UFRRJ

INSTITUTO DE AGRONOMIA

PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA – CIÊNCIA DO SOLO

TESE

Eficiência do autoRA: Algoritmo para Delineamento Automático de Áreas de Referência para Suporte à Amostragem Otimizada de Solos

Hugo Machado Rodrigues

2025

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

EFICIÊNCIA DO AUTORA: UM ALGORITMO PARA DELINEAMENTO AUTOMÁTICO DE ÁREAS DE REFERÊNCIA PARA SUPORTE À AMOSTRAGEM OTIMIZADA DE SOLOS

HUGO MACHADO RODRIGUES

Sob a Orientação do Professor
Marcos Bacis Ceddia

e Coorientação do Pesquisador
Gustavo Mattos Vasques

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

Seropédica, RJ
Fevereiro de 2025

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

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

R696e

Rodrigues, Hugo Machado, 1992-
Eficiência do Autora: Um Algoritmo para delineamento automático de áreas de referência para suporte à amostragem otimizada de solos / Hugo Machado Rodrigues. – Seropédica, 2025.
146 f.: il.

Orientador: Marcos Bacis Ceddia.
Tese (Doutorado). – – Universidade Federal Rural do Rio de Janeiro, Programa de Pós-Graduação em Agronomia Ciência do Solo, 2025.

1. Mapeamento Digital de Solos. 2. Modelagem Preditiva. 3. Estratégias de amostragem. I. Ceddia, Marcos Bacis, 1968-, orient. II Universidade Federal Rural do Rio de Janeiro. Programa de Pós-Graduação em Agronomia-Ciência do Solo III. Título.

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

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA (CIÊNCIAS DO SOLO)

HOMOLOGAÇÃO DE TESE DE DOUTORADO Nº 2 / 2025 - CPGACS (12.28.01.00.00.00.27)

Nº do Protocolo: 23083.012056/2025-04

Seropédica-RJ, 13 de março de 2025.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO

INSTITUTO DE AGRONOMIA

PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA-CIÊNCIA DO SOLO

HUGO MACHADO RODRIGUES

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

TESE APROVADA EM: 17/02/2025.

Marcos Bacis Ceddia. Dr. LASA / DATS / UFRRJ
(Orientador)

Lúcia Helena Cunha dos Anjos. Ph.D. UFRRJ

Helena Saraiva Koenow Pinheiro. Dra. UFRRJ

Diego Nunes Brandão. Dr. CEFET-RJ

Fabricio da Silva Terra. Dr. UFVJM

(Assinado digitalmente em 13/03/2025 13:07)
HELENA SARAIVA KOENOW PINHEIRO
CHEFE DE DEPARTAMENTO
DeptS (12.28.01.00.00.00.33)
Matrícula: 2223668

(Assinado digitalmente em 16/03/2025 14:40)
LUCIA HELENA CUNHA DOS ANJOS
PROFESSOR DO MAGISTERIO SUPERIOR
DeptS (12.28.01.00.00.00.33)
Matrícula: 387335

(Assinado digitalmente em 14/03/2025 11:50)
MARCOS BACIS CEDDIA
PROFESSOR DO MAGISTERIO SUPERIOR
DATS (11.39.00.35)
Matrícula: 1220296

(Assinado digitalmente em 18/03/2025 07:58)
DIEGO NUNES BRANDÃO
ASSINANTE EXTERNO
CPF: 096.083.947-08

(Assinado digitalmente em 13/03/2025 14:56)
FABRÍCIO DA SILVA TERRA
ASSINANTE EXTERNO
CPF: 818.150.690-15

Visualize o documento original em **https://sipac.ufrrj.br/public/documentos/index.jsp** informando seu número: **2**, ano: **2025**, tipo: **HOMOLOGAÇÃO DE TESE DE DOUTORADO**, data de emissão: **13/03/2025** e o código de verificação: **64dc4ec292**

DEDICATÓRIA

Dedicado à minha esposa.

Dedicated to my wife.

AGRADECIMENTOS

Gostaria de expressar minha admiração e gratidão à minha companheira Carolina, que, assim como eu, acredita no caminho acadêmico como profissão para o nosso futuro há 14 anos. Deixo claro que sem o apoio emocional dela para ouvir e sugerir soluções para os desafios da pós-graduação e da vida adulta, essas palavras que apresentam esta tese de doutorado não estariam digitadas. Além disso, deixo claro que o nome do algoritmo autoRA é em sua homenagem.

Estendo meu agradecimento ao meu querido amigo Matheus. Seu constante incentivo, bondade e apoio caloroso foram fundamentais para me ajudar a lidar com o estresse da pós-graduação e da vida. Obrigado, meu amigo! Não poderia deixar de mencionar meus pais, Jorge e Wanda, e aos meus sogros Ary e Alizete, que sempre me apoiaram desde o meu bacharelado, mestrado e doutorado. É uma jornada árdua, e vocês a tornaram mais fácil para mim. Obrigado!

Expresso minha mais profunda gratidão ao meu estimado professor, Prof. Marcos Bacis Ceddia, por sua inspiração, orientação e conselhos inestimáveis ao longo de minha tese de doutorado. Seu apoio, compreensão e flexibilidade inabaláveis desempenharam um papel fundamental em me ajudar a superar os inevitáveis altos e baixos da vida na pós-graduação e alcançar meus objetivos acadêmicos. Ele me orientou desde 2018 e me mostrou como um pesquisador responsável deve lidar com a atmosfera acadêmica. Além disso, sua maneira dura de lidar com as situações buscando ser empático o máximo possível me ensinou a ser humilde em um cenário acadêmico onde às vezes o ego exagera em nossas decisões e pensamentos.

Gostaria também de estender minha sincera gratidão ao coorientador que me orienta desde o meu bacharelado, Dr. Gustavo Vasques. Obrigado por me ajudar na jornada da Ciência do Solo e do Meio Ambiente. Eu o conheço desde 2014 e, desde então, ele sempre me orientou e me ensinou como fazer o melhor trabalho em relação à análise de dados e redação científica.

À Dra. Sabine Grunwald, professora na Universidade da Flórida que me recebeu durante o período do meu doutorado sanduíche entre 2023 e 2024 e que colaborou imensamente com o desenvolvimento dos testes de sensitividade com o algoritmo autoRA.

Agradeço profundamente à Universidade Federal Rural do Rio de Janeiro (UFRRJ) e ao Programa de Pós-Graduação em Ciências do Solo (PPGA-CS), que me proporcionaram a base acadêmica e científica necessária para alcançar este título. Expresso minha gratidão à minha banca examinadora, cujas contribuições foram essenciais para o aprimoramento deste trabalho. Também reconheço o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), cuja concessão da bolsa de doutorado viabilizou minha dedicação integral à pesquisa. Sem essas instituições e pessoas, este percurso não teria sido possível.

Quero reconhecer o generoso apoio financeiro que a CAPES oferece, incluindo bolsas de pós-graduação e bolsas de doutorado sanduíche no exterior. Sem esse financiamento eu não poderia ter um Doutorado em Ciência do Solo. Aprecio a existência da CAPES para o bem dos futuros cientistas do Brasil.

ACKNOWLEDGMENT

I want to express my admiration and gratitude to my companion Carolina, who, like me, has believed in the academic path as a profession for our future for 14 years. I make it clear that without her emotional support to listen and suggest solutions to graduate school and adult life challenges, these words that present this doctoral thesis would not be typed. Also, I make it clear that the name of the autoRA algorithm is in his honor.

I am grateful to extend my appreciation to my dear friend Matheus. His constant encouragement, kindness, and warm support have been instrumental in helping me navigate the stresses of graduate school and life. Thanks, my friend!

I thank my parents, Jorge and Wanda, and my parents-in-law Ary and Alizete, who always supported me, which happens in parallel with my bachelor's, master's, and doctorate time. It is an arduous journey, and you made it easier for me. Thank you!

I express my deepest gratitude to my esteemed professor, Prof. Marcos Bacis Ceddia, for his inspiration, guidance, and invaluable advice throughout my doctoral thesis. Their unwavering support, understanding, and flexibility have played a key role in helping me overcome the inevitable ups and downs of graduate school life and achieve my academic goals. He guided me since 2018 and has shown me how a responsible researcher must handle the academic atmosphere. Also, his tough way of handling situations, while he tries to wear the other's shoes, taught me how to be humble in an academic scenario where sometimes the ego overplays our decisions and thoughts.

I would also like to extend my sincere gratitude to the co-advisor who has been guiding me since my Bachelor's, Dr. Gustavo Vasques. Thanks for helping me through the journey of Soil and Environmental Science. I've known him since 2014, and since then, he has always guided and taught me how to do the best job regarding data analysis and scientific writing.

I am deeply grateful to the Federal Rural University of Rio de Janeiro (UFRRJ) and the Graduate Program in Soil Sciences (PPGA-CS), which provided me with the academic and scientific basis necessary to achieve this title. I express my gratitude to my examining board, whose contributions were essential for the improvement of this work. I also recognize the support of the Coordination for the Improvement of Higher Education Personnel (CAPES), whose granting of the doctoral scholarship made my full dedication to research possible. Without these institutions and people, this path would not have been possible.

To Dr. Sabine Grunwald, a professor at the University of Florida who received me during the period of my sandwich doctorate between 2023 and 2024 and who collaborated immensely with the development of sensitivity tests with the autoRA algorithm.

I want to acknowledge the generous financial support the Coordination for the CAPES provides, including postgraduate scholarships and sandwich doctoral scholarships abroad. Without this sponsorship, I couldn't have a Ph.D. in Soil Science. I appreciate the existence of CAPES for the good of Brazil's future scientists.

RESUMO

Rodrigues, Hugo Machado. **Eficiência do Autora: Um Algoritmo para delineamento automático de áreas de referência para suporte à amostragem otimizada de solos.** 2025. 146 f. Tese (Doutorado em Agronomia, Ciência do Solo) Instituto de Agronomia, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2025.

O algoritmo autoRA é uma metodologia baseada em dados projetada para delinear Áreas de Referência (ARs) que capturam fatores críticos de formação do solo, aprimorando os fluxos de trabalho de mapeamento digital do solo (MDS). Ao aproveitar o Índice de Dissimilaridade de Gower, o autoRA determina sistematicamente o tamanho ideal da área alvo e as configurações de resolução espacial, equilibrando desempenho preditivo e custo-benefício. Para avaliar sua eficácia, o autoRA foi testado em três cenários distintos. No primeiro cenário, realizado no Rio de Janeiro e na Flórida, examinamos o efeito da resolução espacial no cálculo do Índice de Gower e o impacto do tamanho da área-alvo no desempenho do modelo. O modelo preditivo foi desenvolvido para estimar uma Superfície Teórica Simulada (STS) usando uma única iteração. O Modelo de Área de Referência (MAR) ideal com uma área alvo de 50% e um tamanho de bloco de 10 pixels alcançou valores de Distância Euclidiana (DE) (0,15 no Rio de Janeiro e 0,38 na Flórida), aproximando-se dos resultados da amostragem exaustiva e reduzindo os custos em aproximadamente 61% e 63%, respectivamente. No segundo cenário, o autoRA foi aplicado a uma classificação de unidade de solo já mapeada em Sátiro Dias, Bahia, Brasil. Como este estudo teve uma AR delineada manualmente por um especialista e amostras de solo coletadas de acordo com uma estrutura completa de perfil de solo, a autoRA foi aplicada usando as mesmas covariáveis do especialista. Testamos ARs com 10%, 20%, 30%, 40% e 50%, cruzando os limites das ARs propostas com os locais reais das amostras de solo. As amostras internas foram usadas para treinamento do modelo, enquanto as amostras externas validaram a extrapolação das predições. Com 40% de cobertura de AR, o erro de predição usando autoRA foi menor do que o da AR delineada manualmente. Além disso, mapas de classe de solo gerados manualmente e via autoRA foram comparados com 100 iterações de modelagem Random Forest usando a abordagem MDS convencional, onde os conjuntos de dados foram divididos aleatoriamente (70% de treinamento, 30% de validação). Essa validação iterativa confirmou a robustez das previsões do autoRA. O terceiro cenário explorou diferentes proporções de pixels classificados como baixa e alta Dissimilaridade de Gower. A mesma abordagem de treinamento interno e validação externa foi aplicada. Os resultados demonstraram que o foco em regiões de alta dissimilaridade permitiu a redução das áreas amostradas, mantendo a precisão preditiva. A abordagem de Área Total (AT) também foi testada com tamanhos de amostra variando de 100 a 1.000. O modelo AT desenvolvido usando 100 iterações de divisões de 70% a 30% mostrou que a redução do tamanho da amostra resultou em uma melhoria de 20% na DE em comparação com o benchmark de 1.000 amostras. O conjunto de dados de 800 amostras foi identificado como a referência ideal para modelagem de AT. Para testar o efeito do mosaico de áreas de alta e baixa dissimilaridade, mantivemos o limite de melhoria de 20% no DE, usando 800 como uma nova referência. Os resultados demonstraram que um conjunto de dados de 600 amostras focado no RA de 40% delineado pelo autoRA produziu métricas de ED comparáveis ao modelo previsto do AT-STS. A análise de sensibilidade usando simulações STS permitiu testes extensivos de configurações de parâmetros, confirmando que as configurações de autoRA mais eficazes priorizam regiões de alta dissimilaridade. Tão importante quanto determinar o número de pontos para uma previsão precisa é decidir onde colocá-los. A abordagem autoRA responde a essa pergunta simulando várias configurações, fornecendo expectativas de precisão e custo, honrando o conhecimento da dissimilaridade da paisagem e evitando amostragem redundante.

Palavras-chave: Mapeamento Digital de Solos. Modelagem Preditiva. Estratégias de amostragem.

ABSTRACT

Rodrigues, Hugo Machado. **Efficacy of autoRA: Algorithm for automatic delineation of reference areas to support optimized soil sampling.** 2025. 146p. Thesis (Doctorate in Agronomy, Soil Science). Instituto de Agronomia, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2025.

The autoRA algorithm is a data-driven methodology designed to delineate Reference Areas (RAs) that capture critical soil-forming factors, enhancing digital soil mapping (DSM) workflows. By leveraging Gower's Dissimilarity Index, autoRA systematically determines optimal target area size and spatial resolution configurations, balancing predictive performance and cost-effectiveness. To evaluate its efficacy, autoRA was tested in three distinct scenarios. In the first scenario, conducted in Rio de Janeiro and Florida, we examined the effect of spatial resolution in calculating Gower's Index and the impact of target area size on model performance. The predictive model was developed to estimate a Simulated Theoretical Surface (STS) using a single iteration. The optimal Reference Area Model (RAM) with a 50% target area and a 10-pixel block size achieved Euclidean Distance (ED) values (0.15 in Rio de Janeiro and 0.38 in Florida), closely approximating results from exhaustive sampling while reducing costs by approximately 61% and 63%, respectively. In the second scenario, autoRA was applied to an already mapped soil unit classification in Sátiro Dias, Bahia, Brazil. As this study had an RA manually delineated by a specialist and soil samples collected according to a complete soil profile framework, autoRA was applied using the same covariates as the specialist. We tested RAs at 10%, 20%, 30%, 40%, and 50%, intersecting the proposed RAs with the actual sample locations. The inner samples were used for model training, while the outer samples validated the extrapolation of predictions. At 40% RA coverage, the prediction error using autoRA was lower than that of manually delineated RA. Additionally, manual and autoRA-generated RA soil class maps were compared against 100 iterations of Random Forest modeling using the conventional DSM approach, where datasets were randomly split (70% training, 30% validation). This iterative validation confirmed the robustness of autoRA's predictions. The third scenario explored different proportions of pixels classified as low or high Gower's Dissimilarity. The same inner-training, outer-validation approach was applied. The results demonstrated that focusing on high-dissimilarity regions allowed for the reduction of sampled areas while maintaining predictive accuracy. The Total Area (TA) approach was also tested with sample sizes ranging from 100 to 1,000. The TA model developed using 100 iterations of 70%-30% splits showed that reducing the sample size resulted in a 20% ED improvement compared to the 1,000-sample benchmark. The 800-sample dataset was identified as the optimal benchmark for TA modeling. To test the effect of mosaicking high- and low-dissimilarity areas, we maintained the 20% ED improvement threshold, using 800 as a new benchmark. The results demonstrated that a 600-sample dataset focused within the 40% RA delineated by autoRA produced ED metrics comparable to the TA-STS predicted model. Sensitivity analysis using STS simulations allowed for extensive testing of parameter configurations, ultimately confirming that the most effective autoRA settings prioritize high-dissimilarity regions. As important as determining the number of points for an accurate prediction is deciding where to place them. The autoRA approach answers this question by simulating multiple configurations, providing expectations of accuracy and cost while honoring the knowledge of landscape dissimilarity and avoiding redundant sampling.

Keywords: Digital Soil Mapping. Predictive Modeling. Sampling Strategies.

LIST OF FIGURES

LIST OF TABLES

LISTA DE ABREVIAÇÕES

AOI	Area of Interest
autoRA	Automatic Reference Area algorithm
DSM	Digital Soil Mapping
ED	Euclidean Distance
EPM	Exhaustive Prediction Model
MEA	Mean Absolute Error
MU	Mapping Units
OA	Overall Accuracy
PA	Producer Accuracy
R^2	Adjusted Coefficient of Determination
RA	Reference Area
RAM	Reference Area Model
RI	Region of Interest
RMSE	Root Mean Square Error
STS	Simulated Theoretical Surface
TA	Total Area
UA	User Accuracy
WPAI	Weighted Producer Accuracy Index
WUAI	Weighted UserAccuracy Index

SUMMARY

1. GENERAL INTRODUCTION

A soil map can be understood as a model representing the spatial distribution of soil classes and attributes in a specific area of interest. The creation of this map involves considering several fundamental issues, such as the demands of the project and its target audience, the prediction model to be used, and the costs and expected accuracy. While execution time is a factor to consider, it is primarily a consequence of the tools employed and the complexity of the study area rather than a defining element of the mapping process itself (GRUNWALD, S.; THOMPSON; BOETTINGER, 2011). Over time, the soil mapping process has transformed significantly, driven by the advancement of available technologies but keeping intact the final goal: to generate and express reliable and detailed knowledge about soil resources for society (LIMA et al., 2013).

Traditionally, soil mapping was mainly carried out analogously. The pedologist uses topographic maps, aerial photographs, satellite images, and radar images to plan field activities, identify types of soils, and establish relationships between landscape attributes and soils. Based on this data, the pedologist develops a mental model to predict the spatial distribution of the types of soils in the study area (MCBRATNEY; WEBSTER, 1981). This method, although widely used in the field of soil science, is considered by many as subjective, difficult to reproducibility, and often does not offer validation or accuracy metrics for the generated map. Conventional mapping is based on Jenny's equation (1994), known as CLORPT (CL, climate; O, organisms; R, relief; P, parental material; T, time), which considers that soil is a derivation of the interaction of the factors listed in this equation and from this understanding, it is possible to create mental models to map the classes or attributes of the soil. Despite the epistemological and theoretical robustness of the model, the subjectivity implicit in making the maps based solely on human interpretation is incompatible with the search for reproducibility and application of the mapping methodology for extensive areas with faster results.

