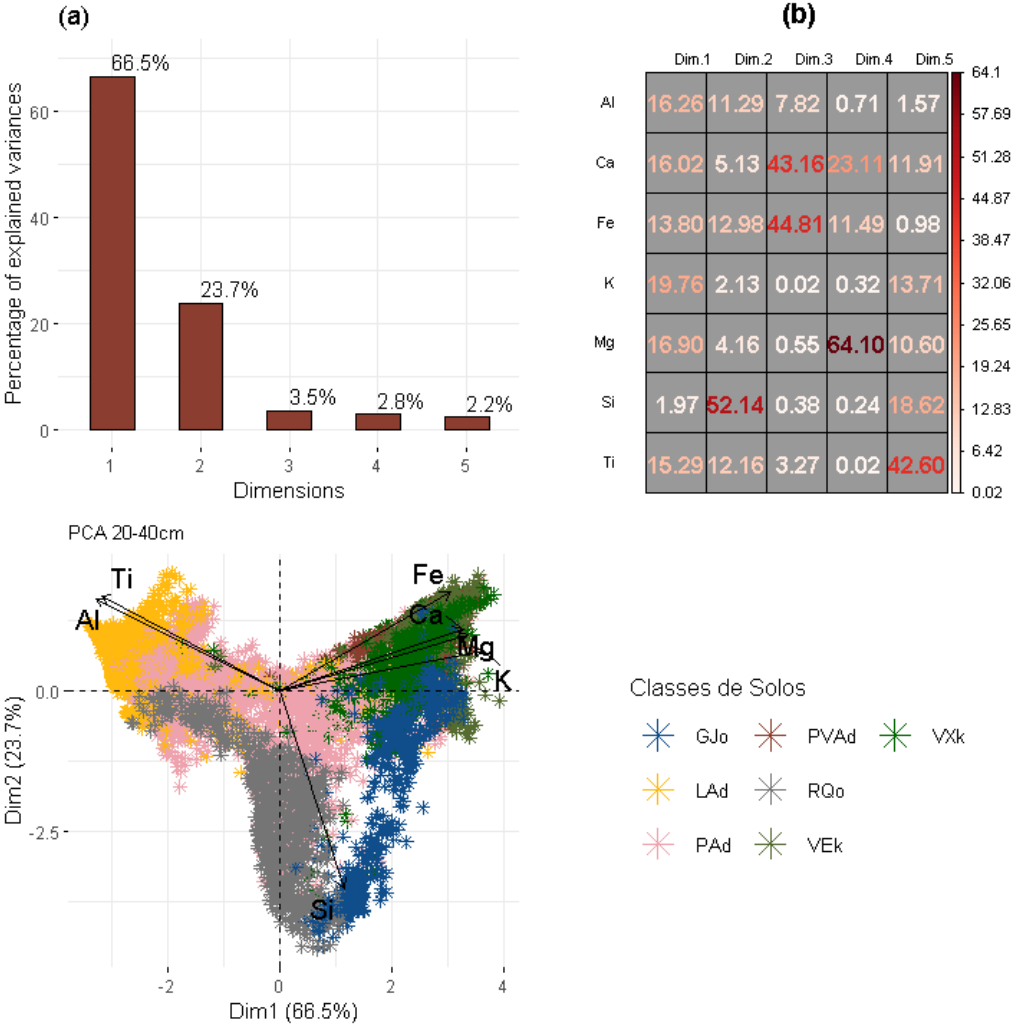


Figura 16. Resultados da análise de componentes principais dos mapas de teores de elementos em solos para a camada de 5-20cm, e do mapa pedológico IBGE (2018) em Santo Amaro/Ba. (a) Percentual de variância explicada por dimensão. (b) Percentual de importância das variáveis por dimensão. (c) Gráfico de dispersão dos indivíduos (classes de solo) e variáveis (elementos).



