

**Predição da variabilidade espacial do teor de argila em
lavouras de cana-de-açúcar no estado de São Paulo utilizando
aprendizado de máquina**

Guilherme Martineli Sanches

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Predição da variabilidade espacial do
teor de argila em lavouras de cana-
de-açúcar no estado de São Paulo
utilizando aprendizado de máquina

Guilherme Martineli Sanches

Guilherme Martineli Sanches

Predição da variabilidade espacial do teor de argila em lavouras de cana-de-açúcar no estado de São Paulo utilizando aprendizado de máquina

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientadora: Prof^a. Dr^a. Kalinka Branco

USP - São Carlos

2022

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi e Seção Técnica de Informática, ICMC/USP, com os dados inseridos pelo(a) autor(a)

MS211p

Martineli Sanches, Guilherme Predição da
variabilidade espacial do teor de argila em
lavouras de cana-de-açúcar no estado de São
Paulo utilizando aprendizado de máquina /
Guilherme Martineli Sanches; orientadora Kalinka
Branco. -- São Carlos, 2022.
56 p.

Trabalho de conclusão de curso (MBA em
Inteligência Artificial e Big Data) -- Instituto de
Ciências Matemáticas e de Computação, Universidade
de São Paulo, 2022.

1. ciência de dados. 2. agronegócio. 3. random
forest. 4. solos agrícolas. I. Branco, Kalinka,
orient. II. Título.

DEDICATÓRIA

À minha estimada esposa Priscilla Sanches e ao meu
amado filho Pedro Henrique pelo apoio,
companheirismo, paciência e amor demonstrados ao
longo desta jornada.

Com imenso carinho,

Dedico

Aos meus amados pais, Affonso e Malu, e irmão

Gustavo e à toda minha família,

Ofereço

AGRADECIMENTOS

Inúmeras são as pessoas a quem gostaria de expressar meu sentimento de gratidão por mais esta etapa em minha vida profissional. No entanto desde já lhe peço desculpa se acabar me esquecendo, mas no fundo do meu coração pode ter certeza de que você está aqui.

Primeiramente gostaria de agradecer a Professora Dr^a. Kalinka Branco pelo apoio e paciência transmitida ao longo desta jornada do MBA. A Prof^a Dr^a Solange Oliveira Rezende e Prof^a Dr^a Roseli Aparecida Francelin Romero pela dedicação à coordenação deste curso que, apesar dos desafios, me proporcionou muito conhecimento e um avanço significativo na minha vida profissional. A todos os professores do MBA em *Big Data* e Inteligência Artificial pela dedicação e aprendizado transmitidos a mim e meus colegas de turma. Também gostaria de agradecer aos meus companheiros de trabalho da “Cropman – Inteligência em Diagnósticos de Solos” pelo auxílio e por me proporcionarem um ambiente agradável e descontraído de trabalho. Aos verdadeiros amigos que compartilharam comigo, com imenso amor, os dias de angústia, ansiedade, tristezas e alegrias.

Agradeço a Deus pela saúde e pelos momentos em que me reestabeleceu quando achava que não teria mais forças, pela esposa amável e pelo meu filho. Agradeço a Deus pela graça de estar vivo em meio ao caos mundial dos dias atuais.

A verdadeira ciência descobre Deus
esperando atrás de cada porta.

Papa Pio XII

RESUMO

SANCHES, G.M. **Predição da variabilidade espacial do teor de argila em lavouras de cana-de-açúcar no estado de São Paulo utilizando aprendizado de máquina.** 2022. 56 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

A expansão da produção de cana-de-açúcar sob novas áreas é um cenário irreversível. Para atender as demandas internas e externas de produção de etanol e açúcar, bem como as metas de descarbonização mundial, a expansão das lavouras deverá ocorrer de maneira sustentável. Desta forma faz-se necessário expandir a produção em áreas de maior potencial produtivo, permitindo o aumento da produtividade agrícola em detrimento de grandes expansões territoriais. Para isso, é fundamental o conhecimento da textura do solo, isto é, do seu teor de argila, um dos principais parâmetros que definem o potencial produtivo das lavouras de cana-de-açúcar. O presente estudo teve como objetivo principal testar e validar o modelo de aprendizado de máquina *Random Forest* (RF) para predição da variabilidade espacial do teor de argila em lavouras de cana-de-açúcar no estado de São Paulo para fins de subsídio a empresas e produtores na expansão de suas áreas de produção. Um *dataset* composto de 2865 amostras de solo, provenientes do setor público e privado, foi avaliado no presente estudo, compreendendo teores de argila variando de solos arenosos (20 g kg^{-1}) a solos muito argilosos (779 g kg^{-1}). O modelo RF foi capaz de estimar o teor de argila com um erro médio quadrático (RMSE) de $56,47 \text{ g kg}^{-1}$ com $R^2 = 0,89$. Entre os atributos preditores avaliados, os “Top 5” que mais impactaram na predição do teor de argila foram as variáveis b6 (banda do SWIR 2), coordenada X, coordenada Y, temperatura e b1 (banda do Red). Por fim, o modelo calibrado foi aplicado, isto é, extrapolado para as áreas de cana-de-açúcar presentes no estado de São Paulo. Portanto, expandir a produção sobre áreas com maior conteúdo de argila significa garantia de melhores e maiores produtividades agrícolas, uma vez que o conteúdo de argila está totalmente ligado ao potencial produtivo das lavouras. Assim, a ciência de dados mostra o potencial de auxiliar governo, produtores e empresas a serem mais assertivos e estratégicos nas escolhas de novas áreas de produção, alcançando maior sustentabilidade do processo produtivo.

Palavras-chave: ciência de dados; agronegócio; florestas aleatórias; solos agrícolas.

ABSTRACT

SANCHES, G.M. **Prediction of clay content spatial variability in sugarcane fields at São Paulo state by machine learning algorithms.** 2022. 56 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2020.

The sugarcane fields expansion is an irreversible scenario. To meet the internal and external demands of ethanol and sugar production, as well as the global decarbonization goals, the expansion of sugarcane fields needs to be done by a sustainable manner. Thus, it is necessary to expand production in areas of greater yield potential, allowing for an increase in agricultural yield to the detriment of large territorial expansions. For this, it is essential to know the soil texture, that is, its clay content, one of the main parameters that define the yield potential of sugarcane crops. The main objective of this study was to test and validate the Random Forest (RF) machine learning model for predicting the spatial variability of clay content in sugarcane fields in São Paulo state, subsidizing companies and growers in the expansion of their production areas. A dataset composed of 2865 soil samples, from the public and private sector, were evaluated in the present study, comprising clay contents ranging from sandy soils (20 g kg^{-1}) to very clayey soils (779 g kg^{-1}). The RF model was able to estimate the clay content with a root mean square error (RMSE) of 56.47 g kg^{-1} with $R^2 = 0.89$. Among the predictor attributes evaluated, the “Top 5” that most impacted the clay content prediction were b6 (SWIR 2 band), X coordinate, Y coordinate, temperature and b1 (Red band). Finally, the calibrated model was extrapolated to the sugarcane fields present in the São Paulo state. Therefore, expanding production over areas with higher clay content means guaranteeing better and greater agricultural yield, since the clay content is totally linked to the yield potential. Thus, data science shows the great potential to help the government, producers and businesses to be more assertive and strategic in choosing new production areas, achieving greater sustainability in the production process.

Keywords: data science; agribusiness; random forest; agricultural soils.

LISTA DE ILUSTRAÇÕES

Figura 1. Mapa dos campos de cana-de-açúcar no Brasil.	30
Figura 2. Mapa com a geolocalização das amostras de solo utilizadas no presente estudo. Amostras de fonte pública (Pronasolos – triângulo preto) e fonte privada (Cropman – quadrado azul).	31
Figura 3. Variação do conteúdo de argila (g kg^{-1}) por fonte de dados.	32
Figura 4. Pipeline do processo de análise de dados e predição do conteúdo de argila no estado de São Paulo.	33
Figura 5. Variabilidade do conteúdo de argila pelas classes de curvatura vertical (a), horizontal (b) e de declividade (c).	35
Figura 6. Dados climáticos de temperatura média anual (a) e precipitação média acumulada anual (b) para o estado de São Paulo.	37
Figura 7. Mapa de reflectância do solo para o estado de São Paulo.	38
Figura 8. Importâncias relativas das variáveis utilizadas no modelo de predição.	42
Figura 9. Scatterplot entre o valor real e predito do conteúdo de argila (g kg^{-1}) da base de dados de teste.	44
Figura 10. Conteúdo de Argila (g kg^{-1}) (a) e classes texturais (b) preditas nas lavouras de cana-de-açúcar do estado de São Paulo.	45

LISTA DE TABELAS

Tabela 1. Produção de cana-de-açúcar, área plantada e produtividade no Brasil e no estado de São Paulo.	30
Tabela 2. Variáveis preditoras utilizadas para predição do conteúdo de argila no estado de São Paulo.	33
Tabela 3. Estatística descritiva das variáveis contínuas avaliadas no estudo.	40
Tabela 4. Estatística descritiva do conteúdo de argila para os datasets de treino e teste.....	40
Tabela 5. Correlação de Pearson das variáveis contínuas do estudo.	41

SUMÁRIO

1. INTRODUÇÃO	20
1.1 Contextualização e justificativa	20
1.2 Objetivos	21
1.3 Organização da Monografia	21
2. REVISÃO BIBLIOGRÁFICA	22
2.1 Produção de cana-de-açúcar	22
2.2 Agricultura de Precisão & Digital	24
2.3 Variabilidade espacial dos solos agrícolas	25
2.4 Aprendizado de máquina no contexto da agricultura	27
2.5 Considerações Finais	29
3 MATERIAL E MÉTODOS	30
3.1 Local do Estudo	30
3.2 Base de Dados	31
3.3 Análise de Dados	32
3.3.1 Modelo <i>Random Forest</i> (RF)	34
3.4 Variáveis Preditoras	34
3.4.1 Dados de Relevância	34
3.4.2 Dados Climáticos	36
3.4.3 Dados de Solo Exposto	37
3.5 Considerações Finais	38
4 RESULTADOS E DISCUSSÃO	39
4.1 Considerações iniciais	39
4.2 Análise dos Dados	39
4.3 Considerações Finais	46
5 CONCLUSÃO	47
5.1 Contribuições	47
5.2 Dificuldades Encontradas e Limitações	47
5.3 Trabalhos Futuros	47
REFERÊNCIAS	48

1. INTRODUÇÃO

1.1 Contextualização e justificativa

A população mundial irá atingir 9 bilhões de pessoas em 2050 segundo estimativas da Organização das Nações Unidas (ONU, 2021), causando um impensável aumento na demanda por alimentos, fibras e energia. A agricultura mundial terá que, ao mesmo tempo, dobrar a produção de alimentos sem expandir sobre reservas naturais e, reduzir a demanda por fertilizantes sintéticos. Para conciliar este desafio serão necessários sistemas integrados de produção, aumentando a intensidade de cultivos, a produtividade das culturas e uma melhor utilização dos recursos naturais (água, solo e nutrientes). Frente ao cenário mundial atual, o Brasil desempenha e continuará desempenhando papel fundamental no suprimento de alimentos, e energia à população mundial.

A expansão da produção agrícola sob novas áreas é um cenário irreversível. Para atender as demandas internas e externas de produção agrícola de maneira sustentável, faz-se necessário expandir a produção em áreas de maior potencial produtivo, permitindo o aumento da produtividade agrícola em detrimento de grandes expansões territoriais. Para isso, é fundamental o conhecimento da textura do solo, isto é, do seu teor de argila, um dos principais parâmetros que definem o potencial produtivo das lavouras de cana-de-açúcar (Sanches et al., 2019). Frente a este desafio, o uso de tecnologias de sensoriamento remoto (Demattê et al., 2009), aliadas às técnicas de aprendizado de máquina (Rezende, 2003), podem contribuir significativamente para obtenção de mapas precisos da variabilidade espacial do teor de argila dos solos brasileiros.

Mapas de solo são fundamentais para o manejo agrícola de terras cultivadas. No entanto, o mapeamento convencional de solo é caro e demorado (Peets et al., 2012), sendo que as cartas e mapas de solos existentes carecem de precisão (Coelho et al., 2021). Existe uma considerável variação de tecnologias utilizadas no Mapeamento Digital de Solos (MDS), sendo que nas últimas décadas foram realizados consideráveis avanços (Demattê et al., 2020). Entre os principais desafios destacam-se que as metodologias atuais precisam ser padronizadas, os erros minimizados e novas estratégias de mapeamento devem ser elaboradas para superar a falta de mapas detalhados (Minasny & McBratney, 2016; Arrouays et al., 2017). O mapeamento da textura dos solos agrícolas é fundamental na determinação do seu potencial produtivo (Demattê et al., 2020), porém o Brasil carece de mapas detalhados desta informação que possam apoiar

produtores e empresas. Conhecer o teor de argila dos solos permitirá aos produtores e grandes empresários estabelecer estratégias de expansão de suas áreas agrícolas, visando aumentar sua produtividade e lucratividade das lavouras para atender a demanda nacional e internacional de alimentos, fibras e energia.

1.2 Objetivos

Diante deste contexto, o presente estudo tem como objetivo principal testar e validar o modelo *Random Forest* para predição da variabilidade espacial do teor de argila em lavouras de cana-de-açúcar para fins de subsídio a empresas e produtores na expansão de suas áreas de produção. Ao final, o modelo será aplicado nas áreas de cana-de-açúcar no estado de São Paulo, principal polo produtor desta cultura do Brasil.

1.3 Organização da Monografia

O presente Trabalho de Conclusão de Curso (TCC) está organizado em cinco capítulos, sendo: Capítulo 1: Introdução e Justificativa, Capítulo 2: Revisão Bibliográfica, Capítulo 3: Material e Métodos, Capítulo 4: Resultados e Discussão e, Capítulo 5: Conclusão.