With the advent of digital technologies, such as programming languages, soil scientists/pedologists started a methodological workflow using information planes such as satellite images and digital elevation models associated with maps created from the interpolation of data obtained on the ground and implemented modern statistical functions starting a new trend in mapping that would come to be known as Pedometry and Digital Soil Mapping (DSM) (LAGACHERIE; MCBRATNEY; VOLTZ, 2007; MCBRATNEY et al., 2000; MCBRATNEY; MINASNY; STOCKMANN, 2018). The DSM is based on the SCORPAN model (S, soils; C, climate; O; organisms; R, relief; P, parental material; A, In time; N, space), which is an extension of the CLORPT model and includes additional variables such as spatial location (N, such as coordinates) and prior knowledge about the soil (S) (MCBRATNEY, A. B; MENDONÇA SANTOS; MINASNY, 2003).

In this sense, the DSM uses digital tools and advanced statistical techniques to improve the process of mapping soils and their correlated properties (MCBRATNEY; MENDONÇA SANTOS; MINASNY, 2003). The DSM's main objectives include producing maps with prediction error metrics that can be reproduced and evaluated by other researchers and have algorithms that allow the identification of relationships between the target property and explanatory properties. The source of information can be satellite images, existing pedological maps, and digital elevation models, among others. Once the statistical model between the variables is defined, it is possible to map extensive areas with spatial resolution from the explanatory variables used.

DSM primarily produces spatially continuous maps of soil properties using digital data sources such as remote sensing, terrain attributes, and climate data. It involves predictive modeling techniques like geostatistics and machine learning to generate maps that are directly useful for land management, agriculture, and environmental assessment (GRUNWALD, 2010; HARTEMINK; MCBRATNEY; MENDONÇA-SANTOS, 2008; LAGACHERIE, P.; MCBRATNEY; VOLTZ, 2007; MALONE; MINASNY; MCBRATNEY, 2017).

Pedometrics, however, focuses on the quantitative analysis and modeling of soil data. It applies statistical, mathematical, and geostatistical methods to understand soil variability, formation processes, and the relationships among soil properties. While pedometrics provides the foundational techniques used in digital soil mapping, its scope is broader, addressing both theoretical and practical aspects of soil science (DE CARVALHO JUNIOR et al., 2024; MCBRATNEY; MINASNY; STOCKMANN, 2018).

DSM does not seek to replace the work of the pedologist but rather to act as a tool that enables the generation of new maps of soils and their properties more efficiently. Traditionally, it requires fieldwork, opening profiles for description and characterization, sample collection, and laboratory analysis. While this methodological flow remains essential, the approach of pedometrics takes these soil samples further by integrating them with a wide array of covariates using advanced statistical and computational techniques. This integration not only preserves the critical role of soil sampling but also refines the mapping process — reducing subjectivity and enhancing the objectivity and robustness of the associations with soil-forming factors. Additionally, this approach can be complemented by proximal sensors and other equipment that indirectly measure properties linked to soil-type transitions or further increase map accuracy (CASA et al., 2013; GOOLEY et al., 2014; GRUNWALD et al., 2024; GRUNWALD; VASQUES; RIVERO, 2015; JI et al., 2019; NAWAR et al., 2017; RODRIGUES; BRAMLEY; GOBBETT, 2015; SHADDAD et al., 2016).

In the case of Brazil, a country with continental dimensions, significant knowledge about tropical soils has been accumulated over decades. However, the availability of detailed soil surveys at appropriate scales remains challenging. The primary cartographic information available is the map of soils of Brazil at a scale of 1:5,000,000 (SANTOS et al., 2011); the pedological maps of the Radam Brasil Project on a scale of 1:1,000,000, covering a large part of the Brazilian territory; and the soil survey carried out by the IBGE on a scale of 1:250,000, available for ten states of the federation (IBGE, 2018b). While some states, such as São Paulo, have more detailed mappings, much of the country still lacks surveys at scales compatible with current demands for land use planning, environmental monitoring, and precision agriculture. Therefore, there is a pressing need to develop methods that enhance the coverage and resolution of soil surveys across Brazil's diverse landscapes.

The selection of locations to sample soil profiles to best represent the heterogeneity of the landscape is an indispensable and laborious step among the many that involve making a map. Some alternatives have been proposed to optimize the planning of the allocation of sampling points based on different techniques, such as the conditioned Latin hypercube method (MINASNY; MCBRATNEY, 2006), cluster sampling, adaptive sampling, which adjusts points based on local variability, and stratified sampling (BRUS, 2014; BRUS et al., 2009), which seeks to represent different soil classes or homogeneous zones previously identified by analysis of environmental covariates or images from remote sensors.

Within the context of the DSM, the Reference Area (RA) approach emerges as an effective strategy to integrate existing knowledge from conventional mappings with digital techniques. RA consists of mapping soils in a small representative region in detail, establishing the main soil classes and the mapping rules that will be applied in the surrounding areas (LAGACHERIE et al., 2001;

LAGACHERIE; LEGROS; BURROUGH, 1995; TEN CATEN et al., 2011). This approach is based on the hypothesis that soil patterns are repetitive and identifiable in different natural regions, allowing the mapping of new areas to be facilitated and accelerated by applying the rules established in the RA. Studies have shown that the RA method can significantly reduce the need for additional fieldwork and maintain high accuracy in the maps produced (ARRUDA et al., 2016; FAVROT, 1981; LAGACHERIE; VOLTZ, 2000; YIGINI; PANAGOS, 2014). However, transferring this knowledge to other teams of pedologists still faces limitations. The experience acquired during the identification and delineation of RA is based on specialist knowledge and, therefore, like conventional soil mapping, is affected by the subjectivity of the specialist.

This work introduces the *autoRA* (Automatic Reference Area) algorithm, designed to streamline DSM by automating the delineation of reference areas. *autoRA* focuses on optimizing soil sampling in regions that are pedologically and geomorphologically complex or difficult to access, enabling data collection from a smaller yet representative subset that can accurately predict soil properties in a broader adjacent area. The mathematical approach is presented in the following chapters of this thesis.

In practical terms, autoRA calculates Gower's Dissimilarity from SCORPAN-based covariates, ranks pixels by their dissimilarity, and systematically selects the optimal candidate RA. By emphasizing maximum entropy—thereby covering the most remarkable environmental diversity, autoRA minimizes generalization error (E_{gen}) and ensures that predictive models built on the chosen RA achieve an extrapolation error below the specified threshold δ.

It is hypothesized that the autoRA method can automatically delineate a reference area where soil sampling captures the most diverse soil data and encompasses a broad range of covariate values. By systematically optimizing RA parameters—such as size, the proportions of high/low dissimilarity values, and repeated sampling within a Random Forest framework—the method is expected to achieve predictive performance comparable to full-coverage sampling while enhancing reproducibility and scalability for broader adoption in pedometrics. Consequently, a predictive model built from this small but representative area can be effectively extrapolated to the larger non-sampled area, with fewer additional soil samples used to validate the resulting soil property map externally.

The following objectives have been developed to test the autoRA algorithm and meet the aim of creating a tool capable of efficiently locating the coordinates for soil sampling. They will be organized in narrative individualized chapters.

Chapter 1 introduces the initial version of the autoRA algorithm by outlining its core concepts and methodological framework. This chapter details a sensitivity analysis conducted using various combinations of input parameters to test the method's robustness. The algorithm's applicability is then evaluated across two distinct study areas—one in Florida, USA, and the other in Rio de Janeiro, Brazil—demonstrating its versatility in different geographic contexts.

In Chapter 2, the focus shifts to refining the delineation of the reference area in the Recôncavo Baiano region using autoRA version 1. This chapter compares soil-unit maps produced by the automated method with those generated from a specialist-defined reference area (RA manual). In addition to employing the same covariates and utilizing a high-dissimilarity Gower's Index to select a smaller yet heterogeneous sampling area, the study tests the DSM approach by regrouping the entire dataset and partitioning it into 70% training and 30% validation subsets,

repeating this process 100 times. This comprehensive evaluation investigates whether the compact domain can yield robust extrapolation models for the broader region.

Chapter 3 presents the improvements made in autoRA version 2.0, which not only delineates the boundaries of the reference area but also identifies the optimal number of collection points within it. This chapter applies the enhanced algorithm in a practical case study along the Equatorial Margin of Brazil, covering the states of Pará, Amapá, and Maranhão. The refined approach is used to plan soil sampling in a way that supports strategic decision-making for implementing an oil and gas extraction base.

Each specific objective is aligned with a corresponding chapter in the thesis structure, ensuring that all aspects of the development, enhancement, and application of the autoRA algorithm are addressed systematically and integrated throughout the work. Chapter 3, in particular, demonstrates the practical application of autoRA in authentic contexts, evidencing its ability to generate accurate maps for categorical data and validating the algorithm's effectiveness in different geographic regions and variables of interest.

2. CHAPTER I

AUTORA: AN INNOVATIVE ALGORITHM FOR AUTOMATIC DELINEATION OF REFERENCE AREAS IN SUPPORT OF SMART SOIL SAMPLING AND DIGITAL SOIL TWINS

2.1 RESUMO

O Mapeamento Digital de Solos (DSM) aperfeiçoou o desenvolvimento e aplicação das informações sobre o solo, mas normalmente demanda dados de campo que são custosos e laboriosos para implementar e desenvolver modelos de predição de solo de maneira precisa. A abordagem da Área de Referência (RA) pode reduzir a intensidade da amostragem do solo; no entanto, devido a subjetividade do processo de lineamento a precisão do modelo de predição pode ficar comprometido. Neste estudo, apresentamos o algoritmo autoRA, um método inovador de amostragem automatizada de solo que utiliza o Índice de Dissimilaridade de Gower para delinear RAs automaticamente. Essa abordagem preserva a variabilidade ambiental, mantendo a precisão em comparação com um modelo preditivo exaustivo (EPM) baseado em amostragem extensiva da área de interesse. Nosso objetivo foi avaliar a sensibilidade e a eficiência do autoRA variando as áreas-alvo (10–50% da área total) e as resoluções espaciais do tamanho do bloco (5–150 pixels) nas regiões da Flórida, EUA, e Rio de Janeiro, Brasil. Modelamos uma propriedade hipotética do solo derivada de uma combinação de covariáveis DSM comumente usadas e entradas do usuário no autoRA. O desempenho do modelo foi avaliado por meio de R^2, raiz quadrática do erro média (RMSE) e Bias, agregados em uma métrica de Distância Euclidiana (ED). Entre todas as configurações, a seleção ideal de RA - caracterizada pelo ED mais baixo - foi alcançada com uma área alvo de 50% e um tamanho de bloco de 10 pixels, correspondendo à precisão do EPM. Por exemplo, o EPM produziu um ED de 0,17 no Rio de Janeiro, enquanto a melhor configuração de RA produziu um ED de 0,15. Na Flórida, o EPM teve um ED de 0,35 em comparação com 0,38 para o RA ideal. Além disso, a 50%-RA com um tamanho de bloco de 10 reduziu significativamente os custos totais em aproximadamente 61% no Rio de Janeiro (US$ 258.491 a US$ 100.611) e 63% na Flórida (US$ 289.690 a US$ 106.296). O AutoRA identificou sistematicamente configurações de amostragem econômicas e reduziu a área de investigação, mantendo a precisão do modelo. Automatizando o delineamento de RA, o autoRA mitigou a subjetividade inerente aos métodos tradicionais, suportando assim fluxos de trabalho DSM mais reprodutíveis, estratégicos e eficientes.

Palavras-chave: Mapeamento Digital de Solos. Modelagem preditiva. Estratégias de amostragem.

2.2 ABSTRACT

Digital Soil Mapping (DSM) enhances the delivery of soil information but typically requires costly and extensive field data to develop accurate soil prediction models. The Reference Area (RA) approach can reduce soil sampling intensity; however, its subjective delineation may compromise model accuracy when predicting soil properties. In this study, we introduce the autoRA algorithm, an innovative automated soil sampling design method that utilizes Gower's Dissimilarity Index to delineate RAs automatically. This approach preserves environmental variability while retaining accuracy compared to an exhaustive predictive model (EPM) based on extensive sampling of the area of interest. Our objective was to evaluate the sensitivity and efficiency of autoRA by varying target areas (10–50% of the total area) and block size spatial resolutions (5–150 pixels) in regions of Florida, USA, and Rio de Janeiro, Brazil. We modeled a hypothetical soil property derived from a combination of commonly used DSM covariates and user inputs into autoRA. Model performance was assessed using R^2, root mean square error (RMSE), and Bias, aggregated into a Euclidean Distance (ED) metric. Among all configurations, the optimal RA selection—characterized by the lowest ED—was achieved with a target area of 50% and a block size of 10 pixels, closely matching the accuracy of the EPM. For example, the EPM produced an ED of 0.17 in Rio de Janeiro, while the best RA configuration yielded an ED of 0.15. In Florida, the EPM had an ED of 0.35 compared to 0.38 for the optimal RA. Additionally, the 50%-RA with a block size of 10 significantly reduced total costs by approximately 61% in Rio (US$258,491 to US$100,611) and 63% in Florida (US$289,690 to US$106,296). AutoRA systematically identifies cost-effective sampling configurations and reduces the investigation area while maintaining model accuracy. Automating RA delineation, autoRA mitigates the subjectivity inherent in traditional methods, thereby supporting more reproducible, strategic, and efficient DSM workflows.

Keywords: Digital Soil Mapping. Predictive Modeling. Sampling Strategies.

2.3 INTRODUCTION

A Digital Twin is a virtual replica of a physical object, system, or process that is continuously updated with real-time data (GRIEVES; VICKERS, 2017). This digital representation enables simulation, analysis, and optimization of the actual entity's behavior and performance, effectively bridging the gap between the physical and digital worlds. It allows predicting outcomes, diagnosing issues, and testing various scenarios without impacting the real system (GRIEVES, 2022). When the digital twin concept is applied to soils, a dynamic virtual model of the soil environment is created (KIM; HEO, 2024). This model integrates data from various sources, such as in situ sensors, satellite imagery, and simulation tools, to replicate key soil characteristics like moisture, nutrient levels, structure, and thermal properties (CESCO et al., 2023; GRIEVES; HUA, 2024). In precision agriculture, this approach assists in optimizing irrigation, fertilization, and crop management by monitoring soil variability in real-time (DEFRAEYE et al., 2021; PELADARINOS et al., 2023; PURCELL; NEUBAUER, 2023; PYLIANIDIS; OSINGA; ATHANASIADIS, 2021; RODRIGUES et al., 2024). It also supports environmental management by predicting soil responses to changes or remediation strategies and assessing risks like erosion or contamination. Additionally, the use of digital twins in research and modeling helps enhance the understanding of complex soil processes and the interactions among soil, climate, vegetation, and land management practices (VERDOUW et al., 2021).

Pedometry and Digital Soil Mapping (DSM) have revolutionized soil science by enabling the prediction of soil properties and classes across extensive and heterogeneous regions using limited site-specific measurements combined with environmental covariates (MALONE; MINANSY; BRUNGARD, 2019; MCBRATNEY; MENDONÇA SANTOS; MINASNY, 2003). Central to DSM's methodology is the SCORPAN model, which identifies the key soil-forming factors—Soil, Climate, Organisms, Relief, Parent material, Age, and spatial Position—that drive soil variability (MCBRATNEY; MENDONÇA SANTOS; MINASNY, 2003). Advances in global positioning systems, remote sensing technologies, proximal sensors, and computational capacities have further empowered DSM, facilitating the development of sophisticated machine learning algorithms that produce high-resolution gridded soil maps for informed land management and agricultural practices globally (CHEN et al., 2024 (p. 2); KHALEDIAN; MILLER, 2020; MALONE et al., 2017). Despite significant progress, creating fine-resolution soil maps remains challenging due to the limited availability of ground-truth soil data necessary for accurate models. Large-scale initiatives like the Harmonized World Soil Database and SoilGrid provide soil maps at resolutions of approximately 1 km and 250 m, respectively (HENGL et al., 2014; POGGIO et al., 2021). However, extending these resolutions to broader scales is hindered by the scarcity of extensive, high-quality soil measurements that capture the intricate spatial and temporal variability in diverse landscapes. This limitation highlights the need for optimized sampling strategies that efficiently allocate limited resources to maximize data representativeness and model accuracy, especially in remote and ecologically complex regions such as the Brazilian Amazon and the Rocky Mountains in the USA.

One promising solution is the Reference Area (RA) approach, which strategically focuses sampling efforts within a sub-region that encapsulates the essential variability of soil-forming factors present in the larger Area of Interest (AOI) (FAVROT, 1981; LAGACHERIE, et al., 2001; LAGACHERIE, P.; LEGROS; BURROUGH, 1995). This method can significantly reduce sampling costs and logistical burdens while maintaining DSM models' integrity and predictive power (ARRUDA et al., 2016; FERREIRA, et al., 2022). Ferreira et al. (2023), using Gower's Dissimilarity Index to assess RA representativeness, effectively identifies areas where

environmental covariates diverge from the broader AOI, indicating regions where model predictions may falter. Integrating dissimilarity metrics into RA delineation can thus enhance DSM efforts' precision and scalability.

However, the RA approach has predominantly relied on subjective expert judgment for delineating RA boundaries, introducing potential biases and limiting reproducibility (JEAN-MARC ROBBEZ-MASSON, 1994; TEN CATEN et al., 2011). Existing algorithms like CLAPAS, a Procédure Interactive Et Itérative De Classement-Classification (Interactive And Iterative Classification-Ranking Procedure), require manual input of candidate RAs and do not fully automate the delineation process, allowing for human error and inconsistency. Additionally, methods such as conditioned Latin hypercube sampling (cLHS) and divergence metrics (e.g., Kullback-Leibler Divergence) have been explored to optimize sample design and size but often lack direct applicability to the RA framework or fail to link sample size with model performance clearly (MALONE; MINANSY; BRUNGARD, 2019; MINASNY; MCBRATNEY, 2006).

These methodological gaps have significant implications for regions with vast spatial extents and diverse soil landscapes, such as Brazil and the USA. With approximately 8.5 million square kilometers in Brazil, soil mapping is challenged by diverse climate zones, varied geomorphology, and remote, ecologically sensitive areas. Current soil maps cover less than 5% of the national territory at scales finer than 1:100,000 (CANAVESI et al., 2020; FILIPPINI-ALBA; FLORES; BERNARDI, 2023; MOURA, Derick Martins Borges De et al., 2020; VASCONCELOS et al., 2023). Similarly, the USA, encompassing around 9.4 million square kilometers, has achieved detailed soil mapping in agriculturally intensive regions through initiatives like the SSURGO database and SOLUS soil maps (NAUMAN et al., 2024) but faces challenges in natural areas such as the Greater Everglades in Florida due to complex geomorphology and difficult sampling conditions. Addressing these challenges requires automated, objective methods for RA delineation to mitigate subjectivity and enhance the reproducibility and scalability of DSM studies. This chapter aims to introduce the automatic Reference Area algorithm (autoRA version 1.0), a novel tool designed to standardize and automate RA delineation by leveraging Gower's dissimilarity index and a comprehensive sensitivity analysis framework. autoRA systematically identifies RAs that capture the full spectrum of environmental covariate variability within an AOI, ensuring accurate and cost-effective soil models without relying on expert intuition.