5.CONCLUSÃO

No presente trabalho objetivamos estudar e mapear elementos relacionados à composição geoquímica e variabilidade espacial dos diferentes materiais de origem dos solos que predominam no município de Santo Amaro/BA. A área estudada apresenta elevada complexidade na cobertura litológica e consequentemente na distribuição espacial dos solos, o que dificulta o mapeamento de classes de solos por métodos convencionais.

Entre os algoritmos avaliados RF, GBM, SVMr, CUB e MARS apresentaram os melhores resultados em termos de ajustes e representação na distribuição espacial dos teores da maioria dos elementos estudados, com destaque para o RF que apresentou constância e performance superior dentre todos os algoritmos testados. ANN e, principalmente kNN, apresentaram ajustes significativamente inferiores aos outros algoritmos na maioria dos elementos mapeados, mesmo assim, em alguns casos, esses algoritmos conseguiram bons ajustes em termos de métrica, porém geravam mapas com baixa representatividade espacial do fenômeno estudado.

Covariáveis representantes dos fatores relevo e posição espacial foram altamente relevantes para explicar a variação espacial nos teores de elementos estudados no presente trabalho. Devido ao padrão litoestratigráfico que ocorre em Santo Amaro é possível correlacionar o relevo com a distribuição dos diferentes materiais de origem que ocorrem na área, esse fato explica a importância da altitude como fator para explicar as variações na distribuição dos teores de elementos que aí ocorrem. Em relação a importância da covariável espacial XY, ocorre na área uma falha sentido longitudinal que separa litologia que ocorrem nas partes mais altas, predominantemente sedimentos não consolidados e arenitos pobres em cátions básicos, das litologias onde predominam argilitos e folhelhos com alta concentração de cátions básicos. Além disso, através dos dados espaciais foi possível rastrear ocorrência locais e pontuais de materiais de origem com composição geoquímica características, como nos casos dos depósitos arenosos localizados ao sul do município, ricos em Si, e dos pontos de afloramento do material rico em Fe de rochas metamórficas máficas e ultramáficas.

Em Santo Amaro a distribuição espacial dos teores mais elevados de K, Mg, Ca e Fe está relacionada as rochas da formação Candeias, e a ocorrência dos Vertissolos, produto do intemperismo do material de origem supracitado. Nas áreas onde a litologia predominante são as Coberturas Detrito Lateríticas, cujo produto de intemperismo são os Argissolos e Latossolos Amarelos, predominam teores mais elevados de Ti e Al, e os menores teores de cátions básicos. Os modelos espaciais de Si apresentam alta relação com a distribuição de Neossolos Quartzarênicos, essa classe de solos predomina nos locais onde foram mapeados os maiores teores de Si na área de estudo. A falta de amostras em regiões de predomínio de Gleissolos Tiomórficos, e baixa quantidade de amostras nas áreas de predomínio dos Argissolos Vermelho-Amarelo, impossibilitaram a determinação de padrões entre essas classes e a distribuição dos elementos.

A associação entre técnicas de machine learning e a análise XRF foi eficiente na identificação diferenças na geoquímica em Santo Amaro/BA, visto que a área apresenta alta complexidade na variação espacial das coberturas litológicas e pedológica. Desse modo, a integração dessas tecnologias apresenta-se como uma ferramenta eficaz para estudar a variação espacial de atributos e classes de solos, principalmente em locais onde há alta variabilidade espacial dos atributos geoquímicas e consequente complexidade na distribuição espacial dos solos.

6.REFERÊNCIAS

ADLER, Karl. Digital soil mapping and portable X-ray fluorescence prediction of cadmium, copper and zinc concentrations as decision support for crop production. 2022. (Tese Doutorado)

ALVARES, C.A; Stape, J.L; Sentelhas, P.C; de Gonçalves, M; Leonardo, J; Gerd, S; Köppen's Climate Classification Map for Brazil. (em inglês). **Meteorologische Zeitschrift**, 2013. 711–728.

APPELHANS, Tim et al. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. **Spatial Statistics**, v. 14, p. 91-113, 2015.