2. REVISÃO BIBLIOGRÁFICA

2.1 Produção de cana-de-açúcar

No último acordo climático em Glasgow no Reino Unido (COP-26), em novembro de 2021, foi assinado mais uma vez um compromisso internacional para limitar o aumento da temperatura global, principalmente pela adoção de uma economia de baixo carbono. O Brasil tem, e continuará tendo, um papel fundamental para as nações ao redor do mundo por meio da produção de etanol, fibra e energia de cana-de-açúcar, uma matriz altamente renovável e de baixo carbono que pode mitigar as mudanças climáticas (Nass et al., 2007). Assim, a produção e exportação de etanol no Brasil, que vem crescendo a cada ano (OCDE-FAO *Agricultural Outlook* 2015-2024, 2015), pode sofrer alterações mais significativas do que nas projeções anteriores.

O Brasil é o maior produtor de cana-de-açúcar (*Saccharum spp.*) do mundo com uma área cultivada de 8,6 milhões de hectares contra 27,3 milhões de hectares cultivados mundialmente (CONAB, 2021; FAOSTAT, 2019). Ficando somente atrás dos Estados Unidos da América, que possui uma produção de 52,6 bilhões de litros (53%) do etanol global, o Brasil é o segundo maior produtor do mundo, com 29,9 bilhões de litros (30%) de etanol (RFA, 2020). Para cumprir os compromissos das metas globais de bioenergia e reduzir as emissões de Gases de Efeito Estufa (GEE) em 43% abaixo dos níveis de 2005 até 2030 (iNDC Brasil 2015), o governo brasileiro deve aumentar a produção de etanol para 54 bilhões de litros até o ano de 2030, valor este 81% superior ao que foi produzido em 2018 (EPE, 2018). Tal fato demandará um total de 945 milhões de toneladas de cana-de-açúcar por safra (EPE, 2018); atualmente são produzidas 654 milhões de toneladas por safra (CONAB, 2021). Portanto, incentivos por meio de políticas públicas e investimentos contínuos em desenvolvimento científico e tecnológico são necessários para orientar a expansão da produção de cana-de-açúcar em um caminho sustentável, respeitando os limites ambientais, sem a necessidade de novos desmatamentos (Bordonal et al., 2018). Isso promoverá uma mudança irreversível no atual quadro energético no Brasil, e o setor sucroenergético tem uma grande oportunidade de substituir os combustíveis fósseis por biocombustíveis de baixo carbono (etanol) para cumprir essas metas.

Existem dois mecanismos principais para aumentar a produção de cana-de-açúcar e etanol, e assim cumprir os compromissos firmados pelo governo brasileiro na COP-26. Se por um lado é possível aumentar a área cultivada para expansão da produção, por outro lado é

possível aumentar a produtividade agrícola, isto é, produzir mais no mesmo local. Uma contribuição combinada de ambos os mecanismos também é uma estratégia viável.

Embora os agricultores brasileiros cultivem cana-de-açúcar em aproximadamente 8,6 milhões de hectares (CONAB, 2021), o Programa de Zoneamento Agroecológico designou que 65 milhões de hectares são adequados para o plantio de cana-de-açúcar sem maior expansão para florestas nativas ou biomas como Amazônia e Pantanal (Manzato et al., 2009). Apesar de muitas controvérsias relacionadas à expansão da cana-de-açúcar e seu impacto na produção de alimentos (Popp et al., 2014), vários estudos apoiam o potencial de expansão sustentável das culturas, especialmente em terras marginais e pastagens degradadas (Loarie et al., 2011; Macedo et al., 2015; Dias et al., 2016; Oliveira et al., 2019).

No entanto, há um grande potencial para melhorar a eficiência do manejo agrícola em sistemas de cultivo de cana-de-açúcar, diminuindo assim as lacunas de produtividade exploráveis (Marin et al., 2016). A produtividade atingível de cana-de-açúcar é de aproximadamente 300 Megagramas (Mg) por hectare (ha^{-1}) (Waclawovsky et al., 2010), que é muito superior à produtividade média atual dos canaviais no Brasil. No entanto, várias restrições de solo, cultura e clima limitam o aumento da produtividade no Brasil, que atualmente estagnou na faixa de 80 Mg ha^{-1} por quase uma década (Bordonal et al., 2018). A mecanização intensiva da colheita é relatada por Franco et al. (2018) como um dos principais motivos da estagnação da produtividade no Brasil. Evidências semelhantes foram observadas na Austrália até a década de 1990, que foram atribuídos ao aumento da mecanização das operações agrícolas (Franco et al., 2018).

Diante deste cenário, o aumento da produtividade agrícola da cana-de-açúcar figura a melhor solução do ponto de vista ambiental. No entanto, a expansão sobre novas áreas agrícolas será necessária para atingir as metas de produção. Para que esta expansão seja eficiente, ela deve ocorrer em áreas de maior potencial produtivo. Nesse contexto, a adoção de novas práticas e tecnologias será fundamental para superar esse cenário, onde a agricultura de precisão e digital (AP&D) é uma abordagem que inclui diversas tecnologias e ferramentas que podem contribuir significativamente no mapeamento do potencial produtivo das lavouras (Demattê et al., 2020), permitindo a superação desses desafios com resultados significativos para agricultores e empresas (Bramley et al., 2008).

2.2 Agricultura de Precisão & Digital

A Agricultura de Precisão & Digital (AP&D) é uma abordagem que busca aumentar a produção agrícola por meio do manejo específico do solo e da cultura, otimizando os insumos agrícolas e minimizando os impactos ambientais (Bullock et al., 2007). As principais tecnologias disponíveis para os usuários de AP&D são monitores de produtividade, sensores remotos e proximais de solo/planta associados a sistemas de navegação global (*Global Navigation Satellite System*) e diversos pacotes de Sistemas de Informações Geográficas (SIG). Apesar de muitas tecnologias disponíveis, no contexto do Brasil, as mais utilizadas são GNSS para orientação automática de máquinas agrícolas e tecnologias de aplicação à taxa variável (ATV) de fertilizantes de acordo com as necessidades específicas da cultura dentro das lavouras (Borghetti et al., 2016). Portanto, as tecnologias de AP&D permitem desde a otimização do tráfego de máquinas agrícolas para limitação da compactação do solo (Sousa et al., 2018; Aguilera et al., 2019) até uma gestão mais eficiente na aplicação de insumos (Sanches et al., 2021). Embora essas tecnologias estejam disponíveis em todo o mundo, as lavouras de cana-de-açúcar carecem da adoção efetiva dessas ferramentas (Silva et al., 2011) e ainda está longe de seu potencial máximo para manejo localizado no Brasil (Sanches et al., 2019).

Por outro lado, a AP&D trouxe uma nova era para a agricultura, a baseada em dados. A crescente adoção da tecnologia da informação (TI) tornou a coleta de dados uma tarefa simples em campos agrícolas. Atualmente, uma grande quantidade de dados e informações associadas à planta, ao solo e ao clima são coletados de forma rápida e fácil. No entanto, a transformação dos dados (coletados em campo) em um conhecimento correto para a tomada de decisão continua sendo um grande desafio.

O desenvolvimento de sistemas adequados de apoio à decisão para a implementação de decisões continua a ser um obstáculo importante para a completa adoção da AP&D (McBratney et al., 2006). No nível estratégico e tático, os dados reunidos sobre o desempenho de vários sistemas de gerenciamento de fazenda devem ser agrupados por séries temporais para construir uma base de dados sistemática, permitindo avaliações "rápidas e preliminares" dos efeitos das estratégias de gerenciamento com base em experiências obtidas em outros lugares em condições similares de solos (Bouma et al., 1999). Para superar este desafio, a tecnologia da informação tem sido amplamente aplicada em todos os aspectos da agricultura, tornando-se uma ferramenta eficaz para aumentar a produtividade agrícola e para fazer uso completo dos recursos disponíveis (Yan-e, 2011).

2.3 Variabilidade espacial dos solos agrícolas

A produção de alimentos, fibra e energia passará por mudanças significativas nos próximos anos (Sanches e Otto, 2022). Para aumentar a produção, será necessário avançar em processos e tecnologias que permitam aumentar a produtividade e lucratividade das culturas agrícolas. O incremento da produção deverá ser acompanhado de um manejo sustentável da cultura, melhorando a aplicação de insumos para reduzir os impactos ambientais (Sanches et al., 2021). O princípio dos 4R (lugar certo, quantidade certa, fonte certa e hora certa) para o manejo de fertilizantes proposto pelo Instituto Internacional de Nutrição Vegetal (IPNI) deve ser uma prioridade para garantir maiores retornos econômicos e ambientais da produção. O conhecimento da variabilidade espacial do solo é essencial para garantia de produção, além do correto manejo de insumos de forma sustentável. Para isso, caracterizar a variabilidade espacial do solo é um processo que deve ser realizado com alta precisão.

A digitalização da agricultura, onde é possível verificar o conceito de Agricultura de Precisão & Digital (AP&D), engloba um pacote de ferramentas e tecnologias que permitem caracterizar a variabilidade espacial e temporal dos solos com alta precisão, extraindo o máximo de conhecimento deles para otimização da produção e de insumos. A AP&D é considerada a abordagem mais viável para alcançar uma agricultura sustentável (Bullock et al., 2007).

Embora diversas tecnologias estejam disponíveis, a amostragem de solos para caracterizar a variabilidade espacial de seus atributos ainda desperta muito interesse na comunidade científica. Mapear a fertilidade do solo é um dos principais procedimentos para garantir um manejo adequado de fertilizantes. Além disso, mapear a variabilidade espacial dos nutrientes do solo é o caminho pelo qual a AP&D pode tomar decisões agronômicas eficientes. No entanto, um dos fatores limitantes para mapear o solo com alta precisão é a quantidade de amostras necessárias, o que muitas vezes inviabiliza física e economicamente o processo de amostragem (Peets et al., 2012). Além disso, estimativas de modelos baseados em uma única variável são caras e demoradas, especialmente quando análises laboratoriais estão envolvidas (Simbahan e Dobermann, 2006).

Para superar este desafio, alguns estudos têm sido realizados sobre procedimentos de amostragem de solo nas últimas décadas (Webster e Oliver, 1992; Machado et al., 2004; Resende et al., 2006; Coelho et al. 2007; Souza et al., 2014). Embora vários procedimentos tenham sido recomendados, a amostragem mais comumente utilizada para mapeamento dos atributos do solo, incluindo o conteúdo de argila, são as grades espaçadas regulares. Embora tenha benefícios, o alto custo, a baixa produtividade e o tempo gasto no mapeamento dos

atributos do solo por malhas regulares densas são questões que ainda precisam ser abordadas. A utilização de informações prévias sobre a variabilidade espacial do solo e parâmetros da paisagem local pode representar uma solução inteligente para superar esse gargalo, melhorando o processo de amostragem do solo e sua caracterização espacial (Sanches et al., 2018); principalmente em canaviais onde a adoção de tecnologias de AP e de taxa variável é baixa (Silva et al., 2011) e requer estudos de longo prazo como nas lavouras de grãos (Colaço e Molin, 2017).

A variabilidade espacial dos solos pode ocorrer por influências naturais ou influências antrópicas, sendo o resultado de cinco fatores principais: o clima, os microrganismos, o relevo, o material de origem e o tempo (Jenny, 1941). Dos cinco fatores, o relevo é o que pode ser mais rapidamente avaliado.

A atividade antrópica de cultivo ocasiona uma diferenciação do solo principalmente devido ao manejo e tratamentos utilizados, gerando alterações nas concentrações de nutrientes dos solos (Souza et al., 2014). A todos estes processos de alteração antrópico dos solos, é necessário dar devida atenção às influências causadas pelo relevo. A relação entre os solos e relevo é intrínseca, sendo que as formas geomorfológicas são precursoras da variação dos parâmetros físicos e químicos dos solos (Moore et al, 1993). Essa variabilidade, quando analisada pela ótica do relevo, é originada por pequenas alterações de declividade que alteram os processos pedogenéticos, o transporte e o armazenamento de água no perfil do solo (Sanchez et al., 2009). Apesar da relação intrínseca entre relevo e os atributos químicos e físicos do solo (Fulton et al., 1996; Chung et al., 2001, Gaston et al., 2001), atualmente o processo de amostragem para caracterização do solo não leva em consideração as diversas formações do relevo, como as propostas por Valeriano e Rosseti (2008), sendo escassos os trabalhos que investigam essa relação nos campos de cana-de-açúcar.

Os atributos do relevo, conhecidos como variáveis geomorfométricas, são informações extremamente importantes no entendimento da variabilidade químico-física do solo (Valeriano e Rosseti, 2008).

Entre as variáveis geomorfométricas, a declividade, a orientação das vertentes, a curvatura vertical, a curvatura horizontal e a própria altitude são as mais amplamente difundidas, sendo estas determinadas localmente pela derivação da superfície topográfica a partir de um Modelo Digital de Elevação (MDE). Estudos apresentam que a distribuição espacial dos atributos topográficos, que caracteriza as trajetórias do fluxo d'água, é um dos fatores responsáveis pela distribuição das propriedades do solo que ocorre em resposta a esse movimento de água dentro e sobre a superfície, permitindo realizar associações entre as

propriedades físico-químicas do solo e a topografia local (Moore et al, 1993). Brubaker et al. (1994) observaram que o conteúdo de argila, matéria orgânica, capacidade de troca catiônica e potássio estão relacionados com a posição no relevo, onde estes elementos tendem a decrescer do interflúvio superior para a direção do sopé. Wilson e Gallant (2000) trouxeram uma boa seleção de estudos que tiveram por objetivo investigar a relação entre as características do relevo com as propriedades do solo, evidenciando a importância das variáveis geomorfológicas na distribuição espacial das propriedades do solo.