To validate autoRA's efficacy, we apply the algorithm to two distinct study areas: the State of Florida (USA) and Rio de Janeiro (Brazil). These regions were chosen for their contrasting pedodiversity patterns and varying sampling difficulties—Florida represents an agriculturally intensive and accessible landscape. At the same time, Rio de Janeiro encompasses remote and ecologically complex terrains. We conduct a sensitivity analysis by varying the spatial resolution of environmental covariate maps and RA sizes to evaluate impacts on model accuracy and sampling costs. Additionally, we use simulated theoretical surface attribute maps to assess autoRA-generated RAs' robustness under different modeling scenarios.

This study presents autoRA as a replicable, data-driven tool for DSM practitioners, contributing to the broader discourse on optimal sampling strategies in soil science. Automating RA delineation, autoRA facilitates efficient and objective soil survey designs, enhancing DSM efforts' scalability and reliability in remote, ecologically complex regions and more accessible, intensively studied landscapes. Ultimately, autoRA represents a significant advancement toward standardized and scalable DSM methodologies, enabling comprehensive soil mapping to inform sustainable land management and agricultural practices globally.

2.4 MATERIAL AND METHODS

2.4.1 The autoRA algorithm

The novel autoRA was developed by the authors' team and is patented under the number BR1020240208676, and the brand autoRA is under the registered trademark number 937505684. This registration took place in Brazil, and soon, they will be registered in the United States Patent Office. The autoRA allows for the automatic delineation of RAs with different dimensions (i.e., RA target area) to implement smart soil sampling designs. A fundamental challenge is whether a delineated RA can generate accurate predictive soil models comparable to the exhaustive simulated soilscape ("on-the-ground truth").

The algorithm involves several processes, each contributing to its overall complexity. First, the Gower Dissimilarity Calculation requires computing pairwise distances between all pixels in the study area, resulting in a worst-case complexity of $O(N^2 \cdot p)$, where N is the total number of pixels and p is the number of covariates. If the number of covariates is as large as N, the complexity can further increase to $O(N^3)$. Next, for the Sorting Process used in selecting the Reference Area (RA), an efficient sorting algorithm such as QuickSort (average $O(N \log N)$) or HeapSort (guaranteed $O(N\log N)$) is applied, leading to a complexity of $O(N\log N)$. The Filtering Step for selecting the RA consists of a single pass through the data, yielding a complexity of $O(N)$.

Following these steps, the Training of the Random Forest Model depends on the number of selected samples M. In the worst case, decision trees require sorting at each node, which results in a complexity of $O(M \log M)$. Finally, Validation Metrics Calculation, which involves computing accuracy measures over the selected samples, requires iterating through M, leading to a complexity of $O(M)$.

Summing all these contributions, the total complexity of autoRA can be expressed as $f(N) = O(N^2 \cdot p) + O(N\log N) + O(N) + O(M\log M) + O(M)$. In the worst case, where $p = N$, this simplifies to $f(N) = O(N^3) + O(N\log N) + O(N) + O(M\log M) + O(M)$. If $M \ll N$, meaning the number of selected samples is significantly smaller than the total number of pixels. The dominant term is the Gower Dissimilarity Calculation, resulting in a worst-case complexity of $O(N^3)$.

The computational bottleneck in autoRA arises from the calculation of Gower dissimilarities, which grow cubically with N in the worst-case scenario. Other steps like sorting, filtering, training, and validation have relatively lower computational costs.

Simulated soilscape rasters ensured that each pixel was populated, providing continuous data across the AOI that exhaustively characterized soil patterns. In contrast, real-world soil measurements were typically sparse, with substantial gaps between pedons/sites. Thus, real-world soil datasets did not allow us to characterize the variability of a soil property of interest exhaustively, precisely, and accurately. Interpolated or estimated soil properties of real-world soilscapes showed uncertainties at unsampled locations. Thus, published soil maps were also ill-suited for assessing the sensitivity and effectiveness of the autoRA. Therefore, we simulated hypothetical exhaustive rasters assumed to represent the "ground truth" of a variable of interest, S_{exh} (i.e., a simulated theoretical surface, STS). The STSs were generated from SCORPAN variables of the two AOIs serving as benchmark maps.

We simulated two soilscape rasters using soil-forming (SCORPAN) factors in two contrasting study areas (Rio de Janeiro and Florida). The simulated soil properties provide an idealistic representation of these soilscapes, allowing the assessment of the autoRA algorithm's

behavior and demonstrating its sensitivity to its optional settings (precisely the parameters block size, representing the resolution of the covariates entered, and target area, representing the desired RA dimensions in the ratio of the AOI to be mapped) on soil predictions accuracy.

The next step was to implement the sensitivity analysis, varying the parameters of the autoRA within upper and lower bounds to generate possible RAs and their associated accuracies of the target soil variable of interest, S_{RA}. The selected block size lower and upper bounds were set to 5 and 150 pixels, respectively.

An overview of our methodology applying the autoRA to perform the sensitivity analysis is presented in Figure 1, while details of each step are described in the section below. The first step was to assemble geodata to represent the soil-forming factors of the SCORPAN model comprising various qualitative (nominal/ordinal) and quantitative (discrete/continuous) data (STEP 1, Figure 1). Once the covariate files were loaded, the autoRA algorithm worked simultaneously along two paths for each AOI. Path 1 generated the STS, predictions of S_{exh} using machine learning (Random Forest) from 500 S_{train} locations from the STS surface (STEP 2.1A and STEP 2.2), validation of the S_{exh} using an independent validation dataset (STEP 2.1B), and model performance assessment (STEP 2.3; metrics: coefficient of determination, R^2; root mean square error, RMSE; and Bias).

Path 2 involved computations of possible RAs via the sensitivity methodology entailing calculation of the Gower's Dissimilarity Index, delineation of the RA boundaries, predictions of S_{RA} using machine learning (Random Forest) with various parameter settings of lower and upper bounds, validation of the various S_{RA} using an independent validation dataset, and model performance assessment (metrics: R^2, RMSE, and Bias).

Path 2 represents the part of the algorithm that effectively creates the RAs by cropping the input raster maps (blue border boxes on the right of Figure 1). It involved STEP 3.1, which calculated the Gower's dissimilarity index for each covariate of the SCORPAN model loaded into autoRA, considering different block sizes. In STEP 3.2, the algorithm delineated the limits of the RAs by mosaicking the highest values of the Gower's dissimilarity index concerning the average Gower's dissimilarity index of the full extension of the study area. In STEP 3.2, the algorithm created RAs with different dimensions (i.e., respecting the size of RAs inputted by the user on the parameter target area), represented in terms of % of the location relative to the full extension of the study area (10 to 50%, increasing at a 10% growth rate). STEP 3.3 used a fixed number of points sampled within each generated RA to build S_{RA} prediction models to be applied for the whole AOI. The prediction models are called Reference Area Model (RAM).

Results from Path 1 (i.e., the exhaustive benchmark, the EPM raster surface) and Path 2 (i.e., RAM raster surfaces for multiple RAs) were compared to each other using evaluation metrics (R^2, RMSE, and Bias) in STEP 4. To compare the metrics and choose the RA that produces the best model, the Euclidean Distance (ED) of the metrics of each RA with an idealized standard vector of these metrics was used ($R^2 = 1$, RMSE = 0, and Bias = 0). The RA with the smallest ED value concerning the standard vector was identified as the best-performing RA for a given study area.

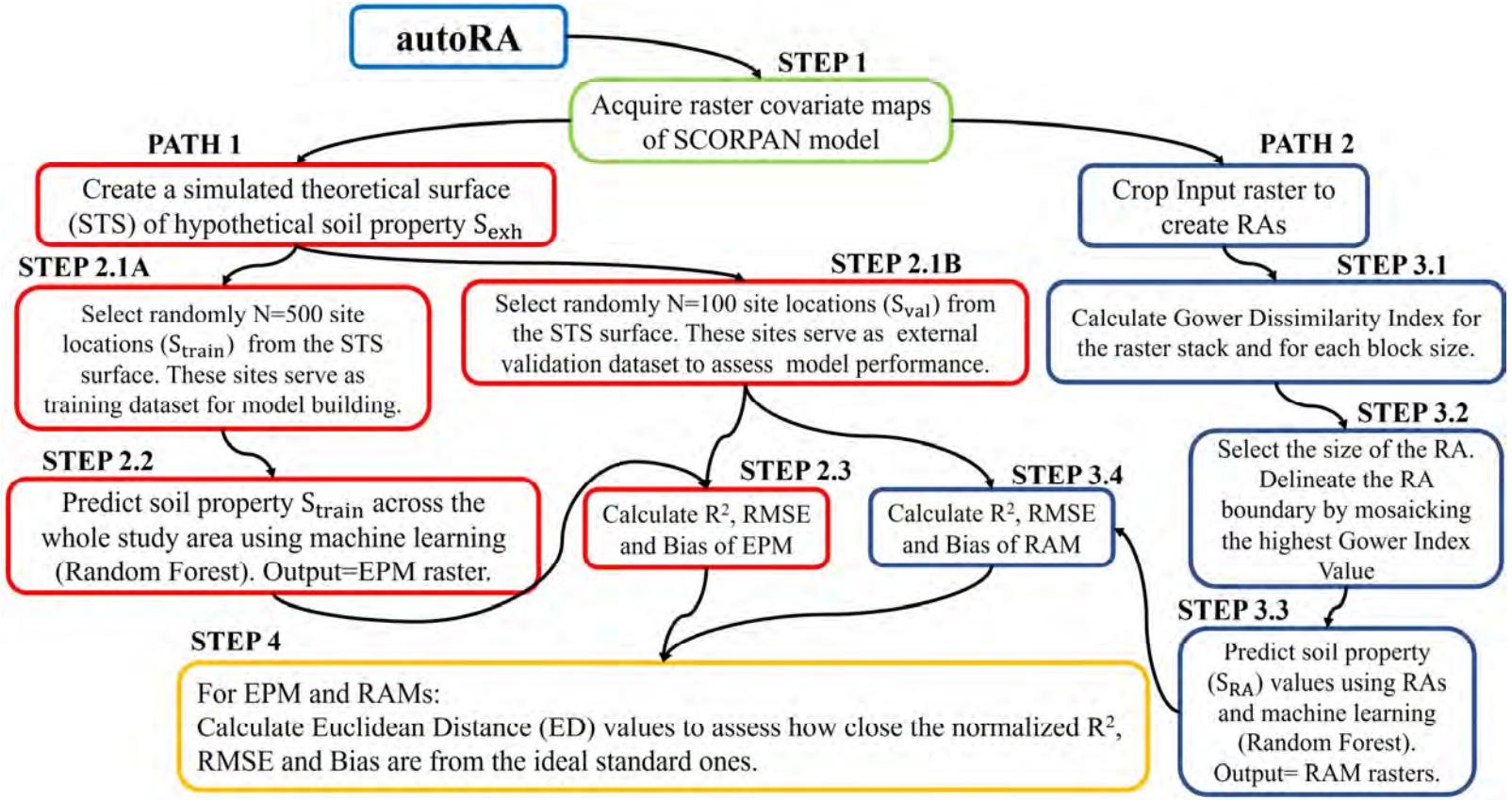


Figure 1. The workflow of the methodology (SCORPAN= Theoretical, quantitative model for soil modeling and mapping (MCBRATNEY; MENDONÇA SANTOS; MINASNY, 2003), S, Soil; C, Climate; O, Organisms including land cover and natural vegetation; R, Relief including terrain attributes; P, Parent material including lithology; A, Age/the time factor; and N space, spatial or geographic position; STS = Simulated Theoretical Surface; RMSE= Root Mean Squared Error; R2 = Coefficient of Adjustment; EPM, Exhaustive Predicted Model; RAM, Reference Area Model Prediction for the target properties STS; RF, Random Forest.

2.4.2 Applying the autoRA in two contrasting soilscapes

a) The study area

Two study areas were selected to evaluate the effectiveness of the autoRA algorithm in different pedological contexts characterized by different soil formation factors. The regions chosen are the State of Florida, located in the USA, with an area of 170,304 km², and the State of Rio de Janeiro in Brazil (BR), with an area of 43,653 km² (Figure 2). Florida is characterized by a predominantly flat terrain, with elevations ranging from sea level up to 110 m. The soil's parent material mainly comprises marine sediments and limestone rocks, resulting in a geology dominated by sedimentary formations (SELLARDS, 1919). The climate is mostly humid subtropical, with annual precipitation ranging between 1,200 mm and 1,800 mm, significantly influencing pedogenetic processes. The primary soil types in Florida are Spodosols, Entisols, and Inceptisols. Ultisols, clayey and more weathered soils, are present in regions with slightly undulated relief. Histosols, carbon-rich soils, are prominent throughout Florida, occurring in isolated wetlands and the Greater Everglades in South Florida (LAPIERRE; IRIZARRY; ANDREU, 2022). The main factors in soil formation in Florida include the parent material (mainly sedimentary), warmer and humid climate, moderate to high precipitation, predominant vegetation of coniferous forests and coastal plains, and flat relief that favors slow drainage and accumulation of organic matter (WATTS; COLLINS, 2008).

The State of Rio de Janeiro has a complex geology composed of igneous and metamorphic rocks, such as granites and gneiss, associated with sedimentary rocks and colluvial and alluvial deposits (HEILBRON et al., 2020). The relief varies from mountainous, with altitudes that reach 1,600 meters, as in the Serra dos Órgãos Mountains to nearly flat terrain at the coast and toward the North region of the state (Figure 2). The climate is humid tropical, with annual precipitation between 1,000 mm and 2,500 mm, influenced by the proximity of the Atlantic Ocean (JUNIOR; NASCIMENTO, 2022). The pedological diversity includes mainly Oxisols and Ultisols, with Inceptisols and Entisols formed on various landscapes (IBGE, 2018a). Soil formation factors in Rio de Janeiro are influenced by geological diversity, relief, humid climate, and dense vegetation of the Atlantic Forest, which contribute to the high rate of weathering and deep soil formation (GELSLEICHTER et al., 2023; PEREIRA; ANJOS, 1999; PEREIRA, et al., 2023).

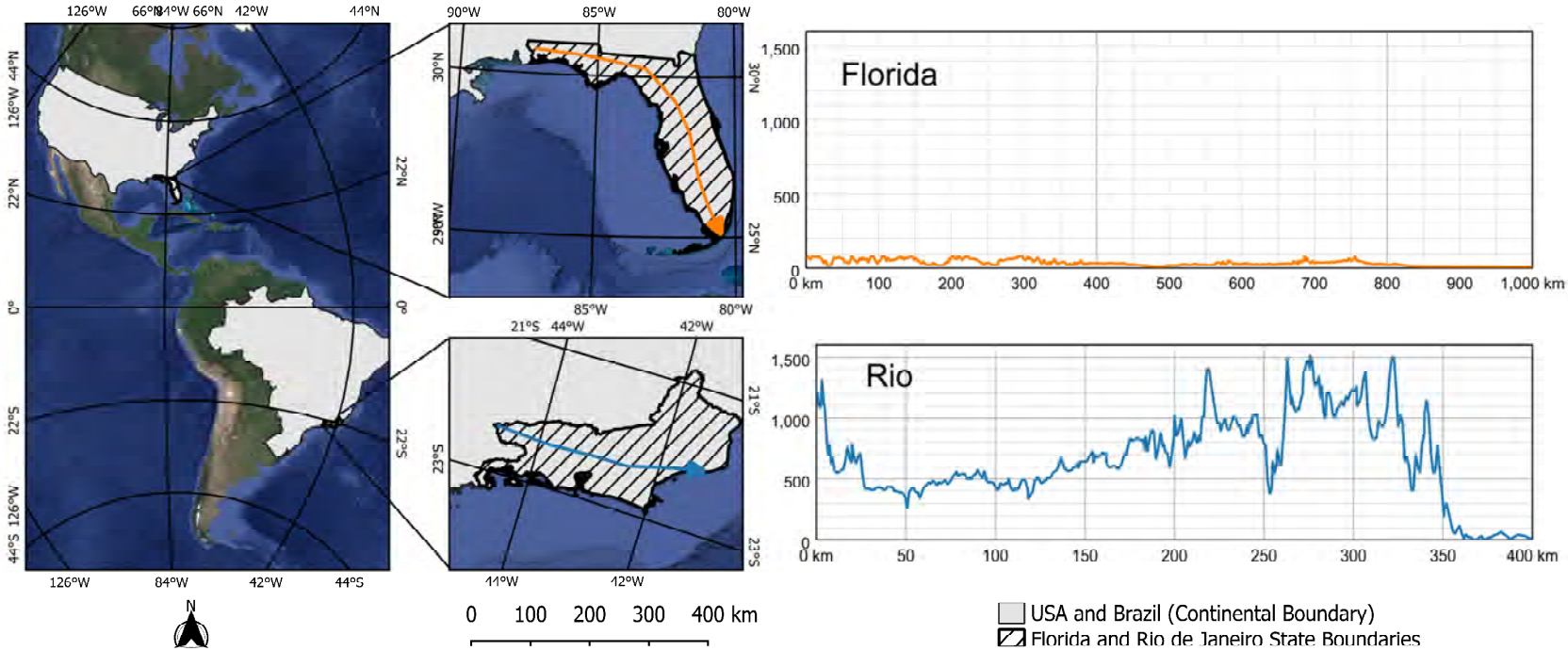


Figure 2. Location map of Florida and Rio de Janeiro and their respective elevation profiles.

b) Environmental covariates

To represent the soil formation factors (JENNY, 1994), five environmental covariates were selected for each study area (STEP 1, Figure 1) to apply the autoRA algorithm and generate the STS (Path 1, Figure 1). The soil map provided by the Natural Resources Conservation Service (NRCS) United States Department of Agriculture (SOIL SURVEY STAFF, NATURAL RESOURCES CONSERVATION SERVICE, 2016) was used to identify the types of soils in Florida at a scale of 1:250,000, containing seven soil levels. The U.S. Geological Survey (USGS MINERAL RESOURCES, 2017) the geologic map containing 35 geologic levels for the Florida region alone at a scale of 1:100,000 was made available. The relief was represented by a digital elevation model (DEM) with a resolution of 30 m and made available by the National Aeronautics and Space Administration using Advanced Spaceborne Thermal Emission and Reflection Radiometer satellite data. The raster data of average annual precipitation and average annual temperature from 1981 to 2010 were obtained from the NRCS (2012a; 2012b) at a final resolution of 1 km.

For Rio de Janeiro, the Brazilian Institute of Geography and Statistics (IBGE, 2018b) made the soil maps available at a scale of 1:250,000, containing 21 soil levels. The IBGE provided the geology map at a scale of 1:250,000 with nine geological levels. To represent the relief of Rio de Janeiro, the DEM of the Shuttle Radar Topography Mission (SRTM) satellite with a spatial resolution of 90 m was used. The precipitation and average annual temperature maps were obtained from the WorldClim database (FICK; HIJMANS, 2017) with a spatial resolution of 1 km from 1980 to 2016. All covariates for both case studies were harmonized to 1 km spatial resolution.

c) The simulated theoretical surface

The STS (Path 1) in Figure 3 was implemented using an adapted methodology from Meyer and Pebesma (2021). The STS acts as a hypothetical target variable (S_{exh}) that is both explainable and plausible, reflecting soil information derived from environmental covariates for a given study area. Before the map algebra, all categorical covariates (e.g., a geology map with multiple classes) were split into separate raster layers using dummy transformations of 0 or 1. Thus, each category becomes an individual map with presence coded as 1 and absence as 0. Numeric covariates, such as digital elevation models (DEM), precipitation, or temperature, were scaled to a 0–1 range to ensure comparability across all variables.