ARAUJO-CARRILLO, Gustavo A. et al. IRAKA: The first Colombian soil information system with digital soil mapping products. **Catena**, v. 196, p. 104940, 2021.

ARNOLDUSSEN, Stijn; VAN OS, B. J. H. The potential of lacquer-peel soil profiles for palaeo-geochemical analysis using XRF analysis. **Catena**, v. 128, p. 16-30, 2015.

ASSAMI, Tarek; HAMDI-AÏSSA, Baelhadj. Digital mapping of soil classes in Algeria–A comparison of methods. **Geoderma Regional**, v. 16, p. e00215, 2019.

AZEVEDO, Ladyanne Pinheiro. Mapeamento geoquimico de solos contaminados por metais (pb, zn, as e cu), Santo Amaro da Purificação, Bahia. 2013. (Dissertação de Mestrado).

BARBOSA, J. S. F.; DOMINGUEZ, J. M. L. Texto Explicativo para o Mapa Geológico ao Milionésimo. **SICM/SGM, Salvador. (Special Edition)**, 1996.

BEGUIN, Julien et al. Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. **Geoderma**, v. 306, p. 195-205, 2017.

BEHRENS, Thorsten et al. Digital soil mapping using artificial neural networks. **Journal of plant nutrition and soil science**, v. 168, n. 1, p. 21-33, 2005.

BEHRENS, Thorsten et al. Spatial modelling with Euclidean distance fields and machine learning. **European journal of soil science**, v. 69, n. 5, p. 757-770, 2018.

BENEDET, Lucas et al. Rapid soil fertility prediction using X-ray fluorescence data and machine learning algorithms. **Catena**, v. 197, p. 105003, 2021.

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

BOMFIM, M.R. Características de ecossistemas manguezais contaminados por metais traços. Salvador, UFBA, Instituto de Geociências, 2014. 107p. (Tese Doutorado).

BRASIL. Ministério das Minas e Energia. Secretaria Geral. Projeto RADAMBRASIL Folha SD. 24 Salvador: Geologia, geomorfologia, pedologia, vegetação e uso potencial da terra. MME/SG/**Projeto RADAM BRASIL**, Rio de Janeiro, 1981.

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

BRENNING, Alexander. Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive

models. **Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie**, v. 19, n. 23-32, p. 410, 2008.

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

CAIN, Meghan K.; ZHANG, Zhiyong; YUAN, Ke-Hai. Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence, and estimation. **Behavior research methods**, v. 49, n. 5, p. 1716-1735, 2017.

CAIXETA, J. M.; BUENO, G. V.; MAGNAVITA, L. P.; FEIJÓ, F. J. Bacias do Recôncavo, Tucano e Jatobá. **Boletim de Geociências da PETROBRAS**, Rio de Janeiro, v. 8, n. 1, p. 163-172, 1994.

CAMPBELL, Patrícia Morais da Matta et al. Digital mapping of soil attributes using machine learning1. **Revista Ciência Agronômica**, v. 50, p. 519-528, 2019.

CHEN, Qi et al. Decision variants for the automatic determination of optimal feature subset in RF-RFE. **Genes**, v. 9, n. 6, p. 301, 2018.

CHEN, Songchao et al. Digital mapping of the soil thickness of loess deposits over a calcareous bedrock in central France. **Catena**, v. 198, p. 105062, 2021.

CHEN, Songchao et al. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. **Geoderma**, v. 409, p. 115567, 2022.

COELHO, Fabrício Fernandes et al. Digital soil class mapping in Brazil: a systematic review. **Scientia Agricola**, v. 78, 2020.

CONRAD, Olaf et al. System for automated geoscientific analyses (SAGA) v. 2.1. 4. **Geoscientific Model Development**, v. 8, n. 7, p. 1991-2007, 2015.

COSTA, Elias Mendes et al. Mapping soil properties in a poorly accessible area. **Revista Brasileira de Ciência do Solo**, v. 44, 2020.