Para garantir uma expansão sustentável das lavouras de cana-de-açúcar para alcançar as metas estabelecidas, o conhecimento do conteúdo de argila e dos seus padrões espaciais de variabilidade na paisagem é extremamente importante, onde técnicas de aprendizado de máquina poderão suprir esta demanda.

2.4 Aprendizado de máquina no contexto da agricultura

A coleta de dados e a extração de informações e conhecimento relevantes a respeito da variabilidade espacial dos solos agrícolas pode contribuir significativamente com a expansão sustentável das lavouras de cana-de-açúcar. O advento massivo da TI nos campos de cultivo cada dia mais traz uma nova era para a agricultura, a baseada em dados (Sanches et al., 2019). Neste contexto, as inúmeras ferramentas da chamada ciência de dados, aliada ao uso de tecnologias de sensoriamento remoto (Demattê et al., 2009) e de técnicas de aprendizado de máquina (Rezende, 2003), permitirão aumentar a produtividade agrícola e fazer uso completo dos recursos disponíveis (Yan-e, 2011).

O uso de ferramentas estatísticas e matemáticas nas lavouras de cana-de-açúcar para predição da produtividade e dos atributos do solo não é recente. Alguns trabalhos demonstram a influência dos atributos do solo na produtividade da cana-de-açúcar com respostas distintas para a camada arável e subsuperficial do solo.

Magalhães e Cerri (2007) avaliaram a possível correlação entre produtividade de cana-de-açúcar com alguns atributos químicos e físicos do solo. Os autores observaram que as correlações entre a produtividade de cana-de-açúcar com os atributos do solo são em geral fracas ($< 0,5$) e concluíram que a correlação simples entre os atributos do solo e produtividade da cana-de-açúcar não é suficiente para explicar a variabilidade espacial da cultura, sugerindo que outras características, além dos atributos de solo, sejam analisadas. Como sugerido por Cerri e Magalhães (2012), uma simples correlação de *Pearson* entre os parâmetros do solo e da

planta não é suficiente para explicar a produtividade. Neste sentido, os avanços em ciência de dados e *big data* (Wolfert et al., 2017) podem resolver esse gargalo.

Alguns estudos relatados na literatura usaram técnicas de mineração de dados, como algoritmos de *Random Forest* (Breiman, 2001), para estimar a produtividade da cana-de-açúcar (Bocca et al. 2015; Everingham et al., 2016; Bocca et al. 2016), mostrando o potencial dessas ferramentas. Os métodos de RF têm sido amplamente adotados para certos problemas agrícolas, como análise de sensoriamento remoto (Lebourgeois et al., 2017; Parente et al., 2017), estimativas de produtividade (Tulbure et al. al., 2012; Fukuda et al., 2013; Newlands et al. 2014), níveis foliares de nitrogênio (Abdel-Rahman et al. 2013) e classificação de variedades de cana-de-açúcar (Everingham et al., 2007). Os algoritmos de RF podem lidar com grandes volumes de dados, usar variáveis categóricas como preditores, medir o grau de importância das variáveis preditivas e gerar a probabilidade de classe e são robustos contra *overfitting*, mesmo para conjuntos de dados levemente desequilibrados (Khoshgoftaar et al., 2007).

Mais recentemente Sanches et al. (2019) utilizaram também algoritmos de RF para investigar os atributos do solo que mais impactam a produtividade da cultura de cana-de-açúcar em uma lavoura de 50 hectares. Como resultado os autores concluíram que a textura do solo, mais especificamente o conteúdo de argila, foi o fator que mais impactou a produtividade, demonstrando o potencial das ferramentas de aprendizado de máquina no contexto da agricultura baseada em dados.

A avaliação da variabilidade espacial do conteúdo de argila tem sido objeto de estudo por diversos autores e em diferentes locais do planeta, como Nigéria (Gobin et al., 2001), China (Wei et al., 2008), Irã (Zeraatpished et al., 2019), Chile (Mashalaba et al., 2020) e Brasil (Pinheiro et al., 2017, Coelho et al., 2021, Helfer et al., 2021). Tais estudos buscam investigar a predição do teor de argila na paisagem por diferentes métodos e técnicas de aprendizado de máquina. Com este objetivo, Zeraatpished et al. (2019) utilizaram amostras de solo, covariáveis do relevo e imagens de satélite na tarefa de predição espacial de atributos do solo na região de Bojuren-Irã. Entre as técnicas avaliadas, incluindo os modelos *Cubist*, *Random Forest*, *Regression Tree* e *Multiple Linear Regression*, os autores encontram que o modelo RF apresentou o menor erro médio quadrático (RMSE) para todos os atributos avaliados, incluindo o conteúdo de argila (RMSE = 7,8 g kg⁻¹). Os atributos declividade e curvatura do relevo foram as variáveis mais explicativas da variabilidade espacial do conteúdo de argila.

Em uma ampla revisão realizada por Coelho et al. (2021) sobre mapeamento de solo e técnicas de aprendizado de máquina, os autores encontraram que 54% dos estudos avaliados utilizaram árvores de decisão (AD) na tarefa de classificação de solos. Em seguida, as técnicas

de Redes Neurais Artificiais (RNA) e Regressão Logística (RL) foram as mais utilizadas, representando 25% e 16% dos estudos avaliados. Os autores ainda concluíram que as RNA são mais eficientes que a RL, não encontrando evidências de que estas seriam superiores as AD.

2.5 Considerações Finais

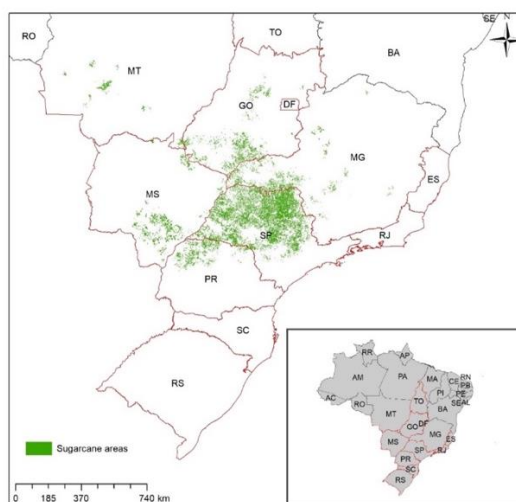
Tendo em vista a importância da cultura de cana-de-açúcar para o Brasil e sua necessária expansão para as próximas décadas, o uso de técnicas de *machine learning* serão cada vez mais necessárias para apoiar a tomada de decisão dos produtores e empresas. Para tal, a predição do conteúdo de argila utilizando tais técnicas poderá auxiliar fortemente na decisão assertiva de locais estratégicos de expansão, isto é, que apresentam maior e mais adequado potencial produtivo (maior conteúdo de argila). Dados públicos e privados para compor a base de dados são fundamentais nesse sentido, conforme detalhado no tópico a seguir, cujos dados serão utilizados para calibração e validação do modelo *Random Forest*.

3 MATERIAL E MÉTODOS

3.1 Local do Estudo

O presente estudo para predição do conteúdo de argila em campos de cana-de-açúcar foi desenvolvido no estado de São Paulo (SP), principal estado produtor do Brasil (Figura 1). O estado de São Paulo fica na principal região produtora de cana-de-açúcar do Brasil, a região Centro-Sul (SCR), que responde por cerca de ~90% da produção brasileira, sendo São Paulo responsável por ~51% desta produção (Tabela 1). Nesta safra, SP foi responsável por 44% da produção de etanol e 61% da produção de açúcar no Brasil (CONAB, 2022). A produtividade média em SP foi de 71,6 Mg ha⁻¹ na safra 2021/2022, enquanto a produtividade média no Brasil foi de 70,3 Mg ha⁻¹.

Figura 1. Mapa dos campos de cana-de-açúcar no Brasil.



Fonte: O autor.

Tabela 1. Produção de cana-de-açúcar, área plantada e produtividade no Brasil e no estado de São Paulo.

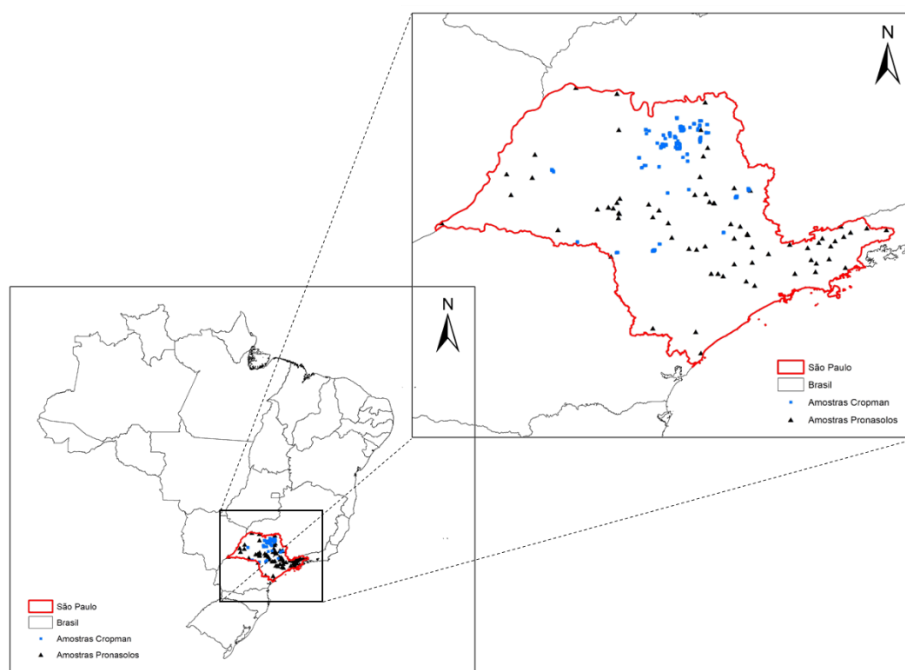
	Produção [10 ⁶ Mg]	Área [10 ⁶ ha]	Produtividade [Mg ha ⁻¹]	Açúcar [10 ⁶ Mg]	Etanol [10 ⁹ L]
Safra 2021/2022					
Brasil	585,2	8,3	70,3	35,1	26,8
São Paulo	298,5	4,2	71,6	21,4	11,9

Fonte: (CONAB, 2022).

3.2 Base de Dados

A base de dados de amostras de solo com o conteúdo de argila foi obtida de duas fontes principais, sendo uma pública e outra privada (Figura 2). A primeira delas corresponde as amostras de solo da empresa Cropman¹ (base de dados privada), contendo um total de 2786 amostras. Já a segunda fonte de dados corresponde a base de dados públicos do Pronasolos², que é o Programa Nacional de Levantamento e Interpretação de Solos do Brasil. Esta segunda contém um total de 79 amostras de solo. Para ambas as bases de dados foram considerados apenas as amostras de solo que estão na camada superficial do solo, isto é, até a profundidade de 0,20m. Apesar do número inferior de amostras de solo da base do Pronasolos, é possível observar sua maior relevância em termos de espacialidade, enquanto as amostras da Cropman apresentam uma maior concentração espacial no estado.

Figura 2. Mapa com a geolocalização das amostras de solo utilizadas no presente estudo. Amostras de fonte pública (Pronasolos – triângulo preto) e fonte privada (Cropman – quadrado azul).



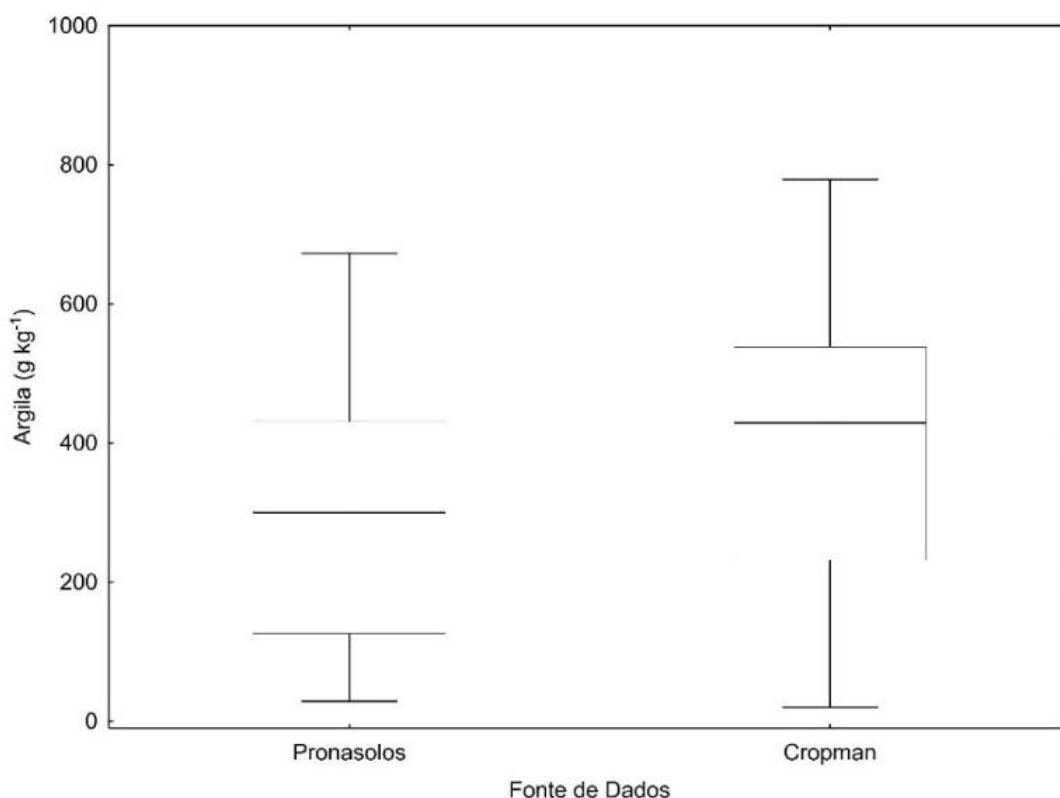
Fonte: O autor.