Using these standardized covariates, the STS is generated via map algebra interactions among the covariate maps (Figure 3), ensuring consistency with their spatial patterns. For example, the STS in Florida (STS_{Flo}) is computed by:

$$STS_{Flo} = Pedo_{Flo} + Geo_{Flo} + (DEM_{Flo} * Precip_{Flo}) + Temp_{Flo},$$

While the STS in Rio de Janeiro (STS_{Rio}) is calculated as:

$$STS_{Rio} = Pedo_{Rio} + Geo_{Rio} + DEM_{Rio} + (Precip_{Rio} * Temp_{Rio}),$$

As a calculated synthetic map, the STS is assumed to be error-free. Figure 3 illustrates this process for both Florida and Rio de Janeiro. To facilitate direct comparison, each STS is subsequently normalized to a 0–1 scale using:

$$STS_{normalized} = \frac{STS - STS_{min}}{STS_{max} - STS_{min}},$$

And then multiplied by 100 to yield a final scale from 0% to 100%. Because this map algebra product is solely an interaction of covariates, it does not represent a physically measured quantity. Instead, it is a spatially plausible synthetic surface that reflects the relative influence of each covariate on a hypothetical soil property. These dimensionless $STS_{normalized}$ values serve as a reference populated with S_{exh} for all subsequent analyses, functioning as a benchmark map to assess the efficiency of parameters of the autoRA algorithm.

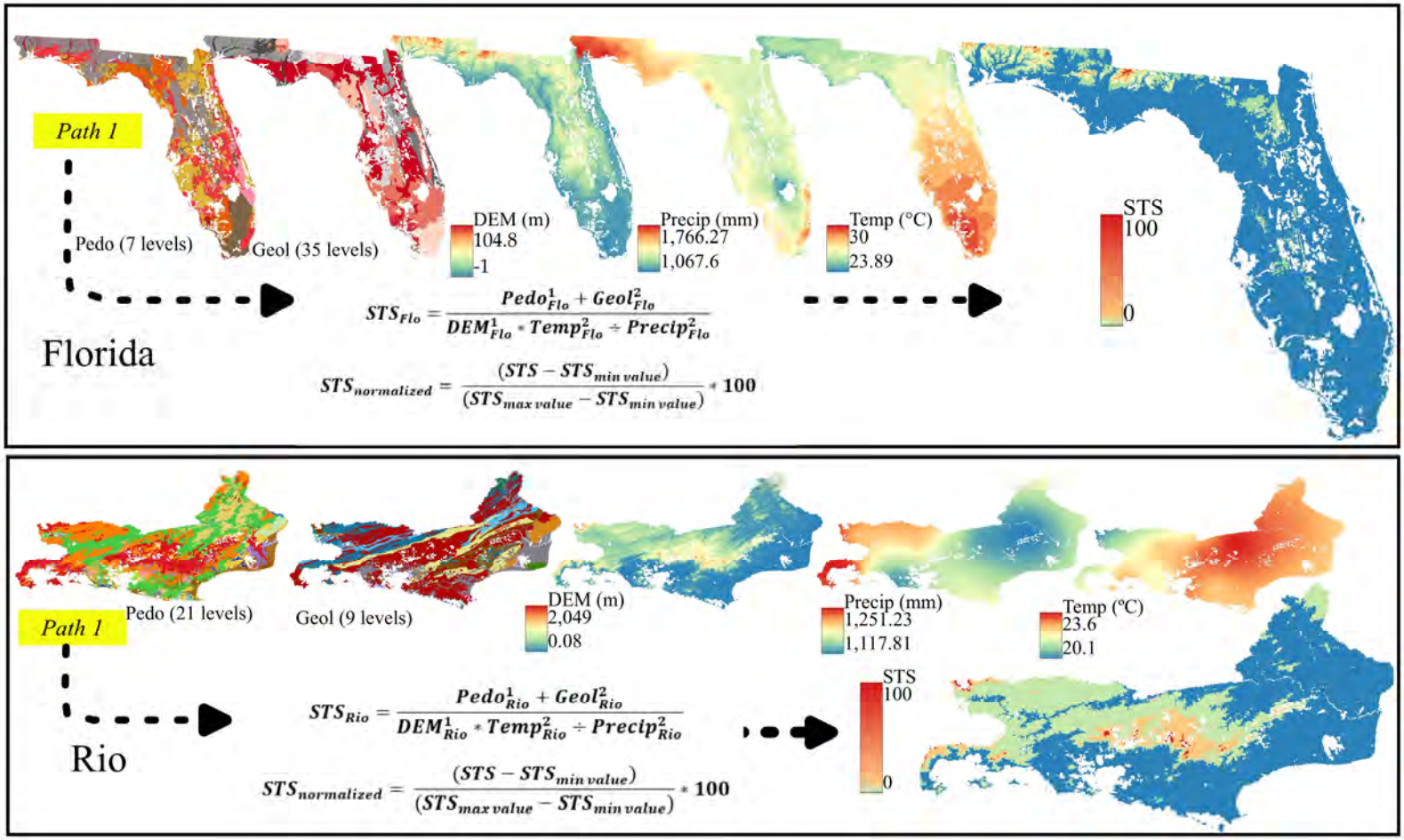


Figure 3. Raster stacking (Path 1) to calculate the STS for Florida and Rio de Janeiro. Pedo - Pedology; Geol - Geology; DEM - Digital elevation model (meters); Precip. - Annual average precipitation in millimeters; Temp. – Annual average temperature in ºC.

d) Training and Validation Datasets

A grid of 500 points (representing site locations) was randomly generated (STEP 2.1A). We chose a random distribution of these points to ensure adequate spatial representativeness and avoid excessive concentration in certain areas that could introduce biases in the predictive model. To allow side-by-side comparisons of Path 1 (EPM) and Path 2 (RAMs), the same number of points (N: 500) were chosen in each of the two study areas. We used the spatial extraction function in ArcGIS Pro to extract the variable S_{train} at the 500 site locations of STS_{Flo} and STS_{Rio}. The S_{train} values were then used as the target (dependent) variable for predictive model development using machine learning (training phase). A second grid of 100 points (representing site locations) was randomly extracted from the STS raster, with variable Sval serving as an external validation dataset in each study area (STEP 2.1B). An independent validation set is essential to verify the model's ability to generalize its predictions to new samples not used during a model's training, thus ensuring the reliability and applicability of the results obtained.

Figure 4 illustrates the spatial distribution of the training and validation points in both study areas (Florida and Rio de Janeiro). We used the Random Forest (RF) machine learning algorithm to develop predictive models using the environmental covariates of the SCORPAN model (pedology, geology, digital elevation model, average precipitation, and average temperature) as input (independent) variables and S_{train} as output (target) variable (STEP 2.2 of Path 1 in Figure 1). Training models were customized to study areas with separate RF training models developed for Florida and Rio de Janeiro (Rio). We employed the "randomForest" package (LIAW; WIENER, 2002) available for the R software (R CORE TEAM, 2024).

In our RF regression modeling for the Florida and Rio datasets, we employed the default parameters provided by the R package randomForest to ensure consistency and reliability across our analyses. Specifically, each model was constructed with 500 trees (ntree = 500), a number sufficient to ensure that every input row is predicted multiple times, thereby enhancing the stability and accuracy of the predictions. The number of variables randomly sampled as candidates at each split (mtry) was set to 1, following the default of one-third of the total predictors (p/3), given that each dataset contained five predictors. Additionally, the minimum size of terminal nodes (nodesize) was maintained at the default value of 5, encouraging the growth of smaller, more computationally efficient trees while preventing overfitting. By adhering to these default parameter settings for the Florida and Rio Random Forest regression models, we ensured a balanced approach that optimizes predictive performance and computational efficiency without requiring extensive parameter tuning.

The RF algorithm was chosen based on its ability to handle large volumes of data, its robustness in noisy data, and its ability to capture highly nonlinear relationships between the covariates and the dependent variable. In addition, RF offers internal validation mechanisms, such as estimating the importance of variables, which contribute to the interpretation and improvement of the model (CLINGENSMITH; GRUNWALD, 2022). The trained machine learning model for Florida using the S_{exh} generated the STS-EPM raster for Florida while the same procedure was used to create the STS-EPM for Rio de Janeiro.

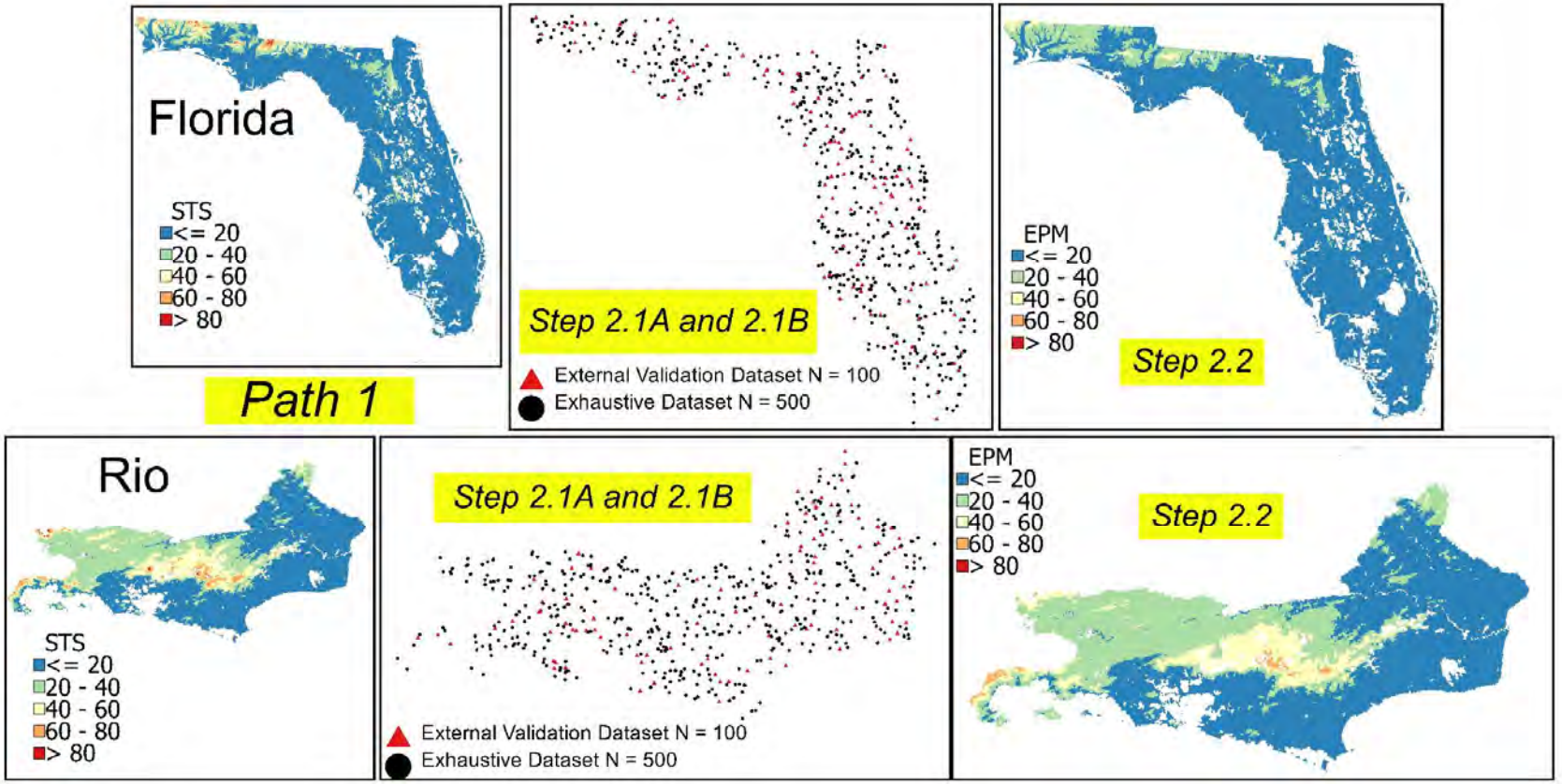


Figure 4. Spatial distribution of the training (N: 500) and validation (N: 100) datasets for the selected study areas. Path 1, Simulated Theoretical surface (STS); sampling training and validation dataset (STEP 2.1A and STEP 2.1B of Figure 1); Exhaustive Prediction Model (EPM) using the training dataset and Random Forest machine learning – STEP 2.2).

2.4.3 Determining the gower index and delineating the reference areas

In Path 2 (STEP 3.1 and 3.2 in Figure 1), the autoRA algorithm was run by cropping the Florida and Rio de Janeiro covariates and then calculating Gower's dissimilarity index. The algorithm offers an argument called block size (block size) dependent on the original spatial resolution of a given input raster. The block size parameter allows the grouping of pixels into a window defined by the number of rows *x* columns of the block (e.g., 5 x 5 pixels = 25 pixels total in a block). For example, if the original resolution of the covariates is 1 km² and the block size is set to 5, each block will have a dimension of 5 km x 5 km. Then, considering each block size value, the entire set of covariates is clipped using the respective block size mask.

Figure 5 shows the schematic mask used to clip the set of covariates from the values used for escalating block size values from lower to upper bounds of 5, 10, 20, 30, 40, 50, 100, and 150 for Florida and Rio. As noted earlier, all covariates for both study areas had been harmonized to 1 km spatial resolution. Applying a block size value of 5 resulted in 5 km x 5 km blocks (total size of 25 km^2); likewise, increased block size values generated larger blocks across the study areas, aggregating the input data.

The Gower's dissimilarity index of the autoRA algorithms was calculated based on Gower (1971) as described in equations (1) to (3) (STEP 3.1 in Figure 1). Each block size value mask cropped a covariate raster, and the Gower's dissimilarity index is calculated (X_{block}) and it was compared to Gower's dissimilarity index of the covariate raster in the total area (X_{total}). Suppose the X_{block} values had a Gower's dissimilarity index value close to the (X_{total}), it means that the dissimilarity is low with a value close to 0, and vice versa. It is high with a value close to 1.

This process is repeated for each covariate present in the data set. The dissimilarity values obtained for each covariable are then summed for each set of covariates and grouped according to the block size. The lower the sum of Gower's dissimilarity indices for a given block size, the lower the diversity between the block investigated and the AOI. On the other hand, as the differences between the Gower indices calculated for each block and the AOI increase, the values tend to approach 1, indicating a high Gower's dissimilarity index between the covariables in the specific pixel aggregation and the covariates of the entire area suggesting that the block captures significant variability that is not represented by the average value of the AOI.

$$\textbf{Gower Dissimilarity}_{\textbf{block,total}} = 1 - \frac{\sum_{k=1}^{p} \delta_k d_k}{\sum_{k=1}^{p} \delta_k} \qquad \text{Eq. (1)}$$

Where p represents the total number of variables considered (e.g., geology, pedology, DEM, precipitation, temperature), and δ_k is an indicator that takes the value 1 if the variable k is valid for the comparison (i.e., relevant and has data available) and 0 otherwise. The term d_k is the normalized difference for variable k, which quantifies the dissimilarity between the block and the total area for that specific variable. The numerator sums the contributions of valid variables ($\delta_k d_k$), while the denominator ensures that only valid variables are included in the normalization. The final value is subtracted from 1 so that the index represents dissimilarity, where higher values indicate more significant dissimilarity between the block and the total area.

For numerical variables, such as temperature, precipitation, or elevation, d_k is calculated as the normalized difference between the block's value ($x_{block,k}$) and the total area's value ($x_{total,k}$). Here, ($x_{block,k}$) represents the average or representative value of a variable k within the block, and ($x_{total,k}$) represents the average or representative value of the same variable across the total area.

These values, $x_{block,k}$ and $x_{,total,k}$ are derived by calculating the average of the variable within the block or across the total area, respectively. The normalized difference is computed by Equation 2, where $Range_k$ is the difference between the dataset's maximum and minimum values of k.

$$\mathbf{d_k} = \frac{\left|\mathbf{x_{block,k}} - \mathbf{x_{total,k}}\right|}{\mathbf{Range_k}} \quad \text{Eq. (2)}$$

Equation 3 defines the values of d_k in the case of categorical variables, such as the maps of geology and pedology.

$$\mathbf{d_k} = \begin{cases} \mathbf{0}, & \mathbf{if\ x_{block,k}} = \mathbf{x_{total,k}} \\ \mathbf{1}, & \mathbf{if\ x_{block,k}} \neq \mathbf{x_{total,k}} \end{cases} \quad \text{Eq. (3)}$$

Another argument the autoRA algorithm has is the target area, which represents the size of the RA that the user would like to delineate. The pixel Gower values with the highest dissimilarity values (closest to 1) are grouped to represent the user-desired ratio. This step of the autoRA algorithm describes STEP 3.2 (Figure 5). The target area argument allows the user to enter a list of percentage values. We selected the area ratios of 10%, 20%, 30%, 40%, and 50% for the sensitivity analysis to demonstrate the behavior of the target area on soil predictions (S_{RA}).

The search process iterates through various block size values used to calculate the Gower dissimilarity index. For instance, it might start with a larger grouping, like 100 × 100 pixels. At this coarser resolution, the smoothed dissimilarity values reveal broad spatial patterns. In contrast, using a smaller block size, such as 5 × 5 pixels, produces a more detailed, fine-resolution map of Gower dissimilarity.

The autoRA algorithm systematically explores different area sizes by cycling through all values provided in the block size parameter. For example, if the target area is set to 10%, the algorithm applies the same pixel grouping defined by the block size. It then increases the target area by 20% and repeats the process, continuing this pattern until all specified block sizes and target area values have been used. As a result, the algorithm generates multiple RA formats by combining each target area value with each block size. This approach allows autoRA to capture a range of spatial patterns at different resolutions and area sizes.