CPRM – Companhia de Pesquisa de Recursos Minerais. Materiais de construção civil na região metropolitana de Salvador. Gonçalves, J. C. V; Moreira, M.D; Borges, V. P. Salvador: CPRM, 2008. 53p**. Informe de Recursos Minerais. Série Rochas e Minerais Industriais**, 2. Programa Geologia do Brasil- PGB.

CPRM – Companhia de Pesquisa de Recursos Minerais. Mapa geológico do estado da Bahia. Versão 1.1. Souza, J. D. de; Melo, R. C. de; Kosin, M. (Coords.). Salvador: CPRM, 2003. Escala 1:1.000.000.

CPRM – Companhia de Pesquisa de Recursos Minerais. Atlas pluviométrico do Brasil.CPRM, 2011. Versão.1. Escala 1.5:000.000. Pinto, E. J. de A.; Azambuja, A. M. S. de; Farias, J. A. M.; Salgueiro, J. P. de B.; Pickbrenner, K. (Coords.).; SIG - versão 2.0 - atualizada em 11/2011; **Levantamento da Geodiversidade**.

DE BENEDETTO, Daniela et al. Prediction of Soil Organic Carbon at Field Scale by Regression Kriging and Multivariate Adaptive Regression Splines Using Geophysical Covariates. **Land**, v. 11, n. 3, p. 381, 2022.

DSG - DIRETORIA DE SERVIÇO GEOGRÁFICO (DSG**). Banco de Dados Geográficos do Exército. Versão 3.0**. 2019.

EMBRAPA, Solos. SIBICS - Sistema brasileiro de classificação de solos. 2018.

ESTÉVEZ, Virginia et al. Machine learning techniques for acid sulfate soil mapping in southeastern Finland. **Geoderma**, v. 406, p. 115446, 2022.

EVERINGHAM, Y. L. et al. An introduction to multivariate adaptive regression splines for the cane industry. In: **Proceedings of the 2011 Conference of the Australian Society of Sugar Cane Technologists**. 2011.

GLOAGUEN, Thomas Vincent; PASSE, José João. Importance of lithology in defining natural background concentrations of Cr, Cu, Ni, Pb and Zn in sedimentary soils, northeastern Brazil. **Chemosphere**, v. 186, p. 31-42, 2017.

GOMES, Lucas Carvalho et al. Modelling and mapping soil organic carbon stocks in Brazil. **Geoderma**, v. 340, p. 337-350, 2019.

GUYON, Isabelle et al. Gene selection for cancer classification using support vector machines. **Machine learning**, v. 46, n. 1, p. 389-422, 2002.

GUZMÁN, Juliana et al. Estratigrafia da Bacia de Jatobá: estado da arte. **Estudos Geológicos**, v. 25, n. 1, p. 53-76, 2015.

HARTEMINK, Alfred E.; KRASILNIKOV, Pavel; BOCKHEIM, J. G. Soil maps of the world. **Geoderma**, v. 207, p. 256-267, 2013.

HEUNG, Brandon et al. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. **Geoderma**, v. 265, p. 62-77, 2016.

HOUNKPATIN, Kpade OL et al. Assessment of the soil fertility status in Benin (West Africa)–Digital soil mapping using machine learning. **Geoderma Regional**, v. 28, p. e00444, 2022.

HSEU, Zeng-Yei et al. Portable X-ray fluorescence (pXRF) for determining Cr and Ni contents of serpentine soils in the field. In: **Digital soil morphometrics**. Springer, Cham, 2016. p. 37-50.

HUDSON, Berman D. The soil survey as paradigm-based science. **Soil Science Society of America Journal**, v. 56, n. 3, p. 836-841, 1992.

IBGE. Instituto Brasileiro de Geografia e Estatística. Pedologia 1:250.000. 2018

JANG, Ho-Jun et al. Spatial pedological mapping using a portable X-ray fluorescence spectrometer at the Tallavera grove vineyard, Hunter Valley. **Korean Journal of Soil Science and Fertilizer**, v. 49, n. 6, p. 635-643, 2016.