¹ www.cropman.com.br

² <https://www.embrapa.br/pronasolos>

O conteúdo de argila apresentou variação semelhante de acordo com a fonte de dados, apesar da alta diferença em termos de número de amostras. Enquanto a Cropman apresentou uma variação no conteúdo de argila de 20 a 779 g kg⁻¹ (mediana de 429 g kg⁻¹), o Pronasolos variou de 29 a 673 g kg⁻¹ (mediana de 300 g kg⁻¹) (Figura 3).

Figura 3. Variação do conteúdo de argila (g kg⁻¹) por fonte de dados.

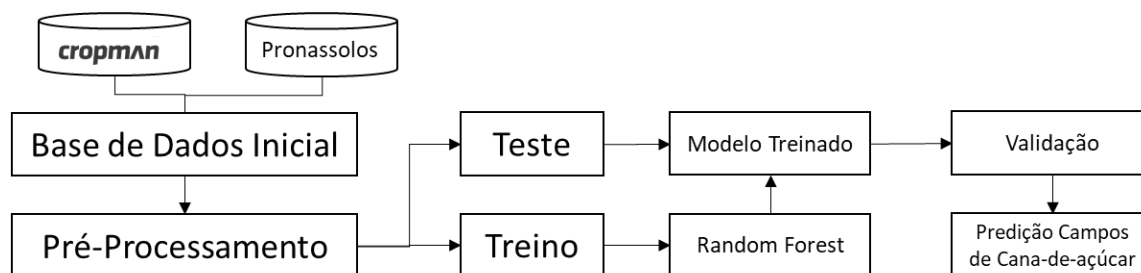


Fonte: O autor.

3.3 Análise de Dados

A análise de dados foi realizada conforme o *pipeline* ilustrado na Figura 4. A base de dados do projeto foi composta por amostras de solo da Cropman e do Programa Pronasolos. Primeiramente foram filtradas as amostras de solo que estavam na camada superficial do solo, isto é, até o limite de 0,20m de profundidade. Em seguida as amostras foram analisadas para remoção dos valores discrepantes, isto é, aquelas que estavam fora do intervalo de ± 3 do desvio padrão foram tratadas como *outliers*.

Figura 4. Pipeline do processo de análise de dados e predição do conteúdo de argila no estado de São Paulo.



Fonte: O autor.

Em seguida, por meio de um *software* de Sistema de Informação Geográfica (SIG) foram extraídos os valores de todas as variáveis preditoras (Tabela 2). Nesta etapa utilizou-se o *software* ArcMaP 10.6[®]. Na etapa de pré-processamento ainda foi avaliado a estatística descritiva e a correlação de Pearson entre as variáveis contínuas do estudo. Após, o conjunto de dados foi dividido em treino e teste, sendo utilizado a proporção, respectivamente, de 70% e 30%. Com o modelo treinado e testado, extrapolou-se o mesmo para os campos de cana-de-açúcar presentes no estado de São Paulo. Os campos de cana-de-açúcar no estado de São Paulo foram obtidos do Projeto de Mapeamento Canasat³.

Tabela 2. Variáveis preditoras utilizadas para predição do conteúdo de argila no estado de São Paulo.

ID	Variável	Descrição	Tipo	Unidade
1	ZN	Elevação do Terreno	Numérica	metros [m]
2	V5	Classes de Curvatura Vertical	Categórica	-
3	H5	Classes de Curvatura Horizontal	Categórica	-
4	SN	Declividade	Numérica	Porcentagem [%]
5	SA	Classes de Declividade	Categórica	-
6	Coord. X	Coordenada X	Numérica	Graus decimais [°]
7	Coord. Y	Coordenada Y	Numérica	Graus decimais [°]
8	b1	Banda 1 - Red	Numérica	Reflectância
9	b2	Banda 2 - Green	Numérica	Reflectância
10	b3	Banda 3 - Blue	Numérica	Reflectância
11	b4	Banda 4 - NIR	Numérica	Reflectância
12	b5	Banda 5 - SWIR 1	Numérica	Reflectância
13	b6	Banda 6 - SWIR 2	Numérica	Reflectância
14	Precip.	Precipitação Média Anual Acumulada	Numérica	milímetros [mm]
15	Temp.	Temperatura Média Anual Acumulada	Numérica	Graus Celsius [°C]

Fonte: O autor.

³ <http://www.dsr.inpe.br/laf/canasat/>

3.3.1 Modelo *Random Forest* (RF)

As árvores de decisão pertencem à classe dos algoritmos supervisionados, onde uma variável dependente é explicada em função de n variáveis independentes, medidas em qualquer escala. Uma árvore de decisão consiste em um conjunto de regras que associam um conjunto de condições com um resultado específico (Breiman, 2001). Estas regras também podem ser representadas em um formato de árvore intuitiva, permitindo a visualização das relações entre os preditores e a saída.

O princípio existente nas árvores de classificação é o de “dividir-para-conquistar”. Deste modo, em cada nível de uma árvore, um problema mais complexo de previsão/classificação (em que existe maior heterogeneidade de valores da variável alvo) é decomposto em subproblemas mais simples. Tal fato se traduz na geração de nós descendentes, nos quais, a heterogeneidade da variável a prever (e explicar) é mais atenuada, podendo as previsões serem efetuadas com menores riscos para cada um desses nós (Santos Rodrigues, 2005). Trata-se de uma investigação que segue o fluxo do nível mais “geral” para o caso “particular”, no sentido em que cada novo nível da árvore os nós descendentes se limitam ao valor de mais um atributo explicativo.

No presente trabalho a árvore de decisão foi utilizada para estimar o conteúdo de argila (variável dependente) com base nos parâmetros do relevo, clima e refletância do solo (variáveis independentes), descritas nas seções seguintes.

Diversos são os algoritmos utilizados nas árvores de decisão, sendo que no presente estudo utilizou-se o algoritmo *Random Forest* (RF) descrito em Breiman (2001). Para aplicação do RF adotou-se como parâmetros 100 árvores para iteração com a relação 70%/30% dos valores para o treinamento e teste do modelo. Todas as análises foram executadas por meio da linguagem de programação *Python*, utilizando o pacote *Scikit-learn*. Ainda, foi utilizado o RF para calcular a importância das variáveis preditoras.

3.4 Variáveis Preditoras

3.4.1 Dados de Relevo

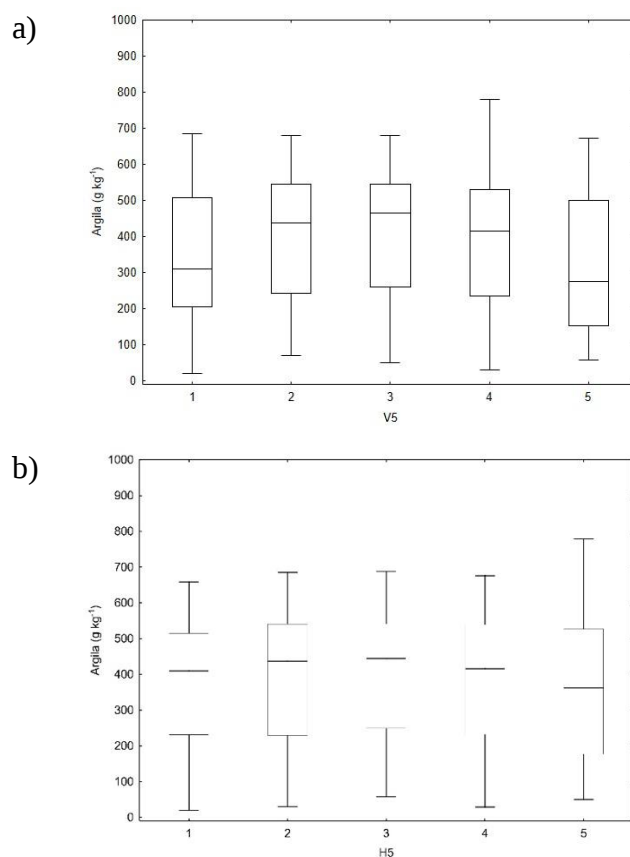
Os parâmetros do relevo (ID 1 ao 5) utilizados no presente projeto foram obtidos do banco de dados Topodata (Brasil, 2008). Esses dados foram gerados por Valeriano e Rosseti (2008) e Valeriano e Albuquerque (2010), onde os dados do SRTM (*Shuttle Radar Topography*

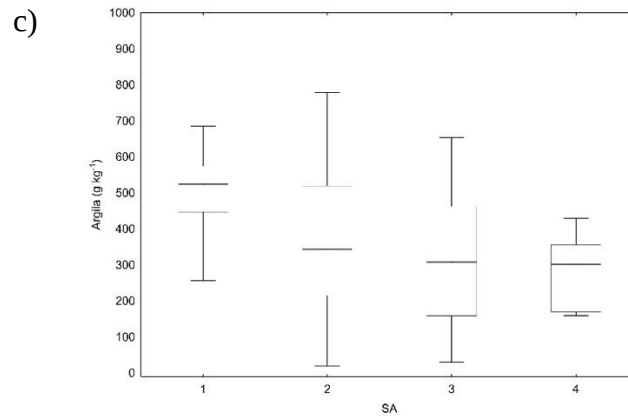
Mission) foram refinados para uma resolução espacial de 30 metros (Rabus et al., 2003; Valeriano e Rossetti, 2012).

As variáveis geomorfológicas das curvaturas vertical (ID 2) e horizontal (ID 3) foram avaliadas. A curvatura vertical é classificada de 1 a 5, correspondendo a terrenos côncavos (1 e 2), retilíneos (3) e convexos (4 e 5). A curvatura horizontal é classificada também de 1 a 5, correspondendo a terrenos convergentes (1 e 2), planares (3) e divergentes (4 e 5); ambos de acordo com a classificação proposta por Valeriano (2008). Já a variável classes de declividade (ID 5) seguiu a classificação proposta por Marques (1971), no qual estão presentes as classes: suave (< 2,5%), moderado (2,5 a 12%), forte (12 a 50%) e muito forte (> 50%).

As variáveis elevação (ID 1) e declividade (ID 4) também foram consideradas. O conteúdo de argila mostrou ser maior na classe 3 de ambas as curvaturas vertical (Figura 5 – a) e horizontal (Figura 5 – b) do relevo analisadas, sendo que para as classes de declividade o conteúdo de argila é maior nas formações suaves e moderado (Figura 5 – c). Por fim, os pontos de amostragem do solo, com suas respectivas coordenadas X (ID 6) e Y (ID 7), foram associados às variáveis do relevo a que pertenciam.

Figura 5. Variabilidade do conteúdo de argila pelas classes de curvatura vertical (a), horizontal (b) e de declividade (c).



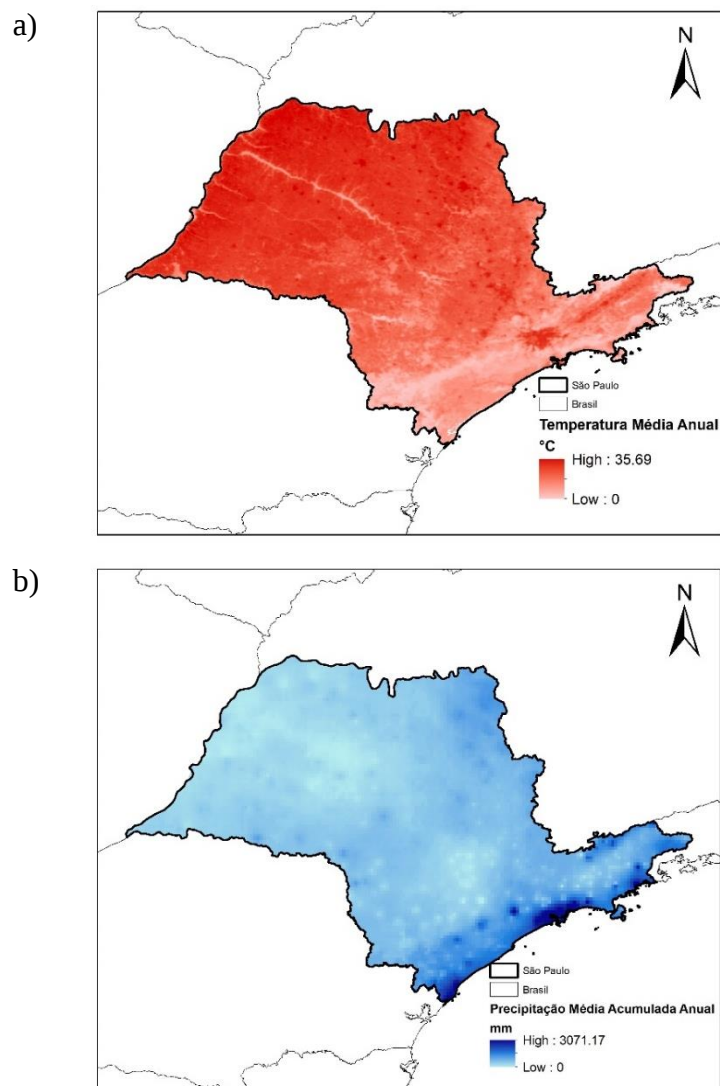


Fonte: O autor.

3.4.2 Dados Climáticos

Os dados climáticos de precipitação (ID 14) e temperatura (ID 15) foram obtidos por meio do *Climate Hazards Group InfraRed Precipitation with Stations* (CHIRPS) (Figura 6). Desenvolvido pelo *United States Geological Survey* (USGS) e pelo *Climate Hazards Group at the University of California, Santa Barbara* (UCSB), o CHIRPS é uma base de dados composta por diferentes fontes de informação, tais como, *The Climate Hazards Group's Precipitation Climatology* (CHPCLim); Observações de satélites com espectroscopia de infravermelho termal (*Thermal Infrared, TIR*), geoestacionárias quase globais da *National Oceanic and Atmospheric Administration* (NOAA); Centro de Previsão Climática (CPC); *National Climatic Data Center Cli* (NCDC); *Coupled Forecast System* da NOAA, versão 2 (CFSv2) e dados observacionais de estações meteorológicas (Funk et al., 2015). Para ambas as variáveis foi considerado o período de 20 anos para composição da variável, sendo de 2002 a 2022. Para a precipitação foi considerado a média do valor acumulado anual no período de 20 anos, enquanto para a temperatura foi considerada a média anual do período avaliado.