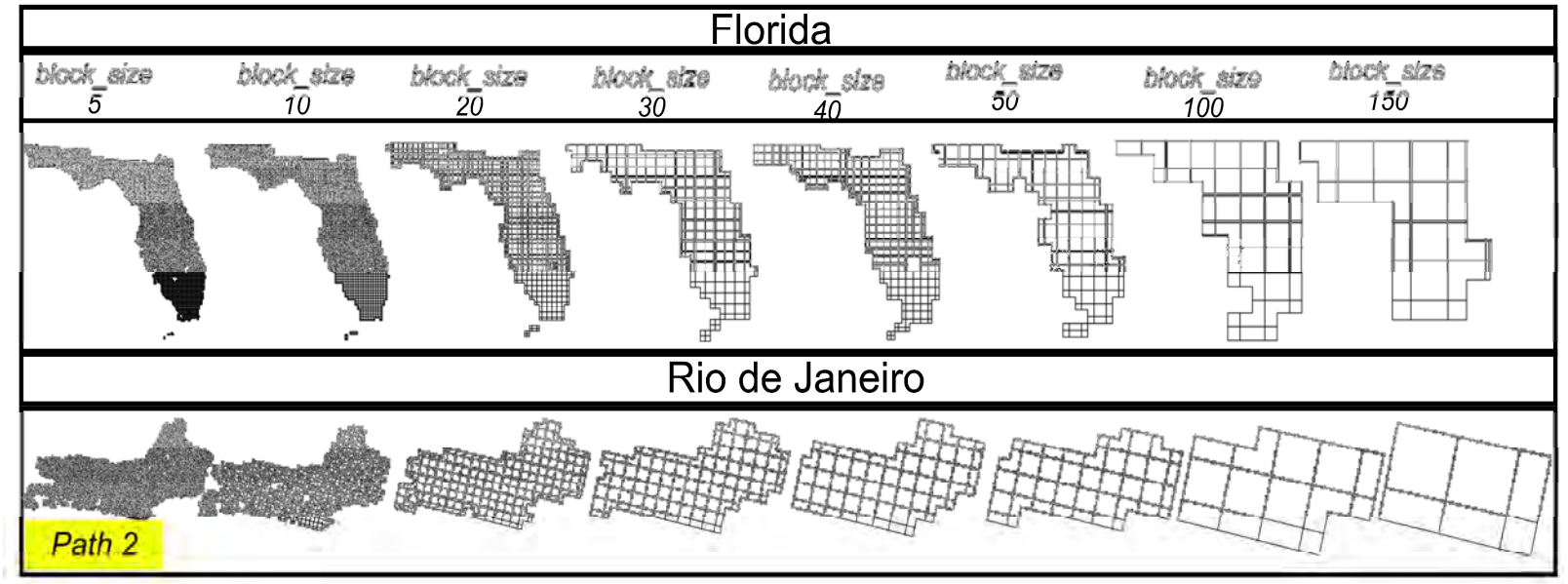


Figure 5. Block sizes and masks were used to calculate the Gower dissimilarity index in Florida and Rio de Janeiro (Path 2, STEP 3.1 of Figure 1).

2.4.4 Prediction and accuracy of modeling the exhaustive and reference area dataset

According to Path 1, the autoRA algorithm used 500 points extracting dimensionless values from the STS for each x and y coordinate, allowing the building of the benchmark model EPM that covered the full extension of the AOI. The exact number of points (N: 500) was also selected in Path 2 to model soil property S_{RA} for each RA (STEP 3.3 in Figure 1). We used the conditioned Latin hypercube sampling method (MINASNY; MCBRATNEY, 2006) with the maps of pedology, geology, temperature, precipitation, and the digital elevation model as inputs to place the 500 site locations for each of the two study areas (Florida and Rio de Janeiro). Random Forest machine learning was used for training prediction models at the 500 sites with covariates as inputs and S_{RA} as output. Separate RF models were created for each of the study areas. Finally, these models were upscaled to the entire Florida and Rio de Janeiro study region, creating RAM rasters.

We used autoRA to validate the EPM and RAM rasters created in STEP 2.3 of Path 1 and STEP 3.4 of Path 2, respectively. The same independent validation dataset (N: 100) identified in STEP 2.1 B of Path 1 was used to assess EPM and RAM via external validation. The metrics used to evaluate accuracy were the Root of Mean Squared Error (RMSE) and Bias, and the Adjusted Coefficient of Determination (R^2) was used to quantify the model fit.

The RMSE (Equation 4) measures the average magnitude of the errors between predicted and observed values; hence, values close to 0 indicate better model accuracy. Bias (Equation 5) quantifies the systematic error in the prediction over an external validation dataset, representing the average difference between the predicted and observed values. A Bias value close to zero indicates the absence of a systematic trend during the adjustment of the prediction model. The R^2 (Equation 6) means the proportion of variance in the training dataset that the model explains. Values close to 1 indicate that the adjusted model has a high explanatory capacity.

$$\mathbf{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}}} \quad \text{Eq. (4)}$$

$$\mathbf{Bias} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{\hat{y}_i} - \boldsymbol{y_i}) \quad \text{Eq. (5)}$$

$$\mathbf{Adjusted\ R^2} = \mathbf{1} - \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \boldsymbol{\bar{y}})} \quad \text{Eq. (6)}$$

where y_i are the values of the target variable simulated surface extracted for each of the 100 points intended for the composition of the dataset for external validation; $\hat{y}_i$ are the simulated surface values predicted for the 100 external validation points from the prediction models for each combination of tested arguments such as block size and target area size; $\bar{y}$ is the average of the 100 observed simulated surface values for the external validation group; n is the number of observations present in the validation set.

The Euclidean Distance (ED) was calculated to synthesize the metrics presented in Equations 4, 5, and 6. To calculate the Euclidean distance of the RMSE, Bias, and Adjusted R^2 metrics, it was essential to first scale them. Normalization ensures that all metrics contribute equally to the distance calculation, regardless of their original units or ranges. The escalation considered the maximum and minimum values present among all combinations of target area and block size used. Equations 7, 8, and 9 were used to normalize the RMSE, Bias, and R^2, respectively.

$$\mathbf{RMSE_{normalized}} = \frac{\mathbf{RMSE - minimum(RMSE)}}{\mathbf{maximum(RMSE) - minimum(RMSE)}} \quad \text{Eq. (7)}$$

$$\mathbf{Bias_{normalized}} = \frac{\mathbf{Bias - minimum(Bias)}}{\mathbf{maximum(Bias) - minimum(Bias)}} \quad \text{Eq. (8)}$$

$$\mathbf{R^2{}_{normalized}} = \frac{\mathbf{R^2 - minimum(R^2)}}{\mathbf{maximum(R^2) - minimum(R^2)}} \quad \text{Eq. (9)}$$

The ED was calculated to assess how close the normalized values are to the ideal standard ones: RMSE = 0, Bias = 0, and R^2 = 1. From the normalized RMSE, Bias, and R^2 values, the distances were calculated using Equation 10. A lower ED value indicates that the metrics are closer to the ideal values, suggesting a more accurate and less skewed model. In contrast, higher distance values signal a more significant discrepancy between the standard values.

$$\mathbf{Euclidian\ Distance} = \sqrt{(\mathbf{0 - RMSE})^2 + (\mathbf{0 - Bias})^2 + (\mathbf{1 - R^2})^2} \quad \text{Eq. (10)}$$

2.4.5 Cost simulations

We conducted a cost simulation to assess the practical efficiency of the autoRA algorithm. This simulation considered standard parameters influencing sampling logistics costs and planning during fieldwork. Specifically, the daily road mileage required to reach each sampling coordinate, the salaries of the necessary personnel, and the number of days needed to complete the sampling project were a product of the calculus between the road length required to reach all the points and the maximum distance threshold to be driven by day. We applied the EPM sampling cost simulation utilizing the whole length of the road network made available for the State of Florida (U.S. CENSUS BUREAU, 2019) and the State of Rio de Janeiro (IBGE, 2018a). The road network for both study areas is shown in Figure 6.

For the RAMs created from the dataset within each delineated RA from each combination of target area and block size, the shape of the roads was clipped using the respective RA extension as a mask to retain the roads inside it. The fuel cost was estimated at US $0.50 per kilometer traveled, with a maximum daily travel limit of 150 km. A Field Technician was considered to receive a salary of US $200 per day each.

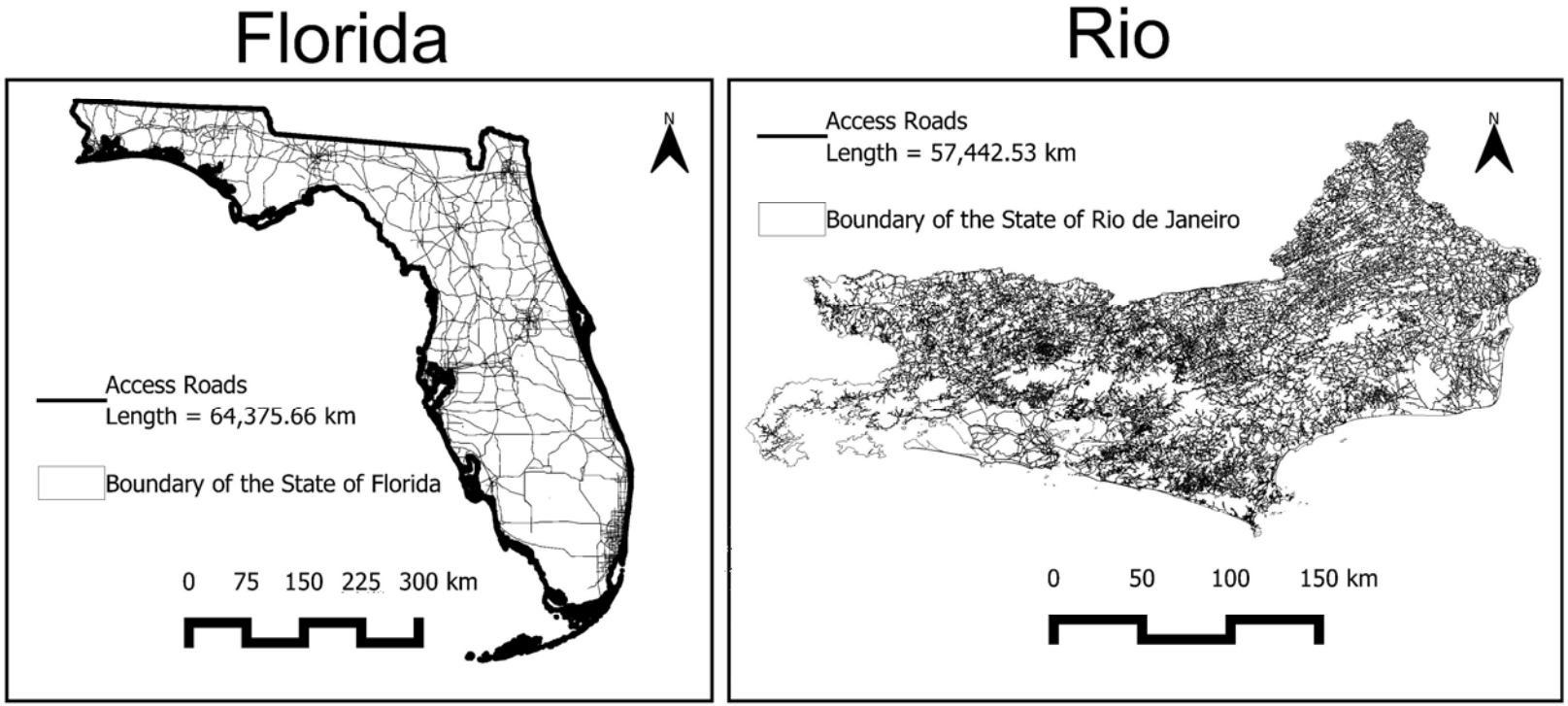


Figure 6. Main access roads for the states of Florida and Rio de Janeiro.

2.5. RESULTS AND DISCUSSION

2.5.1 Gower's dissimilarity index by block size

Results of the effect of varying block sizes on Gower dissimilarity index values were evaluated for Florida (USA) and Rio de Janeiro (Brazil). Different block sizes in the autoRA generate clipping masks from the covariate maps, influencing the aggregation of pixel values and the resulting Gower's Dissimilarity values. As block size increases, the calculated Gower's Dissimilarity Index values become progressively smoother due to the aggregation over larger blocks (Figure 7).

The Gower's Dissimilarity Index values across both regions range from 0.42 to > 0.63, with maps and a unified legend scale for direct comparison, as shown in Figure 7. In Florida, the highest Gower values (0.56 – 0.63) are primarily concentrated in the northwest region, mainly when smaller block sizes (5 – 30) are used. These areas become less distinct with larger block sizes (≥ 50) due to spatial smoothing.

In Rio de Janeiro, the city's western region records the highest Gower's Dissimilarity Index values, ranging between 0.56 and 0.63. This elevated dissimilarity coincides with the city's prominent mountainous landscape, notably Agulhas Negras Peak, which soars to 2,800 meters and is located within the protected Itatiaia National Park. Costa et al. (2024) reported that the western and mountainous areas of Rio de Janeiro were the most dissimilar. When they classified the state of Rio de Janeiro with dissimilar areas using a threshold of 0.34 of Gower's Dissimilarity Index.

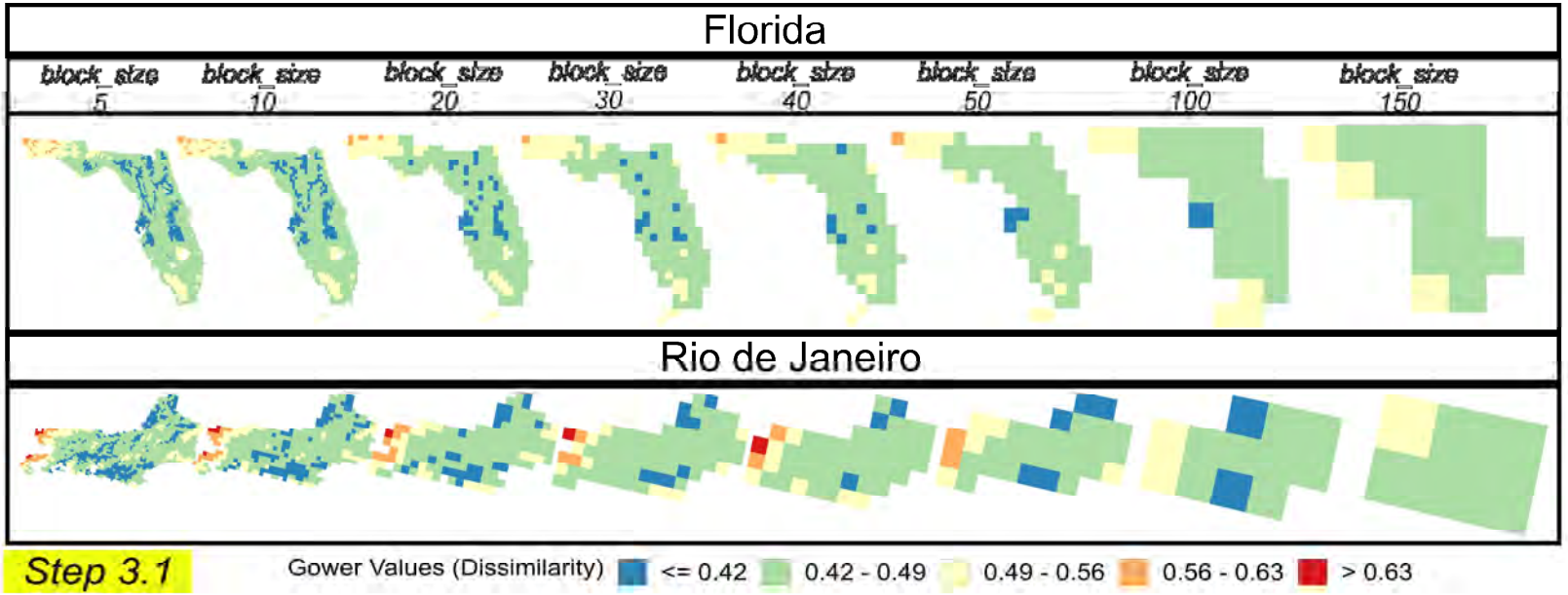


Figure 7. Gower's dissimilarities index map for block sizes and two study areas (Florida and Rio de Janeiro).

2.5.2 Reference areas' spatial distribution by block size and target area

Results from the implementation of STEP 3.2 delineating a variety of RAs within lower and upper bounds of block size and target area are shown in Figure 8. These findings offer a comprehensive perspective on how spatial resolution (block size) and coverage (target area) interact to influence RA delineation and sampling efficiency. For smaller target areas such as 5, 10, 20, and 30%, the delineation of the RA benefits from higher-resolution Gower's Dissimilarity Index values, as evident in the first row of Figure 8 (Florida) and Figure 9 (Rio de Janeiro). In these scenarios, finer block sizes highlight subtle physiographic gradients, producing more intricately defined RA boundaries and enabling a more detailed, granular representation of environmental heterogeneity.

As the target area increases to 40% and 50%, the delineated RAs encompass a broader range of physiographic information, effectively approaching the modal conditions of the entire region. The autoRA algorithm's ability to scale from finer resolutions (yielding more detailed boundaries and subtle distinctions) to broader coverage (capturing widespread physiographic features) sets it apart from existing methods like CLAPAS and conditioned Latin Hypercube Sampling (cLHS). Unlike CLAPAS, which requires manual input of candidate RAs and lacks full automation, autoRA systematically evaluates environmental variability through Gower's dissimilarity index, enabling a more holistic and flexible approach to RA delineation (JEAN-MARC ROBBEZ-MASSON, 1994). Additionally, while cLHS ensures broad initial coverage, it does not dynamically adjust to the spatial heterogeneity of the landscape, potentially leading to redundant sampling or missed environmental gradients (MALONE; MINANSY; BRUNGARD, 2019; MINASNY; MCBRATNEY, 2006).

As the RA encapsulates the more diverse soil-forming factor represented by the environmental variables used as input on the autoRA algorithm, the 500 training points within each RA are exhibited in Figures 8 and 9. They are expected to predict the STS by the RAM and the EPM. By doing so, autoRA enhances the scalability and efficiency of DSM workflows, particularly in diverse and challenging landscapes like those in Florida and Rio de Janeiro. This adaptability is essential for DSM practitioners seeking to optimize sampling designs in regions where traditional exhaustive sampling is neither feasible nor cost-effective (BRUS, 2014; HEIL; SCHMIDHALTER, 2017).

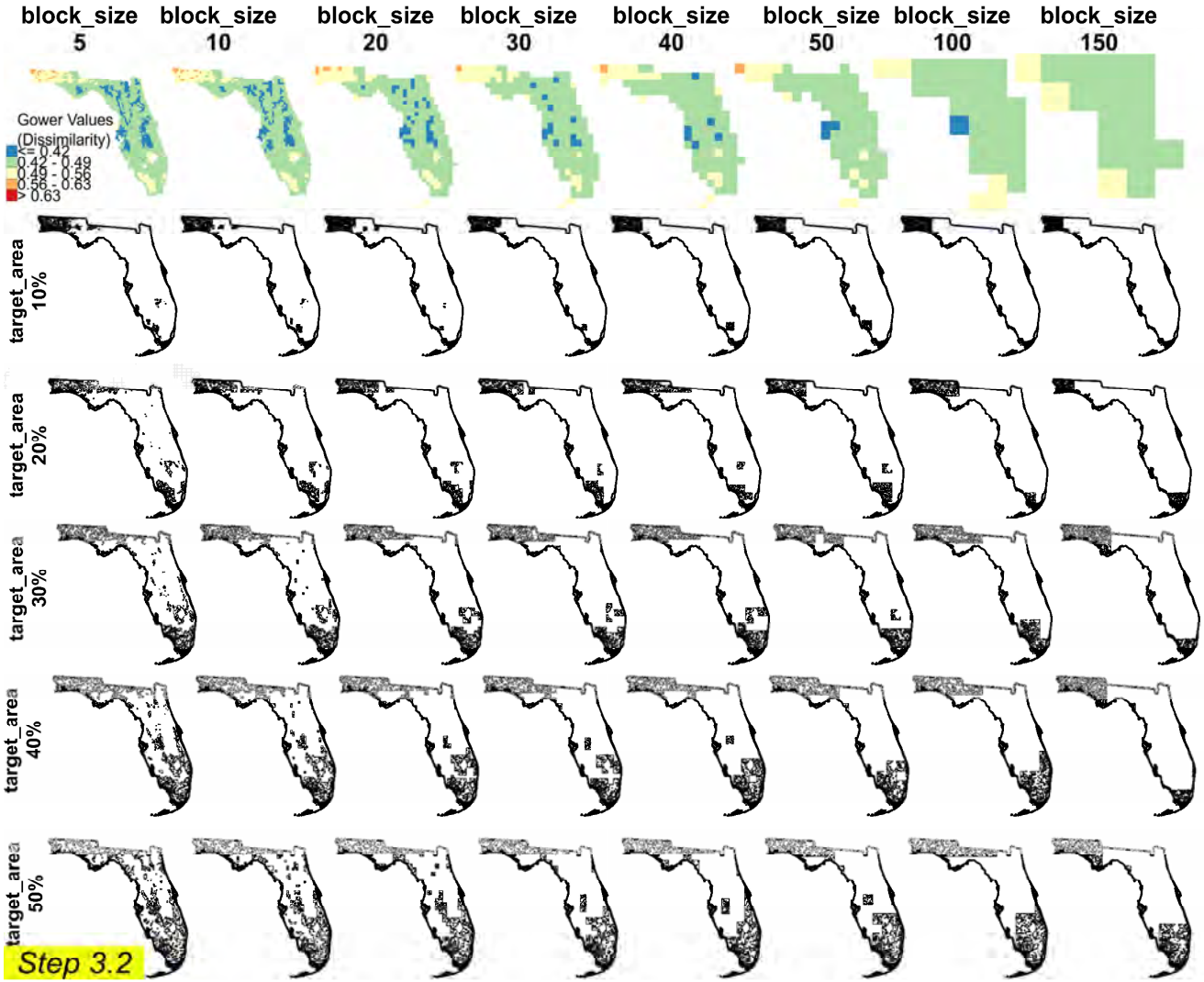


Figure 8. Delimitation of the reference areas (STEP 3.2) and placement of the training points within each reference area for Florida. Combinations of block size (rows) and target area (column) are shown.