JEONG, Gwanyong et al. Spatial soil nutrients prediction using three supervised learning methods for assessment of land potentials in complex terrain. **Catena**, v. 154, p. 73-84, 2017.

JEUNE, Wesly et al. Multinomial logistic regression and random forest classifiers in digital mapping of soil classes in western Haiti. **Revista Brasileira de Ciência do Solo**, v. 42, 2018.

KALAMBUKATTU, Justin George; KUMAR, Suresh; ARYA RAJ, R. Digital soil mapping in a Himalayan watershed using remote sensing and terrain parameters employing artificial neural network model. **Environmental earth sciences**, v. 77, n. 5, p. 1-14, 2018.

KALNICKY, Dennis J.; SINGHVI, Raj. Field portable XRF analysis of environmental samples. **Journal of hazardous materials**, v. 83, n. 1-2, p. 93-122, 2001.

KARATZOGLOU, Alexandros; MEYER, David; HORNIK, Kurt. Support vector machines in R. **Journal of statistical software**, v. 15, p. 1-28, 2006.

KASSAMBARA, Alboukadel. **Practical guide to cluster analysis in R: Unsupervised machine learning**. Sthda, 2017.

KESKIN, Hamza; GRUNWALD, Sabine; HARRIS, Willie G. Digital mapping of soil carbon fractions with machine learning. **Geoderma**, v. 339, p. 40-58, 2019.

KHALEDIAN, Yones; MILLER, Bradley A. Selecting appropriate machine learning methods for digital soil mapping. **Applied Mathematical Modelling**, v. 81, p. 401-418, 2020.

KHANIFAR, Javad. Modeling of soil thickness based on DEM derivatives calculated using different polynomials. **Arabian Journal of Geosciences**, v. 15, n. 7, p. 1-10, 2022.

KUHN, Max et al. Cubist models for regression. **R package Vignette R package version 0.0**, v. 18, p. 480, 2012.

KUHN, M., & JOHNSON, K. **Applied predictive modeling**. New York: Springer, 2013.

KUHN, Max et al. caret: Classification and Regression Training. R package version 6.0-86. **Astrophysics Source Code Library: Cambridge, MA, USA**, 2021.

LACOSTE, Marine et al. High-resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. **Geoderma**, v. 213, p. 296-311, 2014.

LANG, Bernhard. Monotonic multi-layer perceptron networks as universal approximators. In: **International conference on artificial neural networks**. Springer, Berlin, Heidelberg, 2005. p. 31-37.

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

LAGACHERIE, Philippe et al. Evaluating the impact of using digital soil mapping products as input for spatializing a crop model: The case of drainage and maize yield simulated by STICS in the Berambadi catchment (India). **Geoderma**, v. 406, p. 115503, 2022.

LANG, Bernhard. Monotonic multi-layer perceptron networks as universal approximators. In: **International conference on artificial neural networks**. Springer, Berlin, Heidelberg, 2005. p. 31-37.

LEGROS, Jean-Paul. **Mapping of the Soil**. Science Publishers, 2006.

LI, X; DU, L; LEE, S. Use of Topographic Models for Mapping Soil Properties and Processes. **Soil Systems**. 4. 32. 2020.

LIAW, Andy et al. Classification and regression by randomForest. **R news**, v. 2, n. 3, p. 18-22, 2002.

LIMA NETO, José de Almeida et al. Caracterização e gênese do caráter coeso em Latossolos Amarelos e Argissolos dos tabuleiros costeiros do Estado de Alagoas. **Revista brasileira de Ciência do Solo**, v. 33, p. 1001-1011, 2009.

LUO, Chong et al. Regional soil organic matter mapping models based on the optimal time window, feature selection algorithm and Google Earth Engine. **Soil and Tillage Research**, v. 219, p. 105325, 2022.

MA, Yuxin et al. Pedology and digital soil mapping (DSM). **European Journal of Soil Science**, v. 70, n. 2, p. 216-235, 2019.