Figura 6. Dados climáticos de temperatura média anual (a) e precipitação média acumulada anual (b) para o estado de São Paulo.



Fonte: O autor.

3.4.3 Dados de Solo Exposto

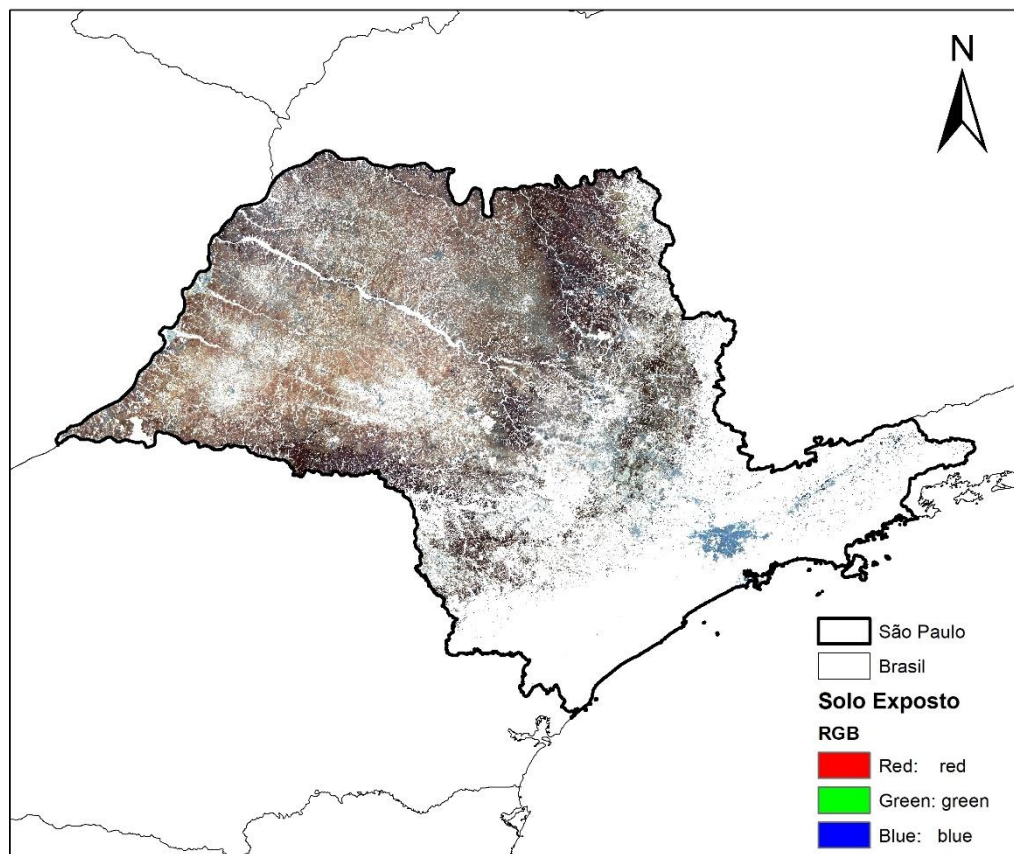
Para as variáveis de solo exposto foram consideradas bandas de reflectâncias (ID 8 ao 13) de imagens de satélite (Figura 7). Utilizou-se para o presente trabalho os dados dos satélites LANDSAT 5, 7 e 8 combinados por meio da Plataforma *Google Earth Engine* (GEE). Para obtenção das reflectâncias das bandas com o solo exposto foi seguido um conjunto de regras combinadas descritas em Demattê et al. (2018), considerando um período de 10 anos (2012 a 2022).

O conceito por trás destas variáveis é que cada alvo da superfície terrestre ou das lavouras (solo ou planta) apresentam um padrão de assinatura característico, que pode ser

avaliado a partir da sua curva espectral. No solo, as formas e intensidades da curva espectral revelam as feições relacionadas à sua textura, teor de matéria orgânica e mineralogia (Demattê et al., 2020).

De maneira geral, os solos mais arenosos apresentam elevado albedo na região do infravermelho médio (SWIR), devido à maior intensidade de reflectância do quartzo. Por outro lado, os solos mais argilosos, apresentam menor albedo devido a certos minerais presentes no solo que absorvem a radiação com maior intensidade. Assim, tanto a planta quanto o solo podem ser analisados por sensores localizados em laboratórios, campo, drones ou satélites, auxiliando na predição do conteúdo de argila.

Figura 7. Mapa de reflectância do solo para o estado de São Paulo.



Fonte: O autor.

3.5 Considerações Finais

De posse dos dados, a próxima seção descreve os resultados obtidos aplicando-se o modelo RF e os discute a luz de estudos presentes na literatura.

4 RESULTADOS E DISCUSSÃO

4.1 Considerações iniciais

Este capítulo apresenta os resultados obtidos a partir da aplicação do modelo RF nos dados do banco de dados.

4.2 Análise dos Dados

Do total de amostras do banco de dados (2865 amostras), as bandas de reflectância do solo exposto apresentaram ~11% de valores faltantes ($N = 2551$) (Tabela 3). De acordo com Demattê et al. (2020), locais que apresentam valores ausentes de solo exposto se devem ao fato de serem áreas de cultivo recentemente abertas, onde as imagens detectam ainda vegetação nativa. Por outro lado, o fato de no presente trabalho termos considerado o período de 10 anos para análise do solo exposto pode ter ocasionado os valores faltantes das bandas. Um período maior de avaliação em trabalhos futuros poderá suprir essas os valores de reflectância para essas variáveis.

A temperatura média anual dos locais amostrados foi de ~30°C, apresentando ainda uma precipitação acumulada anual de 1326 mm (Tabela 1). Ambas as variáveis apresentam um baixo coeficiente de variação (4-5%). A cana-de-açúcar pode apresentar uma necessidade hídrica anual variando de 1500 a 2500 mm (Doorembos & Kassam, 1994), sendo que no presente estudo os locais onde as amostras estavam compreenderam uma variação de 1165 a 2542 mm, valores esses próximos aos destacados pela literatura.

A altitude média das localidades foi de 553 m, variando de locais próximos ao nível do mar (~49 m) a regiões elevadas (1333 m). A declividade foi de locais planos (0%) a íngremes (64%), porém com uma média de 5%.

O conteúdo de argila da base de dados apresentou um coeficiente de variação de ~44%, com teores variando de solos arenosos (20 g kg⁻¹) a muito argilosos (779 g kg⁻¹) (Tabela 1). Os *datasets* de treino e teste apresentaram uma distribuição semelhante no conteúdo de argila, com coeficiente de variação próximo ao da base de dados completa, com média aproximada de 385-387 g kg⁻¹ de argila (Tabela 4). Pelo conteúdo de argila observado nos pontos amostrais avaliados, o presente estudo foi capaz de abranger diferentes tipos de solo. O presente trabalho compreendeu locais muito arenosos (argila < 150 g kg⁻¹) a muito argilosos (argila > 600 g kg⁻¹). Dessa forma, foram contempladas diferentes classes de fertilidade do solo, uma vez que a

classe textural do solo está diretamente relacionada à disponibilidade de água e teores de nutrientes (Raij, 2001).

Tabela 3. Estatística descritiva das variáveis contínuas avaliadas no estudo.

	N	Média	Mediana	Min.	Max.	DP	CV
Coord. X	2865	-48.31	-48.31	-52.98	-44.32	0.69	-1.44
Coord. Y	2865	-21.09	-20.71	-25.03	-19.87	0.70	-3.34
b1	2551	475.03	456.00	247.00	998.00	94.08	19.81
b2	2551	757.75	726.00	399.00	1296.00	140.48	18.54
b3	2551	1157.89	1124.00	711.00	1930.00	179.41	15.49
b4	2551	1639.89	1540.00	1019.00	2896.00	332.45	20.27
b5	2551	2032.75	1882.00	1142.00	4082.00	516.62	25.41
b6	2551	1832.88	1674.00	1088.00	3663.00	439.39	23.97
Temp.	2865	30.30	30.66	19.63	32.71	1.24	4.10
Prec.	2865	1326.48	1308.92	1165.83	2542.38	73.35	5.53
ZN	2865	553.57	517.89	49.47	1333.86	91.34	16.50
SN	2865	5.05	4.00	0.00	64.00	4.14	82.00
Argila	2865	386.73	422.00	20.00	779.00	169.02	43.70

Coord.: Coordenada, Temp.: Temperatura, Prec.: Precipitação, ZN: Elevação, SN: Declividade, N: números válidos, Min.: Mínimo, Max.: Máximo, DP: Desvio Padrão, CV: Coeficiente de Variação. Unidades: Coord. X/Coord. Y: [graus decimais], b1-b6: [reflectância], Temp.: [°C], Prec.: [mm], ZN: [m], SN: [%], Argila: [g kg⁻¹]

Fonte: O autor

Tabela 4. Estatística descritiva do conteúdo de argila para os *datasets* de treino e teste.

	N	Média	Mediana	Min.	Max.	DP	CV
Dataset Treino							
Argila	2006	387.41	422.00	33.00	779.00	168.47	43.49
Dataset Teste							
Argila	859	385.15	421.00	20.00	688.00	170.39	44.24

N: números válidos, Min.: Mínimo, Max.: Máximo, DP: Desvio Padrão, CV: Coeficiente de Variação. Unidades: Argila: [g kg⁻¹]

Fonte: O autor

As bandas do solo exposto foram as que apresentaram as maiores correlações com o conteúdo de argila ($r = -0,39$ a $-0,84$) (Tabela 5). A b5 e b6, corresponde as bandas do SWIR1 e SWIR2, respectivamente, apresentaram a maiores correlações ($r = -0,82$ e $-0,84$). Assim como relatado por Demattê et al. (2009), as bandas do SWIR apresentam alta correlação com o conteúdo de argila do solo por serem espectros que apresentam picos de absorção devido à umidade do solo. Solos mais argilosos tendem a apresentar maiores absorções nestas bandas do espectro, isto é, menores reflectâncias, ocasionando assim a correlação inversa e negativa entre as variáveis.

A elevação do terreno ($r = -0,22$) e a declividade ($r = -0,33$) também apresentaram correlações negativas e significativas com o conteúdo de argila, onde locais menos elevados e mais declivosos apresentaram teores maiores. Estudos mostraram que a relação entre os parâmetros físicos do solo e o relevo é intrínseca (Moore et al., 1993).

A variabilidade dos atributos físicos, e químicos, do solo deve-se às mudanças de declividade que alteram os processos pedogenéticos, transporte e armazenamento de água no perfil do solo (Sanchez et al., 2009). A correlação significativa com a coordenada Y ($r = 0,39$) mostra que o conteúdo de argila é influenciado fortemente pela latitude, mostrando assim a importância de se levar em consideração os parâmetros de localização geográfica. Por sua vez, a temperatura, que apresenta forte correlação com a coordenada Y ($r = 0,73$), apresentou correlação positiva e significativa com o conteúdo de argila também ($r = 0,31$).

Observando-se a Figura 6 - a é possível notar a correlação positiva entre a temperatura média e a latitude, isto é, quanto mais nos aproximamos do Equador (latitudes mais altas), maiores vão ficando as temperaturas médias, influenciando assim na variação do conteúdo de argila.

Tabela 5. Correlação de Pearson das variáveis contínuas do estudo.

	Coord. X	Coord. Y	b1	b2	b3	b4	b5	b6	Temp.	Prec.	ZN	SN	Argila
Coord. X	1,00	0,05*	-0,15*	-0,21*	-0,29*	-0,26*	-0,18*	-0,19*	-0,25*	0,52*	0,35*	-0,15*	0,13*
Coord. Y	0,05*	1,00	-0,44*	-0,44*	-0,34*	-0,44*	-0,52*	-0,53*	0,73*	-0,14*	-0,48*	-0,33*	0,39*
b1	-0,15*	-0,44*	1,00	0,94*	0,81*	0,82*	0,83*	0,84*	-0,27*	-0,01	0,16*	0,29*	-0,68*
b2	-0,21*	-0,44*	0,94*	1,00	0,93*	0,93*	0,89*	0,90*	-0,26*	-0,06*	0,15*	0,32*	-0,75*
b3	-0,29*	-0,34*	0,81*	0,93*	1,00	0,95*	0,85*	0,86*	-0,15*	-0,09*	0,10*	0,30*	-0,70*
b4	-0,26*	-0,44*	0,82*	0,93*	0,95*	1,00	0,94*	0,94*	-0,24*	-0,09*	0,16*	0,36*	-0,80*
b5	-0,18*	-0,52*	0,83*	0,89*	0,85*	0,94*	1,00	0,99*	-0,34*	-0,07*	0,19*	0,33*	-0,82*
b6	-0,19*	-0,53*	0,84*	0,90*	0,86*	0,94*	0,99*	1,00	-0,35*	-0,07*	0,21*	0,33*	-0,84*
Temp.	-0,25*	0,73*	-0,27*	-0,26*	-0,15*	-0,24*	-0,34*	-0,35*	1,00	-0,12*	-0,44*	-0,21*	0,31*
Prec.	0,52*	-0,14*	-0,01	-0,06*	-0,09*	-0,09*	-0,07*	-0,07*	-0,12*	1,00	0,41*	0,12*	0,05*
ZN	0,35*	-0,48*	0,16*	0,15*	0,10*	0,16*	0,19*	0,21*	-0,44*	0,41*	1,00	0,16*	-0,22*
SN	-0,15*	-0,33*	0,29*	0,32*	0,30*	0,36*	0,33*	0,33*	-0,21*	0,12*	0,16*	1,00	-0,33*
Argila	0,13*	0,39*	-0,68*	-0,75*	-0,70*	-0,80*	-0,82*	-0,84*	0,31*	0,05*	-0,22*	-0,33*	1,00

*Significativo à 5%. Coord.: Coordenada, Temp.: Temperatura, Prec.: Precipitação, ZN: Elevação, SN: Declividade.