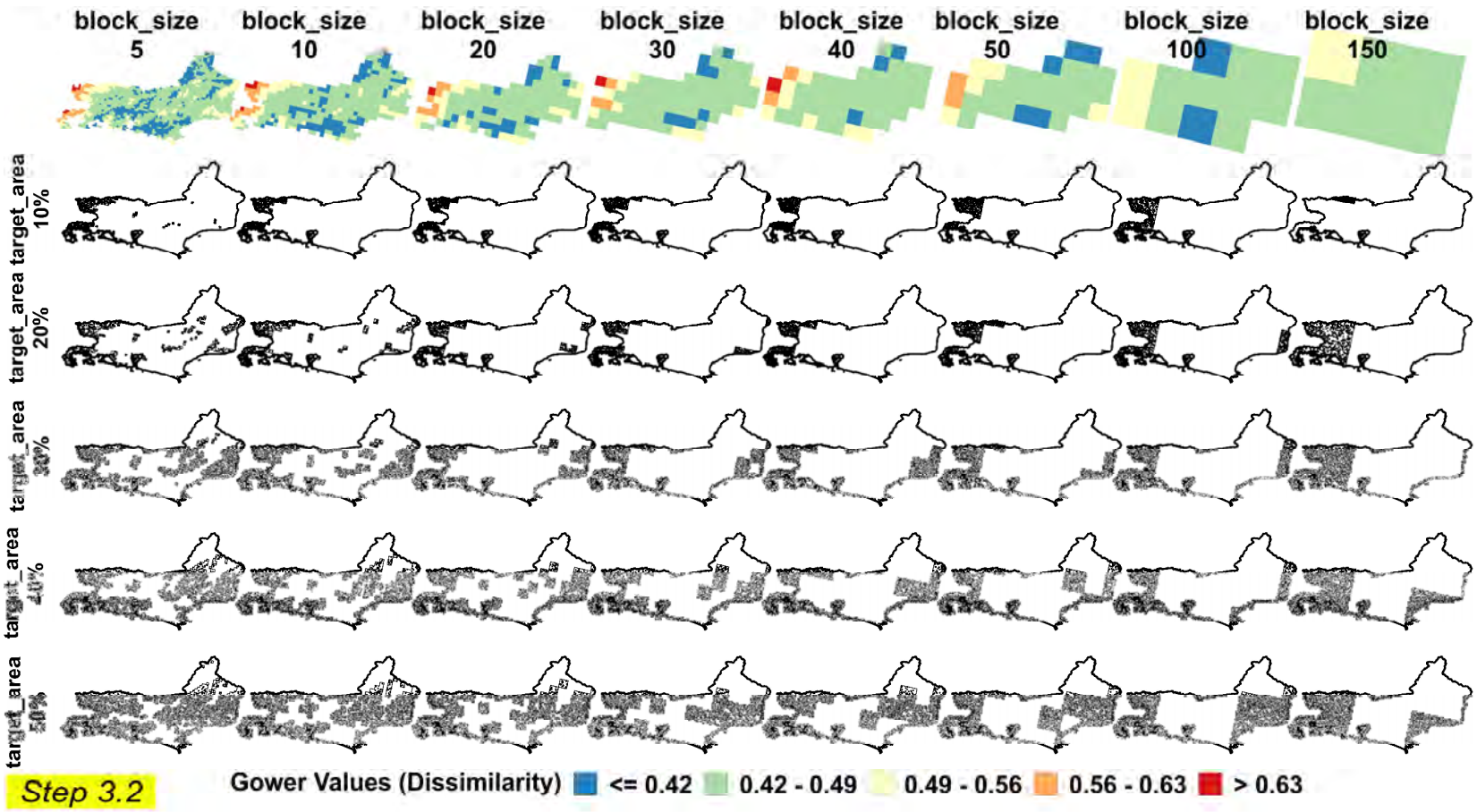


Figure 9. Delimitation of the reference areas (STEP 3.2) and placement of the training points within each reference area for Rio de Janeiro. Block size (rows) and target area (column) are combined.

2.5.3 Reference area selection based on metrics and cost

Finding the optimized RA parameters was fundamental to ensuring the accuracy and efficiency of our RAM predictive models after testing several combinations of target areas and block sizes. This section presents a detailed analysis of various RA configurations based on key performance metrics—R^2, RMSE, and Bias—and incorporates the Euclidean Distance (ED) metric alongside cost simulations to guide the selection process. These configuration results are compared to EPMs, which served as our benchmark by sampling the entire study areas. It is important to remember that all the R^2, RMSE, and Bias in Figure 10 are presented in a scale format (varying from 0 to 1) so they can be compared with the benchmark metric values.

In Figure 10, larger target area sizes consistently exhibit higher R^2 values, demonstrating enhanced explanatory and model fit compared to smaller target area values. For instance, the RAM 50% target area size achieves the highest R^2, closely approaching the EPM benchmark model's performance. The RMSE assesses the average magnitude of prediction errors, with lower values signifying more accurate predictions. Figure 10 shows a clear trend where larger target areas yield lower RMSE values, indicating improved prediction precision. The RAM 50% target area size records the lowest RMSE, suggesting that increasing the target area size significantly reduces prediction errors. The Bias measures systematic errors in predictions, reflecting whether the model overestimates or underestimates the observed values. A Bias value close to zero is desirable, as it indicates systematic minimal misprediction. The analysis reveals that larger target areas tend to have Bias values nearer to zero, highlighting their capability to provide more balanced and unbiased predictions. For example, the 40% and 50% target area sizes exhibit the smallest Bias values, underscoring their reliability for predicting accurately outside the RA-delineated boundaries.

The ED results (Figure 10) calculated for each RAM demonstrated a decreasing trend as the target area size increased. However, costs also rose because a larger target area encompassed more roads and required more time to drive to all recommended sampling points. The ED results using the RA approach were notably similar for Florida and Rio de Janeiro. The smallest ED values for RAM were 0.38 and 0.15, respectively, achieved with a target area of 50% and a block size 10. The EPM benchmark model showed slightly lower ED values of 0.35 for Florida and 0.17 for Rio de Janeiro. The slightly higher metric of the ED compared to the Rio's could be addressed by the randomization process of sampling the 500-training dataset for the EPM.

By limiting the sampling to 50% of the total study area for Florida and Rio de Janeiro, the RAM approach resulted in a total cost reduction of approximately $110,000 compared to the EPM approach. The traditional EPM strategy incurred costs of $258,491 for Rio de Janeiro and $289,690 for Florida. In this way, the RAM provided by autoRA with a target area of 50% and a block size of 10 represents a cost reduction of approximately 57% for Rio de Janeiro and 62% for Florida, highlighting the financial efficiency of the RA-approach supported by the autoRA automatization.

Consequently, the following results and discussion in the paper will focus on the block size of 10 and the target area of 50%, as these parameters yield the lowest ED values. Figure 11 presents the final outlined RAs for Florida and Rio de Janeiro. The access roads and sampling points within the RA are also overlayed in Figure 11.

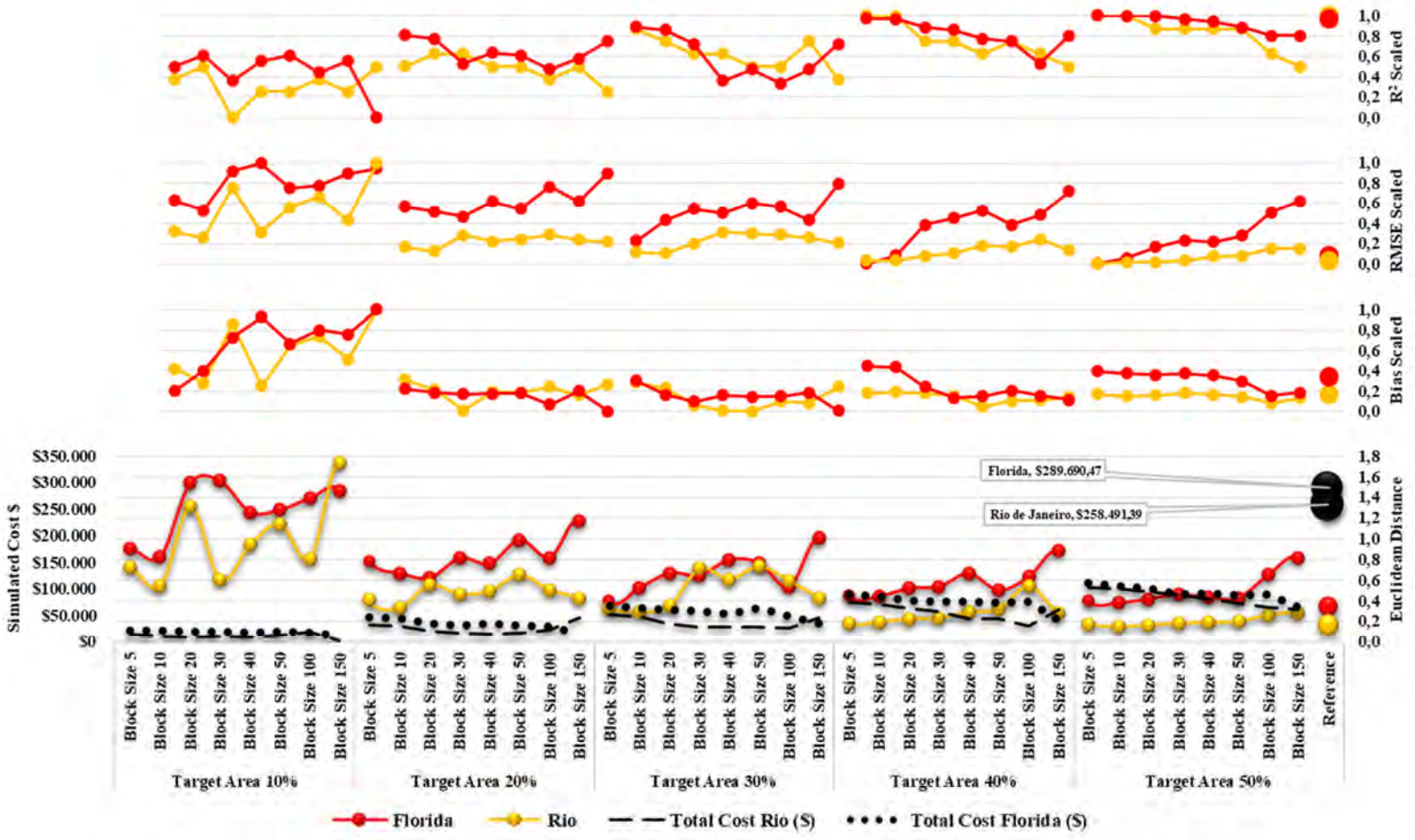


Figure 10. Metrics R2, RMSE, Bias, ED, and simulated cost for each combination of the target area and block size for the autoRA's configuration for Florida and Rio de Janeiro.

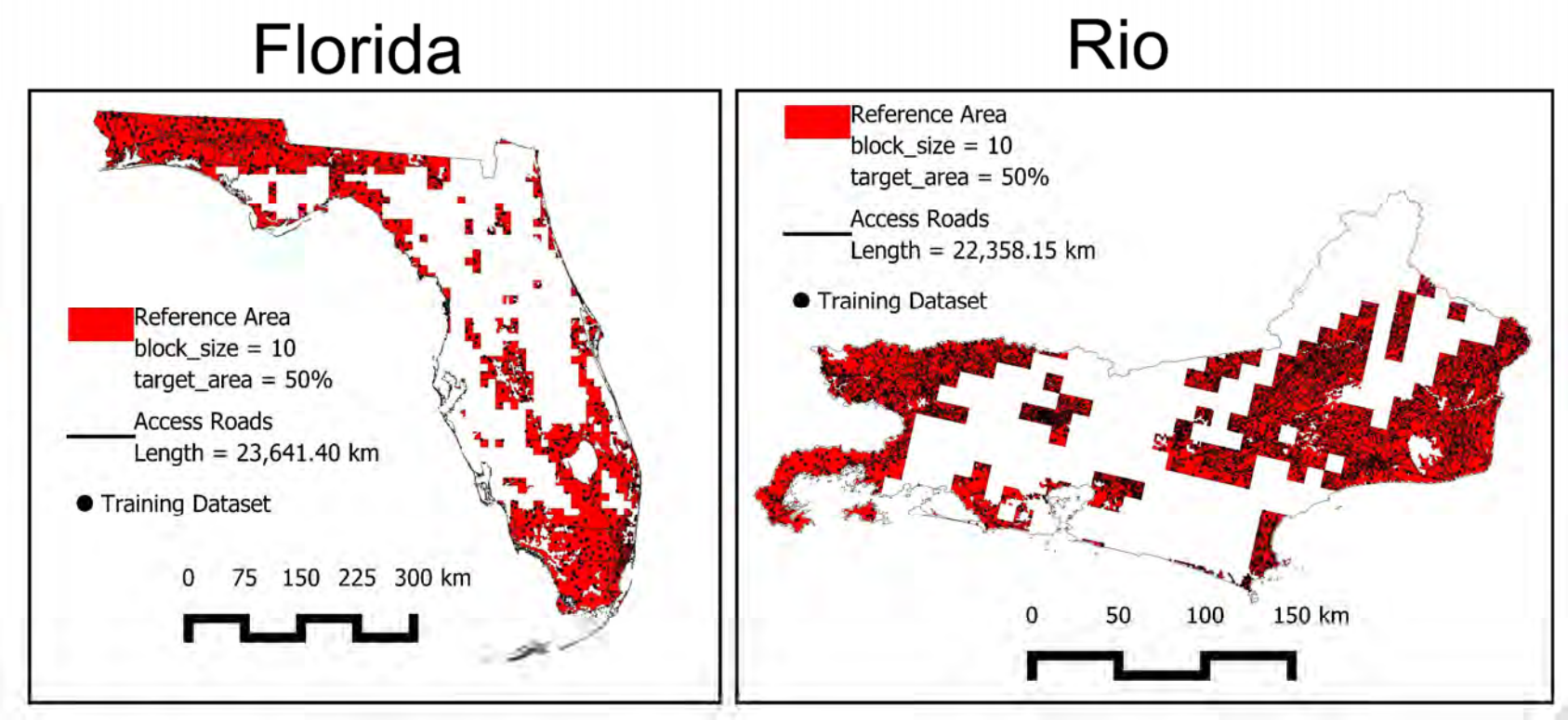


Figure 11. The Reference Area block size = 10, the target area 50% chosen for Florida and Rio, and the soil sampling placement of the training dataset (N: 500).

2.5.4 Florida reference area and predicted simulated theoretical surface analysis

Figure 12 displays the masked covariate maps for Florida's delineated RAM, generated using the autoRA algorithm with a target area of 50% and a block size of 10. These maps illustrate how the autoRA algorithm retrieves the variability of key SCORPAN factors—parent material (Geol), soil type (Pedo), elevation (DEM), precipitation, and temperature—across the state.

Regions dominated by sandy Entisols in well-drained uplands and coastal dunes, such as those captured in the Geol and Pedo maps, contrast sharply with the organic-rich Histosols in the poorly drained Everglades wetland soils in southern Florida. Similarly, the DEM map highlights low-relief areas associated with wetland hydrology and flatwood systems. In contrast, the precipitation and temperature maps emphasize climatic gradients that influence soil development across the state. These masked covariate maps demonstrate the algorithm's ability to prioritize areas with diverse SCORPAN factor interactions while preserving spatial coherence.

Temperature variability in the RAM aligns closely with the EPM, with near-identical frequency distributions across the temperature range, ensuring that climatic gradients influencing soil formation are adequately captured (Figure 13). The precipitation distribution also reflects strong alignment, indicating that both drier and wetter regions are well-represented, which is crucial for capturing hydrologically driven soil patterns. Elevation variability is similarly preserved, with the RAM accurately reflecting the low-lying and upland areas characteristic of Florida's topography, though minor underrepresentation is noted in the higher elevations. Figure 14 demonstrates the maps for the STS predicted for Florida using the RAM with the lowest ED metric from the combination target area of 50% and block size 10 and compares its spatial SPS distribution with the SPS map predicted by the sampling strategy of the EPM that worked as the benchmark map.

Figure 14 demonstrates the percentage frequency of pixel values retrieved in the masked covariate maps, comparing the RAM-encapsulated pixels selected with the lowest ED (target area 50% and block size 10) to the EPM pixel for the whole study area of Florida. The results show that RAM effectively represents the variability of all covariates, ensuring that the selected RA encompasses the diversity observed in the entire dataset. For pedology, RAM preserves the distribution of dominant classes, such as Entisols and Spodosols, while also including less frequent classes, like Histosols, reflecting comprehensive soil variability. Similarly, geological variability is well-represented, with central lithological units such as Holocene sediments and residuum included, although minor deviations are observed for specific formations like the Hawthorn Group.