MCBRATNEY, Alex B.; SANTOS, ML Mendonça; MINASNY, Budiman. On digital soil mapping. **Geoderma**, v. 117, n. 1-2, p. 3-52, 2003.

MEHDIZADEH, Saeid; BEHMANESH, Javad; KHALILI, Keivan. Using MARS, SVM, GEP and empirical equations for estimation of monthly mean reference evapotranspiration. **Computers and electronics in agriculture**, v. 139, p. 103-114, 2017.

MEIER, Martin et al. Digital soil mapping using machine learning algorithms in a tropical mountainous area. **Revista Brasileira de Ciência do Solo**, v. 42, 2018.

MELLO, Fellipe AO et al. Complex hydrological knowledge to support digital soil mapping. **Geoderma**, v. 409, p. 115638, 2022.

MENDES, Wanderson de Sousa et al. Geostatistics or machine learning for mapping soil attributes and agricultural practices. **Revista Ceres**, v. 67, p. 330-336, 2020.

MILLER, Bradley A. et al. Impact of multi-scale predictor selection for modeling soil properties. **Geoderma**, v. 239, p. 97-106, 2015.

MINASNY, Budiman et al. Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands. **Geoderma**, v. 313, p. 25-40, 2018.

MONTE NERO, Michelle Santos. Mapeamento geoquímico dos solos de Santo Amaro, UFRB, Bahia, 2020. (Dissertação de Mestrado)

NAWAR, Said et al. Modeling and mapping of soil salinity with reflectance spectroscopy and landsat data using two quantitative methods (PLSR and MARS). **Remote Sensing**, v. 6, n. 11, p. 10813-10834, 2014.

NAWAR, Said; MOUAZEN, Abdul M. Comparison between random forests, artificial neural networks, and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. **Sensors**, v. 17, n. 10, p. 2428, 2017.

OLIVERA, Elson Paiva et al. The Santa Luz chromite-peridotite and associated mafic dykes, Bahia-Brazil: remnants of a transitional-type ophiolite related to the Paleoproterozoic (> 2.1 Ga) Rio Itapicuru greenstone belt. 2007.

OLIVEIRA, I. A. et al. Caracterização de solos sob diferentes usos na região sul do Amazonas. **Acta amazônica**, Manaus, v. 45, n. 1, p. 1-12, 2015.

PELEGRINO, Marcelo Henrique Procópio et al. Prediction of soil nutrient content via pXRF spectrometry and its spatial variation in a highly variable tropical area. **Precision Agriculture**, v. 23, n. 1, p. 18-34, 2021.

PIIKKI, Kristin et al. Perspectives on validation in digital soil mapping of continuous attributes—A review. **Soil Use and Management**, v. 37, n. 1, p. 7-21, 2021.

POULADI, Nastaran et al. Mapping soil organic matter contents at field level with Cubist, Random Forest and kriging. **Geoderma**, v. 342, p. 85-92, 2019.

QUINLAN, John R. et al. Learning with continuous classes. In: **5th Australian joint conference on artificial intelligence**. 1992. p. 343-348.

R CORE TEAM. **R: A language and environment for statistical computing**, 2022.

RIBEIRO, Bruno Teixeira et al. Portable X-ray fluorescence (pXRF) applications in tropical Soil Science. **Ciência e Agrotecnologia**, v. 41, p. 245-254, 2017.

ROSSEL, RA Viscarra; BEHRENS, Thorsten. Using data mining to model and interpret soil diffuse reflectance spectra. **Geoderma**, v. 158, n. 1-2, p. 46-54, 2010.

SAHIN, Emrehan Kutlug. Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. **SN Applied Sciences**, v. 2, n. 7, p. 1-17, 2020.

SANTOS, P. S. DOS. **Estudo da vulnerabilidade ambiental no município de santo amaro-ba**. [s.l.] UFBA, 2015.