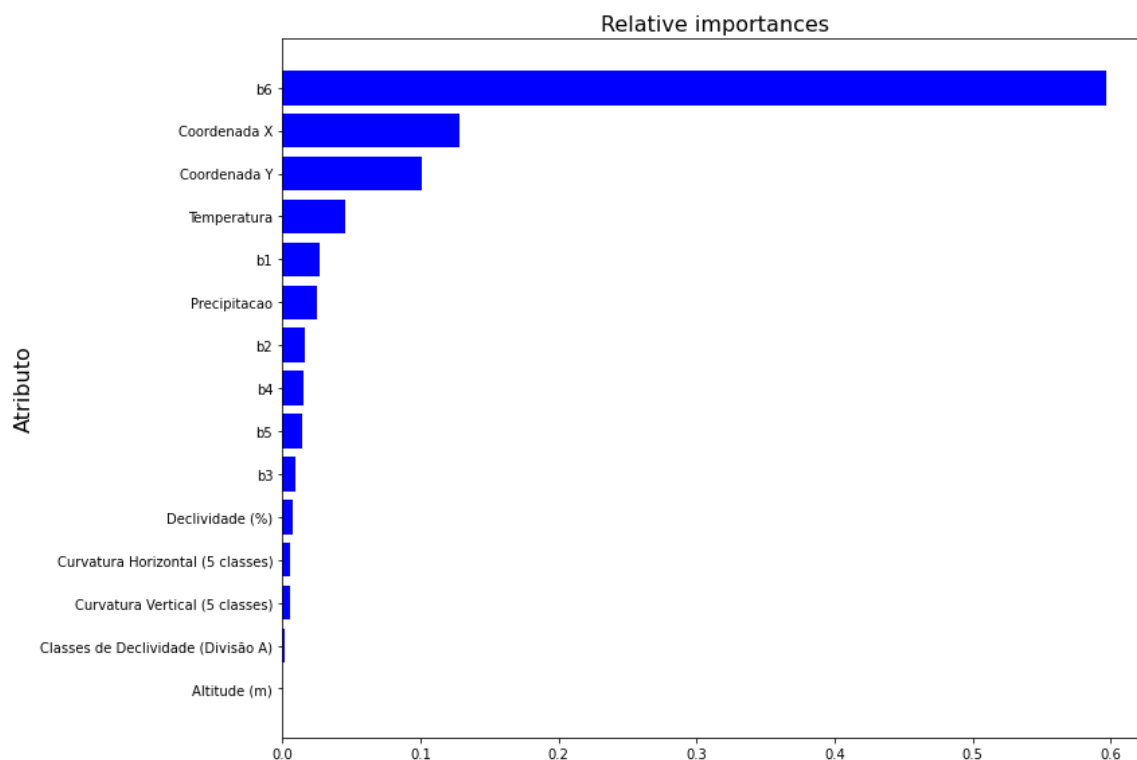
Fonte: O autor

As “Top 5” variáveis que mais impactaram o modelo RF (Figura 8) na predição do conteúdo de argila foram a b6 (banda do SWIR 2), coordenada X, coordenada Y, temperatura e b1 (banda do Red). As “Top 5” variáveis que menos impactaram foram declividade, curvatura horizontal, curvatura vertical, classes de declividade e altitude. Tal fato corrobora com as correlações de *Pearson* apresentadas anteriormente, destacando que a banda do SWIR 2 (b6) teve maior importância quando comparada a importância do SWIR 1 (b5). A banda do SWIR

tem sido utilizada para predição do conteúdo de argila em estudos que vão desde o sensoriamento proximal do solo (Coblinski et al., 2021), realizado em bancadas de laboratório, até estudos que o aplicam utilizando-se de técnicas e dados provenientes de sensoriamento remoto (Gomez et al., 2018).

O grande desafio, assim como relatado por Demattê et al. (2020), quando se trata de predições territoriais por meio do sensoriamento remoto trata-se de descobrir o momento exato em que o solo está exposto. Tal fato torna-se ainda mais desafiador nas lavouras de cana-de-açúcar, onde cada vez menos as áreas estão descobertas devido a presença de palha na superfície causada pela colheita de cana crua; cenário este que mudou nas últimas décadas devido a proibição de queima das lavouras para colheita (Franco et al., 2018).

Figura 8. Importâncias relativas das variáveis utilizadas no modelo de predição.



Fonte: O autor

O modelo treinado apresentou um $R^2 = 0,89$ no *dataset* de teste (Figura 9). O erro médio quadrático (RMSE) foi igual a $56,47 \text{ g kg}^{-1}$ no conteúdo de argila predita. Trabalhos da literatura apresentam estudos onde modelos de predição do conteúdo de argila foram avaliados por espectros do VIS-NIR-SWIR em bancadas de laboratório (Terra et al., 2015; Lacerda et al., 2016 e Nawar et al., 2017). Os autores encontraram R^2 variando de 0,79 a 0,93, com RMSE variando de $77,5$ a $96,5 \text{ g kg}^{-1}$.

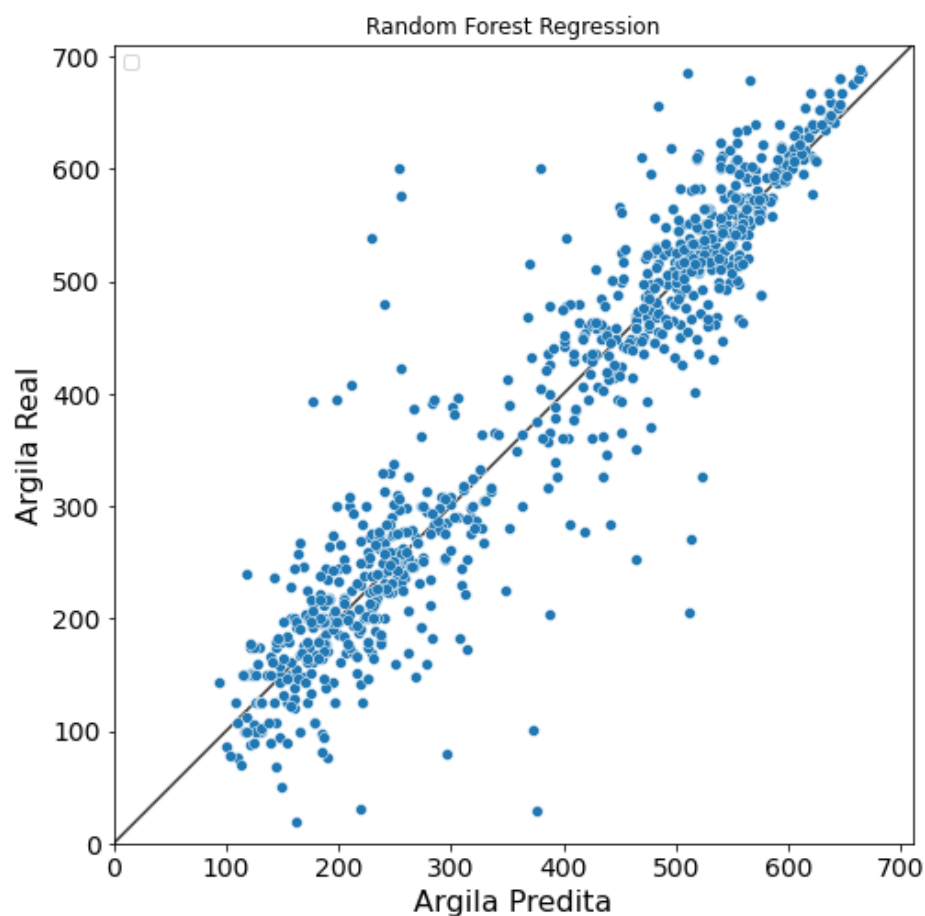
Tal fato mostra a boa performance do modelo calibrado no presente trabalho, onde os erros foram menores quando comparados aos autores mencionados, destacando-se ainda que nenhum deles utilizaram o modelo de RF.

Silva et al. (2019) utilizaram diversos modelos para predição do teor de argila, inclusive o RF. Os autores encontraram $R^2 = 0,68$ e RMSE 116 g kg⁻¹ para o modelo RF. Por outro lado, os mesmos autores concluíram que o modelo *Cubist* performou melhor, apresentando R^2 e RMSE de 0,73 e 106 g kg⁻¹, respectivamente. Apesar dos trabalhos apresentados serem avaliados em estudos de bancada de laboratório, com condições controladas, os achados do presente estudo mostram o grande potencial de utilização do sensoriamento remoto na predição do conteúdo de argila, apresentando bons resultados quando novas variáveis, como localização geográfica e parâmetros do relevo são incluídos no modelo.

Recentemente Silva-Sangoi et al. (2022) encontrou excelentes resultados na predição do conteúdo de argila por espectros no VIS-NIR-MIR utilizando diversas técnicas de aprendizado de máquina, onde a técnica de RF e SGD (*Savitzky-Golay with first derivative*) apresentou a melhor performance. Os autores encontraram $R^2 = 0,95$ e RMSE = 10,6 g kg⁻¹. No entanto, destacamos os autores avaliaram uma quantidade restrita de amostras (148 amostras), com uma variação de 539 a 732 g kg⁻¹, parâmetros estes bem diferentes dos avaliados no presente estudo.

Assim como apresentado anteriormente, a literatura é escassa de trabalhos que apresentam resultados de predição do conteúdo de argila em larga escala no Brasil utilizando-se de técnicas de sensoriamento remoto. Um dos estudos mais recentes de Bellinasso et al. (2021) traz resultados interessantes no uso de sensoriamento remoto para predição do conteúdo de argila utilizando-se de imagens dos satélites Sentinel-2 e Landsat-8. Os autores encontraram $R^2 = 0,75$ e $0,76$ e RMSE = 71,70 e 70,37 g kg⁻¹ de argila para os satélites Sentinel-2 e Landsat-8, respectivamente. Tal fato reforça, mais uma vez, os bons resultados encontrados no presente estudo.

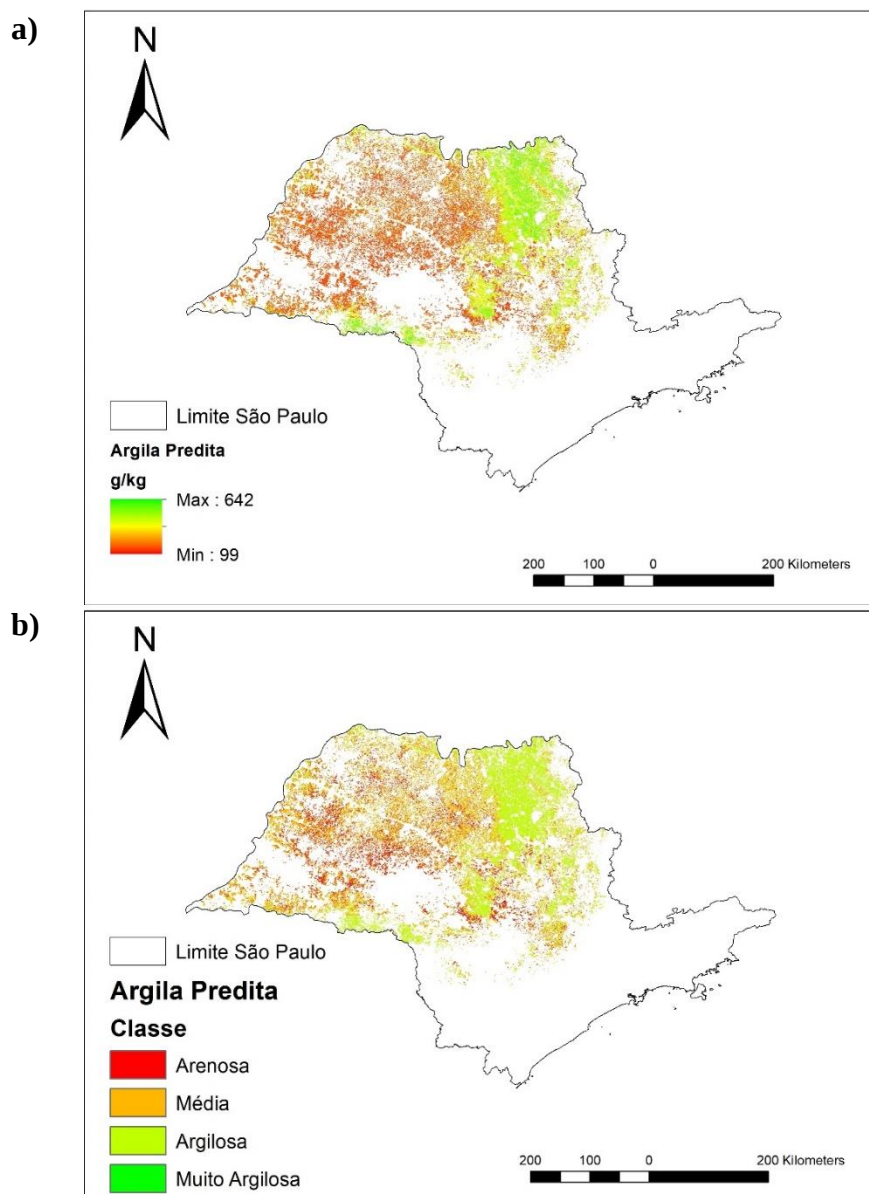
Figura 9. *Scatterplot* entre o valor real e predito do conteúdo de argila (g kg^{-1}) da base de dados de teste.