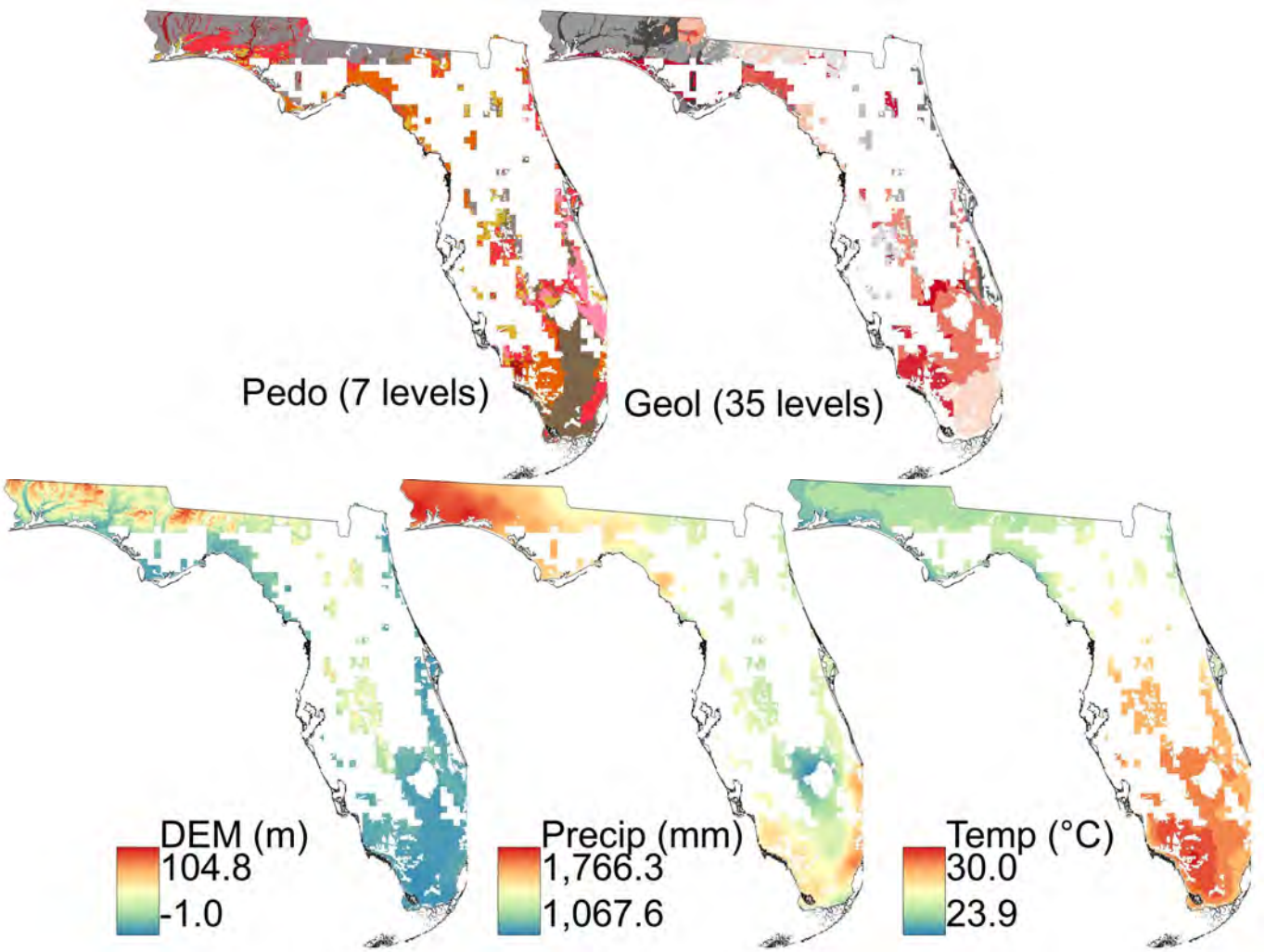


Figure 12. Covariables cropped for Florida's reference area, with a block size of 10 and a target area of 50% delineated by the autoRA.

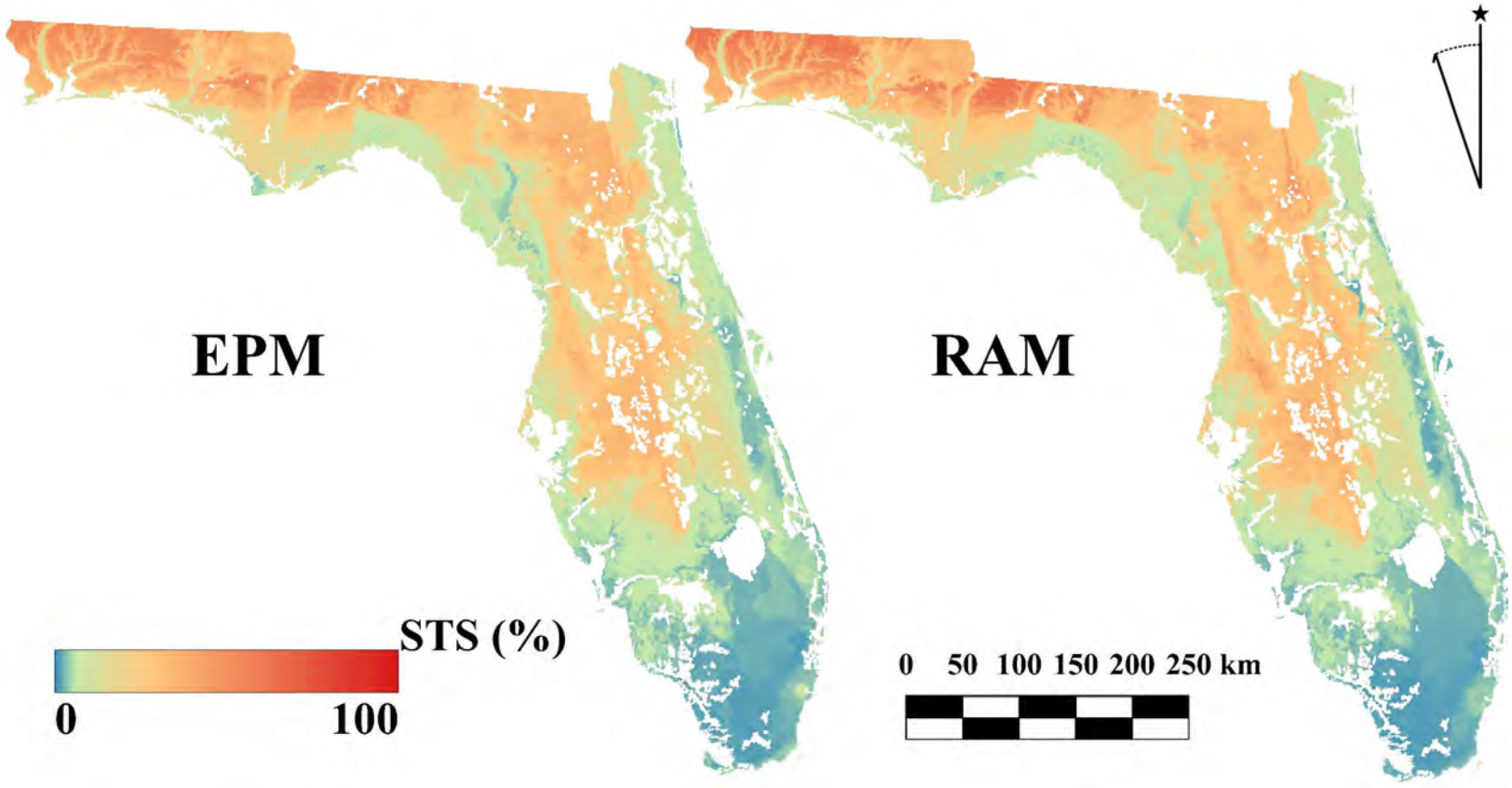


Figure 13. Comparison for the predicted Simulated Theoretical Surface (STS) maps via Exhaustive Prediction Model (EPM) using the whole area sampling strategy and Reference Area Model (RAM) using autoRA best Euclidean Distance metric for Florida.

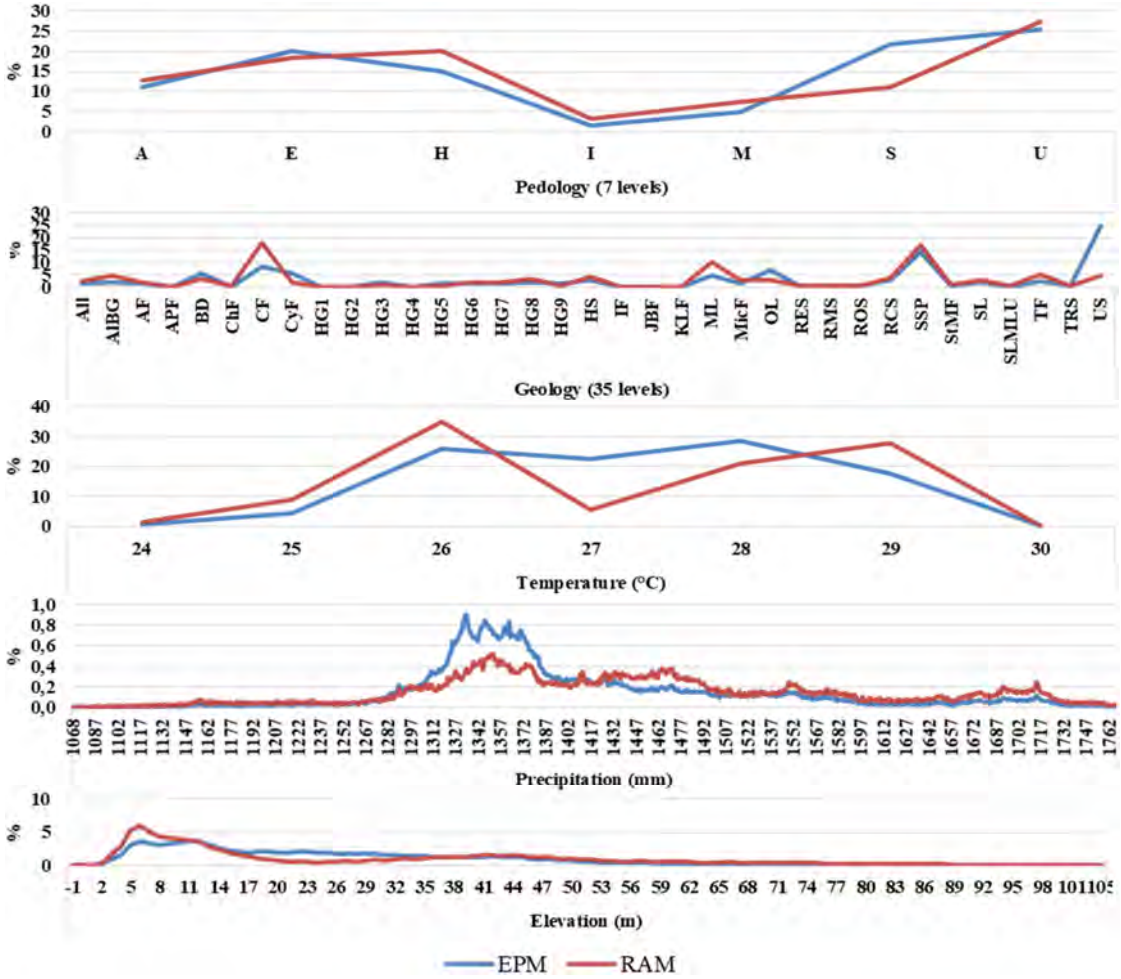


Figure 14. Frequency of pixel information for Exhaustive Prediction Model (EPM) and Reference Area Model (RAM). Pedology map: A, Alfisols; E, Entisols; H, Histosols; I, Inceptisols; M, Mollisols; S, Spodosols; U, Ultisols; Geology map: All, Alluvium; AlBG, Alum Bluff Group; AF, Anastasia Formation; APF, Avon Park Formation; BD, Beach ridge and dune; ChF, Chattahoochee Formation; CF, Citronelle Formation; CyF, Cypresshead Formation; HG1, Hawthorn Group, Arcadia Formation; HG2, Hawthorn Group, Arcadia Formation, Tampa Member; HG3, Hawthorn Group, Coosawhatchie Formation; HG4, Hawthorn Group, Coosawhatchie Formation, Charlton Member; HG5, Hawthorn Group, Peace River Formation; HG6, Hawthorn Group, Peace River Formation, Bone Valley Member; HG7, Hawthorn Group, Statenville Formation; HG8, Hawthorn Group, Torreya Formation; HG9, Hawthorn Group, Undifferentiated; HS, Holocene sediments; IF, Intracoastal Formation; JBF, Jackson Bluff Formation; KLF, Key Largo Limestone; ML, Miami Limestone; MicF, Miccosukee Formation; OL, Ocala Limestone; RES, Residuum on Eocene sediments; RMS, Residuum on Miocene sediments; ROS, Residuum on Oligocene sediments; RCS, Reworked Cypresshead sediments; SSP, Shelly sediments of Plio-Pleistocene age; StMF, St Marks Formation; SL, Suwannee Limestone; SLMLU, Suwannee Limestone-Marianna Limestone undifferentiated; TF, Tamiami Formation; TRS, Trail Ridge sands; US, Undifferentiated sediments.

2.5.5 Rio de Janeiro reference area and predicted simulated theoretical surface analysis

The delineation of RA with the target area of 50% and block size 10 provided the lowest ED metric for Rio de Janeiro, and its respective masked covariate maps are shown in Figure 15. The masked maps of pedology, geology, elevation, precipitation, and temperature illustrate how the RA prioritizes regions with distinct soil-forming factors.

The temperature and precipitation gradients are driven by the state's varied elevation and climatic patterns (NEIVA; DA SILVA; CARDOSO, 2017). Parent material, including crystalline rocks in the Serra do Mar and sedimentary deposits in the coastal plains, further drives the variability in soil mineralogy and texture, as highlighted in the geology map (PINHEIRO JUNIOR et al., 2021) (Figure 15, Geol). The pedology map (Figure 15, Pedo) underscores the diversity of soil types, from highly weathered Oxisols in upland regions to sand Entisols in the coastal plains, capturing the stark transitions driven by relief and parent material. Additionally, the elevation map reflects the role of topography in shaping soil formation, where steep slopes favor shallow soils like Entisols, while flatter areas support deeper, weathered soils (FONTANA et al., 2017; GONÇALVES, Rogério Victor S. et al., 2022).

The frequency distributions of pixel values for the RA-masked covariates and the EPM in Rio de Janeiro are shown in Figure 16. It reveals that the pixels inside the RA chosen for Rio de Janeiro (RAM, target area 50% and block size 10) encapsulated the same pixel variability in the whole Study Area (EPM) extension. RAM represented all 21 soil classes for pedology, including dominant soil orders such as Oxisols and Ultisols and less prevalent ones like Entisols. This ensures that the RAM dataset encompasses the full range of soil variability, preserving critical transitions between highly weathered upland soils and less developed sandy soils in the coastal plains.

Similarly, the geology covariate, which includes nine lithological classes, is well-represented in RAM by capturing the central geological units, such as crystalline rocks and sedimentary deposits, strongly influencing soil mineralogy and texture across the landscape. The RAM effectively captures the full range of variability observed in the EPM for climatic variables, such as temperature and precipitation. Temperature gradients from 20°C to 24°C and precipitation values between 1,116 mm and 1,250 mm are consistently represented, ensuring that the climatic influences on soil formation, such as leaching and organic matter accumulation, are adequately accounted for (Figure 16).

The elevation covariate (Figure 16, DEM), which spans from sea level to 2,049 meters, is similarly preserved in RAM. The frequency distribution indicates a proportional representation of low-lying areas, mid-elevations, and higher terrains, reflecting the dynamic role of relief in shaping soil properties. Steep slopes associated with shallow, eroded soils (e.g., Entisols) and flatter regions where deep weathering occurs (e.g., Oxisols) are included in the RAM dataset, ensuring that topographically driven variability is maintained.

Figure 17 demonstrates that the RAM delineated by autoRA with a target area covering 50% of the total Study Area can effectively map the SPS for Rio de Janeiro. It produces results nearly identical to EPMs, significantly saving time and resources. Visual comparisons highlight the similarity between the two approaches, underscoring RAM's potential for efficient and accurate environmental mapping and offering a cost-effective alternative to EPM without compromising quality.

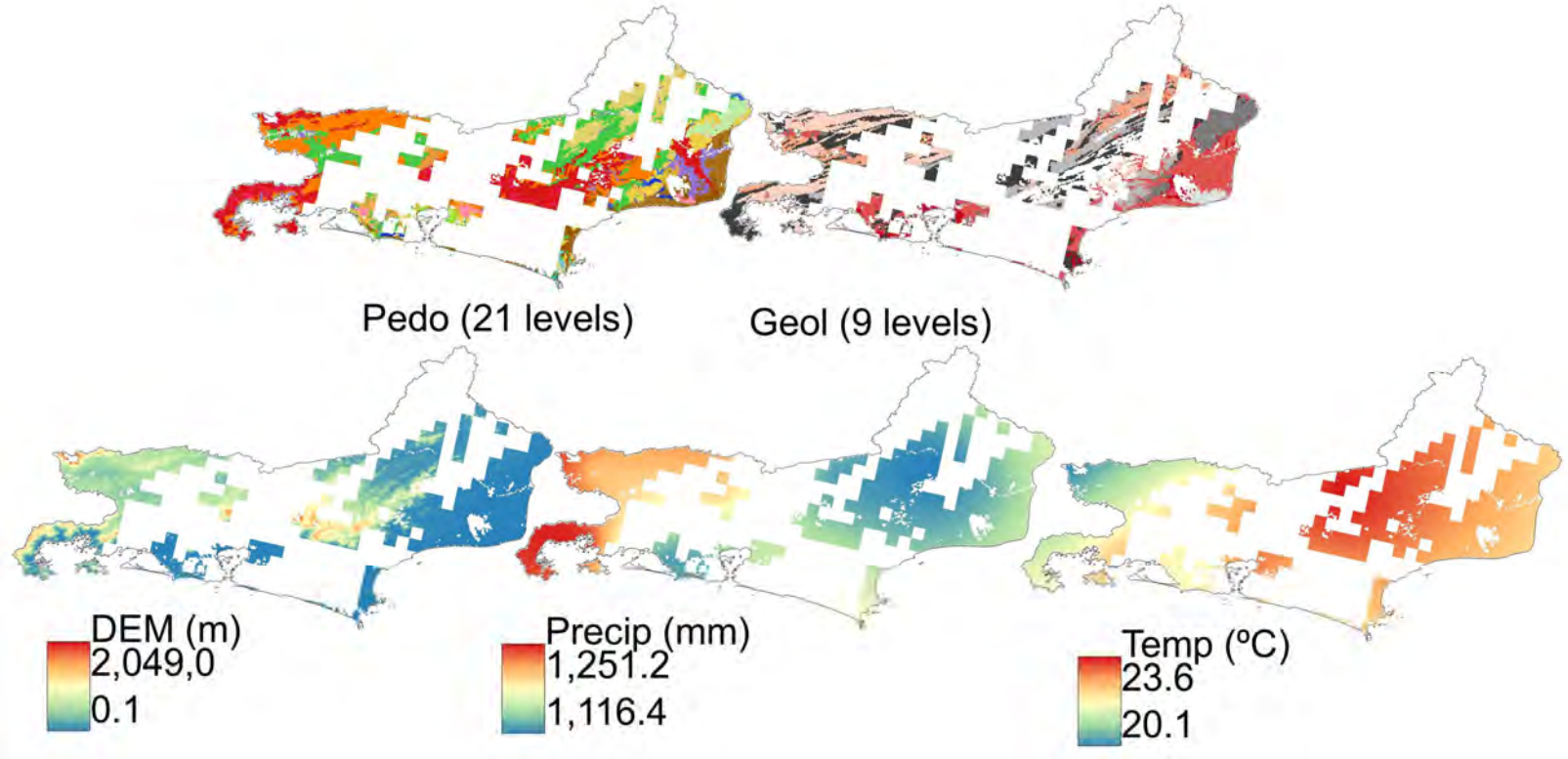


Figure 15. Covariables cropped for Rio de Janeiro's reference area, with a block size of 10 and a target area of 50% delineated by the autoRA.