SEDGWICK, Philip. Pearson's correlation coefficient. **Bmj**, v. 345, 2012.

SEI. Superintendência de Estudos Econômicos e Sociais da Bahia. Divisão Político-Administrativa do Estado da Bahia-Vetor. Versão1.Salvador: SEI-BA, 2018.Escala: 1:100.000.

SILVA, Sérgio Henrique Godinho et al. Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence spectrometer (pXRF). **Ciência e Agrotecnologia**, v. 41, p. 648-664, 2017.

SIQUEIRA, Rafael G. et al. Evaluation of machine learning algorithms to classify and map landforms in Antarctica. **Earth Surface Processes and Landforms**, v. 47, n. 2, p. 367-382, 2022.

SHI, X. et al. Integrating different types of knowledge for digital soil mapping. **Soil Science Society of America Journal**, v. 73, n. 5, p. 1682-1692, 2009.

SMOLA, Alex J.; SCHÖLKOPF, Bernhard. A tutorial on support vector regression. **Statistics and computing**, v. 14, n. 3, p. 199-222, 2004.

TAGHIZADEH-MEHRJARDI, Ruhollah et al. Synthetic resampling strategies and machine learning for digital soil mapping in Iran. **European Journal of Soil Science**, v. 71, n. 3, p. 352-368, 2019.

TIGHE, Matthew et al. Georeferenced soil provenancing with digital signatures. **Scientific Reports**, v. 8, n. 1, p. 1-9, 2018.

TOUZANI, Samir; GRANDERSON, Jessica; FERNANDES, Samuel. Gradient boosting machine for modeling the energy consumption of commercial buildings. **Energy and Buildings**, v. 158, p. 1533-1543, 2018.

USEPA, UNITED STATES ENVIRONMENTAL PROTECTION AGENCY. Method 6200: Field portable X-ray fluorescence spectrometry for the determination of

elemental concentrations in soil and sediment. **Test Methods For Evaluating Solid Waste, US Environmental Protection Agency, Washington, DC, USA**, 2007.

WADOUX, Alexandre MJ-C.; MINASNY, Budiman; MCBRATNEY, Alex B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. **Earth-Science Reviews**, v. 210, p. 103359, 2020.

WADOUX, Alexandre MJ-C.; MCBRATNEY, Alex B. Digital soil science and beyond. **Soil Science Society of America Journal**, v. 85, n. 5, p. 1313-1331, 2021.

WEINDORF, David C.; BAKR, Noura; ZHU, Yuanda. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications. **Advances in agronomy**, v. 128, p. 1-45, 2014.

WERE, Kennedy et al. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. **Ecological Indicators**, v. 52, p. 394-403, 2014.

YAALON, Dan H. Soil science in transition: soil awareness and soil care research strategies. **Soil Science**, v. 161, n. 1, p. 3-8, 1996.

YANG, Yujian; TONG, Xueqin; ZHANG, Yingpeng. Spatial variability of soil properties and portable X-Ray fluorescence-quantified elements of typical golf courses soils. **Scientific Reports**, v. 10, n. 1, p. 1-14, 2020.

VASQUES, Gustavo M. et al. Field proximal soil sensor fusion for improving high-resolution soil property maps. **Soil Systems**, v. 4, n. 3, p. 52, 2020.

ZERAATPISHEH, Mojtaba et al. Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. **Geoderma**, v. 338, p. 445-452, 2019.

ZHANG, Mo; SHI, Wenjiao. Systematic comparison of five machine-learning methods in classification and interpolation of soil particle size fractions using different transformed data. **Hydrology and Earth System Sciences Discussions**, p. 1-39, 2019.

ZHANG, Wengang; GOH, Anthony TC. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. **Geoscience Frontiers**, v. 7, n. 1, p. 45-52, 2016.

ZHOU, Tao et al. Mapping of soil total nitrogen content in the middle reaches of the Heihe River Basin in China using multi-source remote sensing-derived variables. **Remote Sensing**, v. 11, n. 24, p. 2934, 2019.