Fonte: O autor

Por fim, o modelo calibrado foi extrapolado para as áreas de cana-de-açúcar presentes no estado de São Paulo (Figura 10). O mapa predito apresentou conteúdo de argila variando de 99 a 642 g kg^{-1} (Figura 10 – a). As classes de textura arenosa (16,79%) e média (47,31%) contemplam mais de 60% das áreas, sendo a classe argilosa presente em 35,87% das lavouras e a classe muito argilosa em apenas 0,03% das áreas. A classe muito argilosa apresentou baixa porcentagem nas lavouras de cana-de-açúcar. No entanto, cabe destacar que apenas 8% das amostras (255 amostras) do banco de dados eram dessa presente classe, justificando assim uma possível baixa predição dessa classe no estado de São Paulo.

Figura 10. Conteúdo de Argila (g kg^{-1}) (a) e classes texturais (b) preditas nas lavouras de cana-de-açúcar do estado de São Paulo.



Fonte: O autor

Para atender as demandas previstas de produção de etanol e açúcar das próximas décadas (EPE, 2018), a região sudeste brasileira (especialmente o estado de São Paulo) deverá continuar sendo o principal polo de produção, uma vez que as demais regiões brasileiras sofrem com a falta de investimentos em novas pesquisas e tecnologias para aumentar sua capacidade produtiva. Do ponto de vista agrônomo, duas alternativas são possíveis. Se por um lado a expansão das lavouras de cana-de-açúcar em novas áreas é possível, o aumento da produtividade agrícola também é uma alternativa. Apesar das muitas controvérsias sobre a expansão dos canaviais e seu impacto na produção de alimentos (Popp et al., 2014), estudos

mostraram que ainda há potencial para a expansão da lavoura, especialmente em campos de pastagens (Loarie et al., 2011). No entanto, uma alternativa mais sustentável do ponto de vista econômico e ambiental é aumentar a produtividade da cultura.

Para isso, será necessário investir em tecnologias que permitam produzir mais em uma mesma área, ou seja, a produtividade da cultura deve atingir novos patamares e superar a atual média brasileira de $\sim 73 \text{ Mg ha}^{-1}$ (CONAB, 2022).

Desta forma, o presente estudo mostra uma excelente alternativa para auxílio na expansão e aumento da produtividade de cana-de-açúcar. Empresas e produtores poderão se beneficiar do presente estudo na prospecção de novas áreas, seja para arrendamento ou aquisição. Expandir a produção sobre áreas com maior conteúdo de argila significa garantia de melhores e maiores produtividades agrícolas, uma vez que o conteúdo de argila está totalmente ligado ao potencial produtivo das lavouras.

Assim, a ciência de dados mostra o enorme potencial de auxiliar governo, produtores e empresarias a serem mais assertivos e estratégicos nas escolhas de novas áreas de produção, alcançando maior sustentabilidade do processo produtivo.

4.3 Considerações Finais

Neste capítulo foram apresentados e discutidos os resultados obtidos a partir da base de dados analisada. Foram comparadas a implementação usando RF com os resultados apresentados na literatura. O próximo capítulo apresenta as conclusões do trabalho.

5 CONCLUSÃO

5.1 Contribuições

O presente trabalho alcançou um modelo capaz de estimar o teor de argila no estado de São Paulo com $R^2 = 0,89$ e um RMSE de $56,47 \text{ g kg}^{-1}$. Os resultados encontrados apresentam grande potencial para auxiliar novas políticas públicas, empresas e produtores a serem mais estratégicos e assertivos na expansão da produção de cana-de-açúcar de suas lavouras. Expandir a produção em áreas mais argilosas permitirá alcançar produtividades mais elevadas, reduzindo assim a necessidade de abertura de novas fronteiras. As metas de aumento de produção de etanol e açúcar, bem como as demandas por descarbonização, poderão ser beneficiadas dessa expansão estratégica.

5.2 Dificuldades Encontradas e Limitações

Durante o trabalho a principal dificuldade encontrada foi a extrapolação dos resultados do modelo calibrado para as áreas de cana-de-açúcar do estado de São Paulo. Neste ponto, a memória RAM da máquina utilizada (32Gb) não foi capaz de ler o *dataset* de pontos, que continha 149 milhões de registros. Desta forma, foi necessário adquirir uma máquina virtual por meio do Google Colab com 52Gb de memória. Por fim, mesmo com a máquina do Google Colab, foi necessário dividir o *dataset* em 4 partes iguais, permitindo a escalabilidade do modelo para predizer o teor de argila.

5.3 Trabalhos Futuros

Para projetos futuros sugere-se endereçar novos modelos de aprendizado de máquina além do RF, avaliando a capacidade de melhorar os resultados encontrados aqui. Ainda, novas variáveis poderão ser avaliadas, como índices de vegetação. Parâmetros do relevo não utilizados aqui, como divisores e talvegues, poderão ser avaliados também buscando melhorar a performance dos resultados.

REFERÊNCIAS

- Abdel-Rahman, E.M., Ahmed, F.B., 2008. The application of remote sensing techniques to sugarcane (*Saccharum* spp. hybrid) production: a review of the literature. *Int. J. Remote Sens.* 29, 3753-3767.
- Aguilera, E., Diego, A.; de Souza, Z.M., Tormena, C.A., Lovera, L. H., de Souza Lima, E., de Oliveira, I.N., de Paula Ribeiro, N. (2019). Soil compaction, root system and productivity of sugarcane under different row spacing and controlled traffic at harvest. *Soil & Tillage Research* 187, 60-71.
- Arrouays, D., Lagacherie, P., Hartemink, A.E. 2017. Digital soil mapping across the globe. *Geoderma Regional* 9, 1-4.
- Bellinaso, H., Silvero, N.E., Ruiz, L.F.C., Accorsi Amorim, M.T., Rosin, N.A., de Sousa Mendes, W., de Sousa, G.P.B, Sepulveda, L.M.A., de Queiroz, L.G., Nanni, M.R., Demattê, J.A. 2021. Clay content prediction using spectra data collected from the ground to space platforms in a smallholder tropical area. *Geoderma*, 399, Article 115116
- Bocca, F.F., Rodrigues, L.H.A., 2016. The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. *Computers and Electronics in Agriculture* 128, 67-76.
- Bocca, F.F., Rodrigues, L.H.A., Arraes, N.A.M., 2015. When do I want to know and why? Different demands on sugarcane yield predictions. *Agricultural Systems* 135, 48-56.
- Bordonal, R.O., Carvalho, J.L.N., Lal, R., Figueiredo, E.B., Oliveira, B.G., La Scala Jr., N. 2018. Sustainability of sugarcane production in Brazil. A review. *Agronomy for Sustainable Development* 38, 13.
- Borghi, E., Avanzi, J.C., Bortolon, L., Luchiari Jr., A., Bortolon, E.S.O. 2016. Adoption and Use of Precision Agriculture in Brazil: Perception of Growers and Service Dealership. *Journal of Agricultural Science* 8(11), 89-104.
- Bouma, J., Stoorvogel, J., van Alphen, B. J., Booltink, W. G. 1999. Pedology, Precision Agriculture, and the Changing Paradigm of Agricultural Research. *Soil Science Society of America* 63(6), 1763-1768.
- Bramley, B.G.V., Hill, P.A., Thorburn, P.J., Kroon, F.J., Panten, K. 2008. Precision agriculture for improved environmental outcomes: Some Australian perspectives. *Agriculture and Forestry Research* 3, 161-178.

- Brasil. Instituto Nacional de Pesquisas Espaciais (INPE). Topodata: banco de dados geomorfométricos do Brasil. Variáveis geomorfométricas locais. São José dos Campos, 2008. <<http://www.dsr.inpe.br/topodata/>>.
- Breiman, L., 2001. Random forests. *Machine Learning* 45, 5-32.
- Brubaker, S.C., Jones, A.J., Frank, K., Lewis, D.T. 1994. Regression models for estimating soil properties by landscape position. *Soil Sci. Soc. Am. J.* 58 (6), 550 1763–1767.
- Bullock, D.S., Kitchen, N., Bullock, D.G. 2007. Multidisciplinary Teams: A Necessity for Research in Precision Agriculture Systems. *Crop Science* 47(5), 1765.
- Cerri, D.G.P., Magalhães, P.S.G., 2012. Correlation of physical and chemical attributes of soil with sugarcane yield. *Pesquisa Agropecuária Brasileira* 47, 613-620.
- Chung, S.O., Sung, J.H., Sudduth, K.A., Drummond, S.T., Hyun, B.K. 2001. Spatial variability of yield, chlorophyll content, and soil properties in a Korean rice paddy field. In P.C. Robert et al (ed.) *Proc. 5th International Conf. on precision agriculture*. [CD-rom] ASA, CSSA, Madison, WI.
- Coblinski, J.A., Inda, A.V., Demattê, J.A.M., Dotto, A.C., Gholizadeh, A.; Giasson, E. 2021. Identification of minerals in subtropical soils with different textural classes by VIS-NIR-SWIR reflectance spectroscopy. *Catena*. doi:10.1016/j.catena.2021.105334
- Coelho, A.M., Inamassu, R.Y., Vianna, P.A., Matoso, M.J., Ceccon, G. 2007. *Rede de conhecimento em agricultura de precisão para condições de Cerrado e Campos Gerais*. Embrapa Milho e Sorgo, Sete Lagoas, 166p.
- Coelho, F. F., Giasson, E.; Campos, A. R., Tiecher, T., Costa, J. J. F., Coblinski, J. A. 2021. Digital soil class mapping in Brazil: a systematic review. *Scientia Agricola* 78(5). doi: 10.1590/1678-992X-2019-0227
- Companhia Nacional do Abastecimento (CONAB). 2021. *Acompanhamento da safra Brasileira. Cana-de-açúcar, SAFRA 2021/22 Terceiro Levantamento Dezembro/2021*. Brasília, DF, 19p.
- Companhia Nacional do Abastecimento (CONAB). 2022. *Acompanhamento da safra Brasileira. Cana-de-açúcar, SAFRA 2021/22 Quarto Levantamento Abril/2022*. Brasília, DF, 60p.
- Demattê, J. A. M.; Fongaro, C. T.; Rizzo, R.; Safanelli, J. L. Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. *Remote Sensing of Environment*, Amsterdam, v. 212, p. 161-175, 2018.

- Demattê, J.A.M., Ben-Dor, E., Chabrillat, S., Taylor, G., Hill, J., Whiting, M., Sommer, S. 2009. Using imaging spectroscopy to study soil properties. *Remote Sensing of Environment* 113, 38-55.
- Demattê, J.A.M., Poppiel, R.R., Silvero, N.E.Q., Bellinaso, H. 2020. A importância das geotecnologias na caracterização do meio físico com vistas ao aumento da produtividade. *Nutrição de Plantas Ciência e Tecnologia (NPCT)*, v.6.
- Dias L.C.P., Pimenta, F.M., Santos, A.B., Costa, M.H., Ladle, R.J. 2016. Patterns of land use, extensification, and intensification of Brazilian agriculture. *Global Change Biology* 22(8), 2887–903.
- Doorenbos, J.; Kassam, A.H. *Efeito da água no rendimento das culturas*. Campina Grande: UFPB. 1994. 306p. (Estudos FAO: Irrigação e Drenagem, 33).
- Empresa de Pesquisa Energética (EPE). 2018. *Cenários de oferta de etanol e demanda de ciclo otto 2018-2030*. Ministério de Minas e Energia, Rio de Janeiro. 41 p.
- Everingham Y.L., Lowe, K.H., Donald, D.A., Coomans, D.H., Markley, J. 2007. Advanced satellite imagery to classify sugarcane crop characteristics. *Agronomy Sustainable Development* 27(2), 111-117.
- Everingham Y.L., Sexton, J., Skocaj, D., Inman-Bamber, G., 2016. Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy Sustainable Development* 36.
- Food and Agriculture Data (FAOSTAT), Food and Agriculture Organization of the United Nations (FAO). Available at: <http://www.fao.org/faostat/>.
- Food and Agriculture Organization. OECD-FAO. *Agricultural Outlook 2015-2024*. 11 July 2014. Pg. 148. Available at: <http://www.fao.org/3/a-i4738e.pdf>. Access: 14/02/2016.
- Franco, H.C.J., Castro, S.G.Q., Sanches, G.M., Kölln, O.T., Bordonal, R.O., Borges, B.M.M.N., Borges, C.D. 2018. *Alternatives to Increase the Sustainability of Sugarcane Production in Brazil under High Intensive Mechanization*. In. Singh, P., Tiwari, A.K. *Sustainable Sugarcane Production*. Apple Academic Press – 426 p.
- Fukuda S., Spreer, W., Yasunaga, E., Yuge, K., Sardud, V., Müller, J., 2013. Random Forests modelling for the estimation of mango (*Mangifera indica* L. cv. Chok Anan) fruit yields under different irrigation regimes. *Agricultural Water Management* 116, 142-150.
- Fulton, J.P., Wells, L.G., Shearer, S.A., Barnhisel, R.I. 1996. Spatial variation of soil physical properties: Aprecursor to precision tillage. ASAE Paper No. 961002. International Meeting, Phoenix, Arizona. 14–18 July 1996. ASAE, St. Joseph, MI.

- Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., Michaelsen, J., 2015. The climate hazards infrared precipitation with Stations a new environmental record for monitoring extremes. *Scientific data* 2, 1-21.
- Gaston, L.A., Locke, M.A., Zablotowicz, R.M., Reddy, K.N. 2001. Spatial variability of soil properties and weed populations in the Mississippi delta. *Soil Science Society America Journal* 65, 449–459.
- Gobin, A., Campling, P., Feyen, J. 2001. Soil-Landscape Modelling to Quantify Spatial Variability of Soil Texture. *Phys. Chem. Earth* 26(1), 41-45.
- Gomez, C., Adeline, K., Bacha, S., Driessen, B., Gorretta, N., Lagacherie, P., Roger, J.M., Briottete, X. 2018. Sensitivity of clay content prediction to spectral configuration of VNIR/SWIR imaging data, from multispectral to hyperspectral scenarios. *Remote Sensing of Environment* 204:18-30.
- Helfer, G.A., Barbosa, J.L.V., Alves, D., da Costa, A.B., Beko, M., Leithard, V.R.Q. 2021. Multispectral Cameras and Machine Learning Integrated into Portable Devices as Clay Prediction Technology. *Journal of Sensor and Actuator Networks*, 10. doi: 10.3390/jsan1003004.
- Intended nationally determined contribution (iNDC) towards achieving the objective of the United Nations Framework Convention on Climate Change. 2015. Available at: http://www.itamaraty.gov.br/images/ed_desenvsust/BRAZILiNDC-english.pdf. Date: 09/11/2019.
- Jenny, H. 1941. *Factors of soil formation*. McGraw-Hill. New York.
- Khoshgoftaar, T.M., Golawala, M., Hulse, J. Van, 2007. An Empirical Study of Learning from Imbalanced Data Using Random Forest. *19th IEEE Int. Conf. Tools with Artif. Intell.* 310–317.
- Lacerda, M.P.C., Demattê, J.A.M., Sato, M.V., Fongaro, C.T., Gallo, B.C., Souza, A.B. 2016. Tropical texture determination by proximal sensing using a regional spectral library and its relationship with soil classification. *Remote Sensing* 8 (9) p. 701.
- Lebourgeois, V., Dupuy, S., Vintrou, É., Ameline, M., Butler, S., Bégué, A., 2017. A Combined Random Forest and OBIA Classification Scheme for Mapping Smallholder Agriculture at Different Nomenclature Levels Using Multisource Data (Simulated Sentinel-2 Time Series, VHRS and DEM). *Remote Sensing* 9, 259.
- Loarie, S.R., Lobell, D.B., Asner, G.P., Mu, Q., Field, C.B. (2011). Direct impacts on local climate of sugarcane expansion in Brazil. *Nature Climate Change* 1, 105-109.
- Macedo, I.D.C., Nassar, A.M., Cowie, A.L., Seabra, J.E.A., Marelli, L., Otto, M.Wang,

- M.Q.Tyner, W.E. 2015. Greenhouse gas emissions from bioenergy. *In: Bioenergy & Sustainability: Bridging the Gaps* (eds Souza GM, Victoria RL, Joly CA, Verdade LM), pp. 582–616. SCOPE, São Paulo.
- Machado, P.L.O.A., Silva, C.A., Bernardi, A.C.C. *et al.* 2004. Variabilidade de atributos de fertilidade e espacialização da recomendação de adubação e calagem para a soja. *In: MACHADO, P.L.O.A.; SILVA, C.A.; BERNARDI, A.C.C. (Eds.) Agricultura de precisão para o manejo da fertilidade do solo em sistema de plantio direto*. Rio de Janeiro: Embrapa Solos. p.115-129.
- Magalhães, P.S.G., Cerri, D.G.P., 2007. Yield Monitoring of Sugar Cane. *Biosystems Engineering* 96, 1–6.
- Manzatto, C.V., Assad, E.D., Baca, J.F.M., Zaroni, M.J., Pereira, S.E.M. 2009. Zoneamento agroecológico da cana-de-açúcar: expandir a produção, preservar a vida, garantir o future. Documentos 110. Embrapa Solos, 2009. Available at: <https://ainfo.cnptia.embrapa.br/digital/bitstream/CNPS-2010/14408/1/ZonCana.pdf>
- Marin, F.R., Martha, G.B., Cassman, K.G., Grassini, P. 2016. Prospects for increasing sugarcane and bioethanol production on existing crop area in Brazil. *Bio Science* 66:307–316.
- Mashalaba, L., Galleguillos, M., Seguel, O., Poblete-Olivares, J. 2020. Predicting spatial variability of selected soil properties using digital soil mapping in a rainfed vineyard of central Chile. *Geoderma Regional* 22. doi: 10.1016/j.geodrs.2020.e00289.
- McBratney, A., Whelan, B., Ancev, T., 2006. Future directions of precision agriculture. *Precision Agriculture* 6, 7-23.
- Minasny, B.; Mcbratney, A.B. 2016. Digital soil mapping: a brief history and some lessons. *Geoderma* 264, 301-311.
- Moore, I. D., Gessler, P. E., Nielsen, G.A., Peterson, G.A. 1993. Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal* 57 (2), 443-452.
- Nass, L.L., Pereira, P.A.A., Ellis, D. (2007). Biofuels in Brazil: An overview. *Crop Science* 47, 2228–2237.
- Nawar, S., Mouazen, A.M. 2017. Predictive performance of mobile vis-near infrared spectroscopy for key soil properties at different geographical scales by using spiking and data mining techniques. *Catena*, 151:118-129.
- Newlands, N.K., Zamar, D.S., Kouadio, L.A., Zhang, Y., Chipanshi, A., Potgieter, A., Toure, S., Hill, H.S.J., 2014. An integrated, probabilistic model for improved seasonal forecasting

- of agricultural crop yield under environmental uncertainty. *Frontiers in Environmental Science*. doi:10.3389/fenvs.2014.00017.
- Oliveira, D.M.S., Cherubin, M.R., Franco, A.L.C., Santos, A.S., Gelain, J.G., Dias, N.M.S., Diniz, T.R., Almeida, A.N., Feigl, B.J., Davies, C.A., Paustian, K., Karlen, D.L., Smith, P., Cerri, C.C., Cerri, C.E.P. 2019. Is the expansion of sugarcane over pasturelands a sustainable strategy for Brazil's bioenergy industry? *Renewable and Sustainable Energy Reviews* 102, 346-355.
- Organização das Nações Unidas. 2021. *População mundial deve chegar a 9,7 bilhões de pessoas em 2050, diz relatório da ONU*. Disponível em: <https://brasil.un.org/pt-br/83427-populacao-mundial-deve-chegar-97-bilhoes-de-pessoas-em-2050-diz-relatorio-da-onu>. Acesso em: 01/10/2021.
- Parente, L., Ferreira, L., Faria, A., Nogueira, S., Araújo, F., Teixeira, L., Hagen, S., 2017. Monitoring the brazilian pasturelands: A new mapping approach based on the landsat 8 spectral and temporal domains. *International Journal of Applied Earth Observation and Geoinformation* 62, 135–143.
- Peets, S., Mouazen, A.M., Blackburn, K., Boyan, K., Wiebensohn, J. 2012. Methods and procedures for automatic collection and management of data acquired from on-the-go sensors with application to on-the-go soil sensors. *Computers and Electronics in Agriculture* 81, 104 – 112.
- Pinheiro, H.S.K., Carvalho Junior, W., Chagas, C.S., dos Anjos, L.H.C., Owens, P.R. 2017. Prediction of Topsoil Texture Through Regression Trees and Multiple Linear Regressions. *Revista Brasileira de Ciência do Solo*, e0170167. doi: 10.1590/18069657rbcs20170167.
- Popp, J., Lakner, Z., Harangi-Rákos, M., Fári, M. (2014). The effect of bioenergy expansion: food, energy, and environment. *Renewable and Sustainable Energy Reviews* 32, 559-578.
- Rabus, B., Eineder, M., Roth, R., Bamler, R. 2003. The Shuttle Radar Topography Mission—a new class of digital elevation models acquired by spaceborne radar. *ISPRS Journal of Photogrammetry & Remote Sensing* 57, 241-262.
- Raij B, Andrade JC, Cantarella H, Quaggio JA (2001) Análise química para avaliação da fertilidade de solos tropicais. Campinas, Instituto Agronômico.
- Renewable Fuels Association (RFA). 2020. Markets and Statistics. Available at: <https://ethanolrfa.org/statistics/>.
- Resende, A.V., Shiratsuchi, L.S., Sena, M.C., Krah, L.L., Oliveira, J.V.F., Corrêa, R.F., Oro, T. 2006. Grades amostrais para fins de mapeamento da fertilidade do solo em área do

- cerrado. In : Congresso Brasileiro de Agricultura de Precisão, 2, São Pedro-Piracicaba. *Anais. USP_ESALQ*.
- Rezende, S. O. 2003. *Sistemas inteligentes: fundamentos e aplicações*. Editora Manole Ltda. 550 p.
- Sanches, G.M., Magalhães, P.S.G., Remacre, A.Z., Franco, H.C.J. 2018. Potential of apparent soil electrical conductivity to describe the soil pH and improve lime application in a clayey soil. *Soil & Tillage Research* 175, 217-225.
- Sanches, G.M., Otto, R. 2022. A novel approach for determining nitrogen requirement based on a new agronomic principle—sugarcane as a crop model. *Plant Soil*. doi: 10.1007/s11104-021-05263-7.
- Sanches, G.M.; Franco, H.C.J.; Magalhães, P.S.G. 2019. Site-specific assessment of spatial and temporal variability of sugarcane yield related to soil attributes. *Geoderma* 334, 90-98.
- Sanches, G.M.; Magalhães, P.S.G.; Kolln, O.T.; Otto, R.; Rodrigues Jr., F.; Cardoso, T.F.; Chagas, M.F.; Franco, H.C.J. 2021. Agronomic, economic, and environmental assessment of site-specific fertilizer management of Brazilian sugarcane fields. *Geoderma Regional* (24)e00360.
- Sanchez, R.B., Marques Júnior, J., de Souza, Z.M., Pereira, G.T., Martins Filho, M.V. 2009. Spatial variability of soil attributes and of erosion factors in different landforms. *Bragantia* 68 (4), 1095-1103.
- Santos Rodrigues, M. A. 2004. *Árvores de Classificação*. Monografia. Universidade dos Açores – Departamento de Matemática. 34 p.
- Silva, C.B., Moraes, M.A.F.D., Molin, J.P. 2011. Adoption and use of precision agriculture technologies in the sugarcane industry of São Paulo state, Brazil. *Precision Agriculture* 12(1), 67–81.
- Silva, E.B., Giasson, É., Dotto, A.C., ten Caten, A., Demattê, J.A.M., Bacic, I.L.Z., da Veiga M. 2019. A Regional Legacy Soil Dataset for Prediction of Sand and Clay Content with Vis-Nir-Swir, in Southern Brazil. *Rev. Bras. Cienc. Solo* 43, Article e0180174.
- Silva-Sangoi, D.V., Horst, T.Z., Moura-Bueno, J.M., Dalmolin, R.S.D., Sebem, E., Gecler, L., Santos, M.S. 2022. Soil organic matter and clay predictions by laboratory spectroscopy: Data spatial correlation. *Geoderma Regional* 28 Article e00486.
- Simbahan, G.C., Dobermann, A., Goovaerts, P., Ping, J., Haddix, L. 2006. Fine-resolution mapping of soil organic carbon based on multivariate secondary data. *Geoderma* 132, 471-489.

- Sousa, A.C.M., Farhate, C.V.V., de Souza, Z.M., Torres, J.L.R., da Silva, R.B. 2018. Soil Load-Bearing Capacity and Development of Root System in Area Under Sugarcane with Traffic Control in Brazil. *Sugar Tech* 20, 1-9.
- Souza, Z.M. de, Souza, G.S. de, Marques Júnior, J., Pereira, G.T. 2014 Número de amostras na análise geoestatística e na krigagem de mapas de atributos do solo. *Ciência Rural* 44, 261-268.
- Terra, F.S., Demattê, J.A.M., Rossel, R.A.V. 2015. Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis-NIR and mid-IR reflectance data. *Geoderma*. 255-256:81-93
- Tulbure, M.G., Wimberly, M.C., Boe, A., Owens, V.N., 2012. Climatic and genetic controls of yields of switchgrass, a model bioenergy species. *Agriculture, Ecosystems & Environment* 146(1), 121-129.
- Valeriano, M.M., Rosseti, D.F. 2008. TOPODATA: seleção de coeficientes geoestatísticos para o refinamento unificado de dados SRTM. São José dos Campos: *INPE*, 50p.
- Valeriano, M.M., Rosseti, D.F. 2008. TOPODATA: seleção de coeficientes geoestatísticos para o refinamento unificado de dados SRTM. São José dos Campos: *INPE*, 50p.
- Valeriano, M.M., Rossetti, D.F. 2012. Topodata: Brazilian full coverage refinement of SRTM data. *Applied Geography* 32, 300-309.
- Waclawovsky, A.J., Sato, P.M., Lembke, C.G., Moore, P.H., Souza, G.M., (2010). Sugarcane for bioenergy production: an assessment of yield and regulation of sucrose content. *Plant Biotechnology Journal* 8, 263-276.
- Webster, R., Oliver, M. A. 1992. Sample adequately to estimate variograms of soil properties. *Soil Science* 43(1), 117-192.
- Wei, J., Xiao, D., Zeng, H., Fu, Y. 2008. Spatial variability of soil properties in relation to land use and topography in a typical small watershed of the black soil region, northeastern China. *Environ Geol* 53:1663–1672.
- Wilson, J.P., Gallant, J.C. 2000. *Terrain analysis: principles and applications*. 1st Ed. New York: John Wiley & Sons, Inc. 478p.
- Wolfert, S., Ge, L., Verdouwa, C., Bogaardt, MJ., 2017. Big Data in Smart Farming – A review. *Agricultural Systems* 153, 69-80.
- Yan-e, D., 2011. Design of Intelligent Agriculture Management Information System Based on IoT. In: *Intelligent Computation Technology and Automation (ICICTA)*, Guangdong, Shenzhen. 1:1045-1049.

Zeraatpished, M., Ayoubi, S., Jafari, A., Tajik, S., Finke, P. 2019. Digital mapping of soil properties using multiple machine learning in a semiarid region, central Iran. *Geoderma* 338, 445-452.