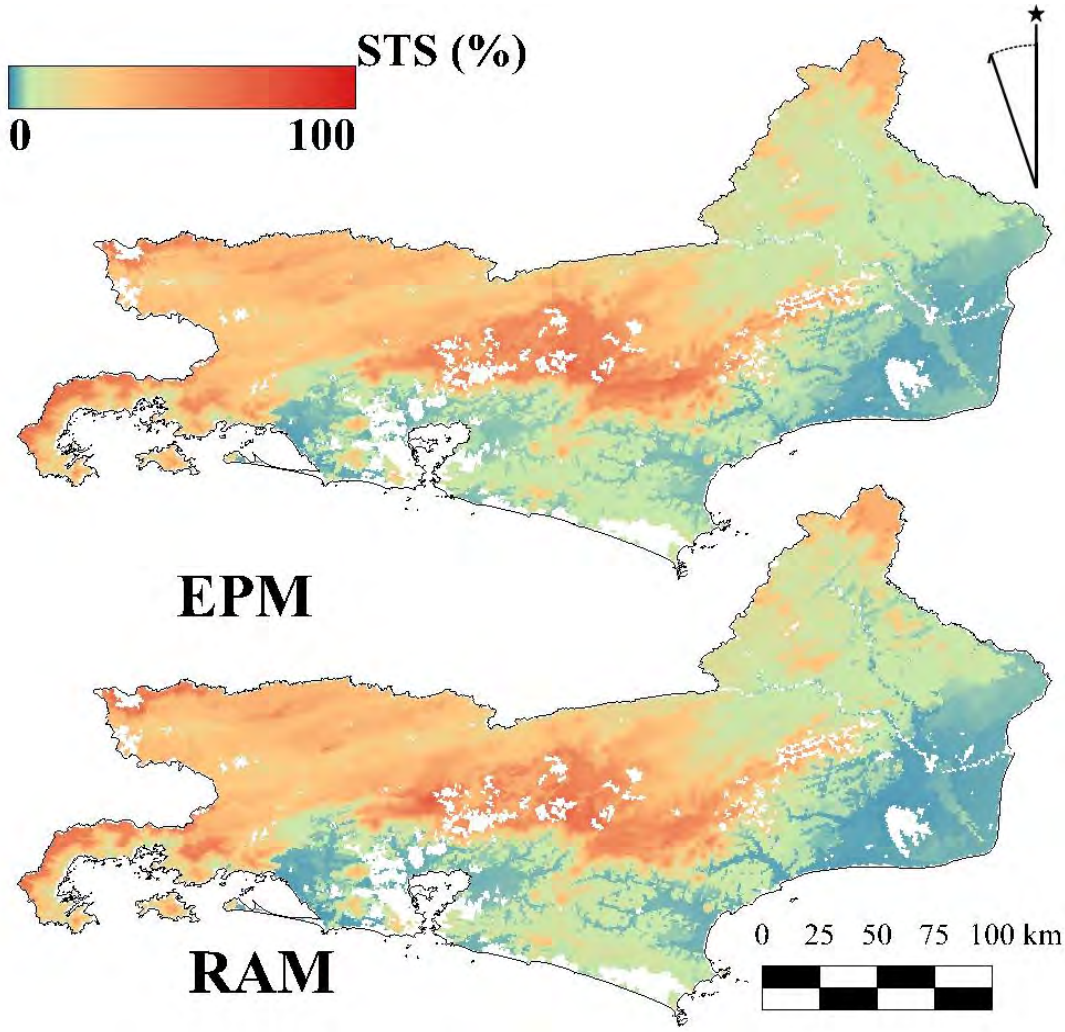


Figure 16. Comparison for the predicted Simulated Theoretical Surface (STS) maps via Exhaustive Prediction Model (EPM) using the whole area sampling strategy and Reference Area Model (RAM) using autoRA for Rio de Janeiro.

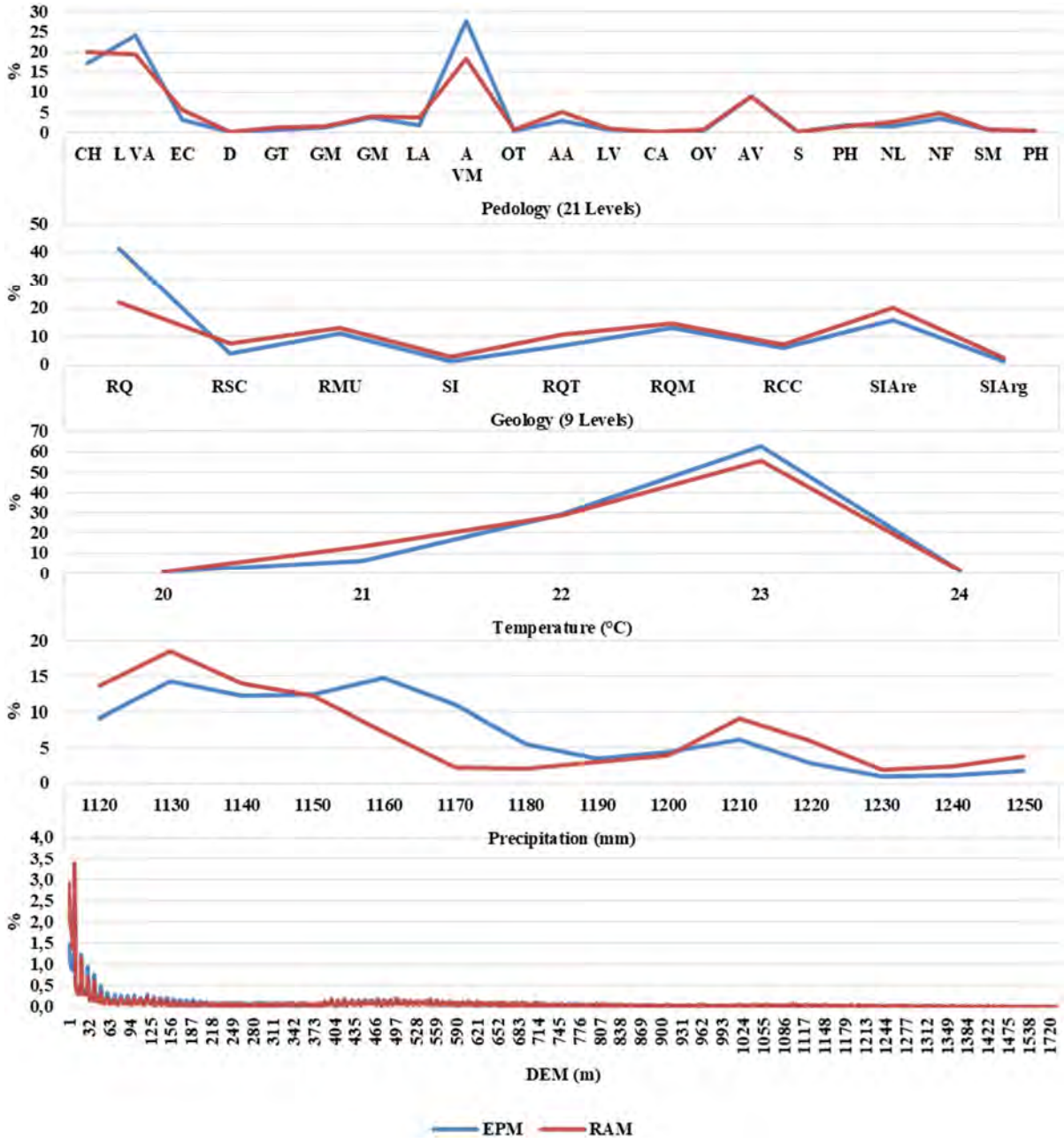


Figure 17. Frequency of pixel information for Exhaustive Prediction Model (EPM) and Reference Area Model (RAM). Pedology map: CH, Cambissolo Háplico; L VA, Latossolo Vermelho-Amarelo; EC, Espodossolo Cárbico; D, Dunas; GT, Gleissolo Tiomórfico; GM, Gleissolo Melânico; GH, Gleissolo Háplico; LA, Latossolo Amarelo; A VM, Argissolo Vermelho-Amarelo; OT, Organossolo Tiomórfico; AA, Argissolo Amarelo; LV, Latossolo Vermelho; CA, Chernossolo Argilúvico; OV, Organossolo Háplico; AV, Argissolo Vermelho; S, Salinas; PH, Planossolo Hidromórfico; NL, Neossolo Litólico; NF, Neossolo Flúvico; SM, Solos Ind. Mangues; PH, Planossolo Háplico. Geology map: RQ, Quartz-feldspathic rocks; RSC, Clastic sedimentary rocks; RMU, Mafic, and ultramafic rocks; SI, Unconsolidated sediments; RQT, Quartzose rocks; RQM, Micaeous quartz-feldspathic rocks; RCC, Carbonatic and calcium-silicate rocks; SIAre, Sandy unconsolidated sediments; SIArg, Clayey unconsolidated sediments.

2.5.6 Evaluating autoRA: contrasts and synergies with established sampling approaches in digital soil mapping

The autoRA represents a novel strategy within the broader field of DSM, which has seen numerous methodologies proposed to optimize sampling designs, balance cost efficiency, and maintain robust predictive performance. To better understand how autoRA aligns with or diverges from the current sampling approaches, this section contrasts autoRA's methodology with four main lines of work: (i) conditioned Latin hypercube sampling (cLHS), (ii) Homosoils, (iii) sampling designs optimizing variance between population and sample sets (STUMPF et al., 2016), and (iv) divergence-based approaches for determining sample size (MALONE; MINANSY; BRUNGARD, 2019; SAURETTE et al., 2023)Although each approach seeks to capture environmental heterogeneity effectively, it differs substantially in its theoretical underpinnings, implementation, and adaptability to varying soil-landscape contexts.

2.5.7 Contrasts with conditioned Latin hypercube sampling

Conditioned Latin hypercube sampling (cLHS) has long been recognized as a robust technique for generating a stratified random sample across relevant covariates (MINASNY; MCBRATNEY, 2006). By projecting environmental variables into a multidimensional feature space, cLHS endeavors to sample each stratum equally, thereby ensuring coverage of the covariate distribution (SENA et al., 2021). However, cLHS presupposes a priori fixed number of samples, which can be problematic in large or heterogeneous regions. Once the number of samples is decided, cLHS does not inherently recalibrate or refine its sampling plan based on new information about soil variability (MALONE; MINANSY; BRUNGARD, 2019). This limitation may lead to the oversampling of relatively uniform areas or failure to capture underrepresented yet pedologically significant zones if the initial stratification proves suboptimal.

In contrast, autoRA continuously gauges spatial heterogeneity via Gower's Dissimilarity Index. By systematically delineating RAs and testing them against performance metrics such as R^2, RMSE, and Bias, combining them with a sensitivity analysis of different target areas and block sizes, autoRA iteratively refines its sampling scope before the fieldwork itself. Hence, while cLHS frontloads the sampling design process, autoRA incorporates feedback loops that dynamically adjust reference-area boundaries and sampling densities. This responsive mechanism is particularly beneficial in regions where heterogeneity is high, it is not easy to access the sampling locations or is not evenly distributed, such as mountainous areas or wetlands (SAURETTE et al., 2023). Additionally, cLHS and autoRA are not mutually exclusive; in principle, autoRA's final delineated RAM could incorporate a cLHS-type scheme for spatial allocation of actual sample sites, if desired. Nonetheless, autoRA's core advantage is its real-time adaptability, which mitigates reliance on static sampling targets and helps manage logistical constraints (e.g., limited field operability and safety concerns).

2.5.8 Contrasts with homosoils

Homosoils, as presented by Nenkam et al. (2022, 2023), similarly uses Gower's dissimilarity index to cluster pedologically similar areas and direct sampling toward zones of maximum dissimilarity. In doing so, Homosoils aims to avoid oversampling landscapes that exhibit relative uniformity while focusing resources on capturing critical variability. This conceptual foundation parallels autoRA's objective of capturing diverse soil formation factors by prioritizing environmental heterogeneity.

However, Homosoils typically pre-specifies a sampling density or cluster count without explicitly integrating metrics of predictive accuracy—such as R^2, RMSE, or Bias—into its final selection process. In contrast, autoRA explicitly uses a Random Forest model on a simulated theoretical surface (STS), running multiple sensitivity analyses (block sizes, target areas) and computing an Euclidean Distance (ED) aggregate of prediction metrics. AutoRA chooses the "optimal" RA arrangement only after running these iterations. Consequently, autoRA identifies the region(s) of highest dissimilarity and verifies that sampling these regions demonstrably leads to predictive gains, thereby establishing a direct link between sampling design and model performance. This systematic feedback mechanism differentiates autoRA from Homosoils, offering a more robust criterion for deciding the optimal coverage while leveraging the core principle that areas of high Gower's Dissimilarity Index merit more careful sampling.

2.5.9 Contrasts with variance-based sampling designs

Stumpf et al. (2016) addressed the challenge of defining an optimal sample size by comparing the variance of the covariate population to that of the sample set. Their methodology incrementally increased sample size, using box plots and density plots of relevant covariates to identify a threshold at which additional sampling yielded diminishing returns in capturing population variance. This approach offers a transparent and intuitive means of sampling: once the sample's variance profile sufficiently approximates that of the population, the design is considered "good enough."

While straightforward and conceptually appealing, variance-based sampling primarily hinges on matching statistical moments of the covariate distribution, which may not always capture deeper or more complex pedological relationships (MALONE; MINANSY; BRUNGARD, 2019). For instance, variance equivalence does not necessarily account for underlying spatial patterns or the combined effect of covariates, which might be crucial in regions where soil properties are influenced by intricate interactions of climate, relief, and parent material (MCBRATNEY, A. B; MENDONÇA SANTOS; MINASNY, 2003). In contrast, autoRA's reliance on Gower's Dissimilarity Index and comprehensive metrics (R^2, RMSE, Bias) ensures that the final RA delineation does more than match a univariate or bivariate variance profile; it also demonstrates robust predictive fidelity for the soil attributes of interest by offering a derivate Simulated Theoretical Surface that is mapped, and the accuracy is evaluated before the fieldwork starts. Another distinguishing factor lies in autoRA's reliance on sensitivity analysis across multiple parameter settings rather than a single stepwise approach to sample size increments. This approach simultaneously refines both the size and shape of the RA, reducing the risk of focusing solely on variable variance while missing other dimensions of soil heterogeneity (STUMPF et al., 2016).

2.5.10 Contrasts with divergence-based approaches for determining sample size

Divergence-based approaches have gained attention for their potential in determining optimal sample size by comparing differences in probability distributions. Malone et al. (2019) employed the Kullback-Leibler Divergence (DKL) statistic to evaluate how closely a sample's empirical distribution function (EDF) approximates that of the larger population. By finding the point of "diminishing returns" in the DKL curve, one can infer an optimal sample size that balances coverage with practical resource limitations.

Building on this concept, Saurette et al. (2023) introduced the Jensen-Shannon Divergence (DJS) and the related Jensen-Shannon Distance ($Dist_{JS}$) as more robust, symmetric metrics for appraising how well a given sample distribution matches the population distribution. These divergence metrics require binning the data into histograms or probability distribution functions (PDFs) and comparing how closely the sample's PDF aligns with the entire domain.

In principle, DKL, DJS, or $Dist_{JS}$ can reveal the "breakpoint" beyond which additional sampling yields marginal improvements in distribution matching. Divergence-based methods thus offer a mathematically elegant solution to determining an "optimal" sample size that captures the principal features of the covariate space.

Yet, like variance-based techniques, divergence-based approaches often treat each covariate or histogram dimension independently (MALONE; MINANSY; BRUNGARD, 2019; SAURETTE et al., 2023). While they are more holistic than a single variance measure, they may not fully capture spatial autocorrelation patterns or complex covariate interactions that strongly influence soil genesis and variability. In contrast, autoRA applies a Random Forest framework to evaluate how well the delineated RAM can predict an STS, encapsulating multiple covariates simultaneously in a respective smaller region. The final selection of target area and block size is thus informed by direct modeling performance, not solely distribution matching. Consequently, autoRA can integrate the strengths of divergence-based analyses—identifying representativeness thresholds—while ensuring that this representativeness translates into tangible predictive accuracy. Indeed, future versions of autoRA could incorporate DJS or $Dist_{JS}$ as complementary indices alongside Gower's dissimilarity, providing an even more refined synergy between statistical distribution matching and predictive modeling.

2.5.11 Synthesis and outlook

Taken together, these comparisons underscore the distinctiveness and adaptability of autoRA. Conditioned Latin hypercube sampling (cLHS) ensures an even distribution of samples across covariate space but does not dynamically adjust to local heterogeneity or feedback from model performance. Homosoils likewise leverages Gower's Dissimilarity Index to detect uniform vs. highly variable areas but does not explicitly integrate predictive metrics into the sampling density decision nor shows the smaller area capable of being sampled and representing the interest study area. Variance-based sample size selection (STUMPF et al., 2016) provides a straightforward mechanism for aligning sample distributions with population variance but can overlook complex multidimensional relationships. It also does not consider the hypothesis of searching and minimizing the investigation area based on the Reference Area approach. Divergence-based approaches (MALONE; MINANSY; BRUNGARD, 2019; SAURETTE et al., 2023) offer mathematically rigorous methods for defining optimal sample sizes by comparing distribution functions. However, they may underserve spatial context or joint covariate interactions and do not consider the hypothesis of minimizing the sampling area to produce a model for extrapolation.

By contrast, autoRA weaves together the strengths of spatial dissimilarity assessment (via Gower's Dissimilarity Index), iterative modeling (via Random Forest) and Simulation Theoretical Surface, and sensitivity analyses (varying target areas and block sizes) into a single workflow. This ensures that representativeness, cost-effectiveness, and predictive reliability are simultaneously prioritized. Moreover, autoRA's capacity to include other divergence metrics or sampling heuristics signals a pathway for future enhancements, making it a flexible platform for integrating new advances in DSM. As a result, autoRA stands out not merely as another sampling design tool but as a dynamic framework that combines data-driven delineation of RAs with tangible model performance evaluation—critical for robust and scalable soil mapping in the face of limited ground-truth data.

2.6 CONCLUSIONS

The autoRA algorithm demonstrated a robust, data-driven approach for delineating RAs that represented critical soil-forming factors, enabled more efficient and accurate digital soil mapping workflows, and systematically identified configurations of target area size and spatial resolution (block size) that balanced predictive performance and cost by employing Gower's Dissimilarity Index to capture environmental heterogeneity. In the optimal RAM with a 50% target area and a block size of 10, autoRA achieved ED values (0.15 in Rio de Janeiro and 0.38 in Florida) that closely approximated the benchmarks obtained using exhaustive sampling (0.17 and 0.35, respectively) while reducing total costs by approximately US$110,000. This translated to cost reductions of about 61% in Rio de Janeiro and 63% in Florida compared to the traditional reference approach.

Beyond this optimal setting, several other combinations offered even more significant cost savings, albeit with marginal trade-offs in accuracy. For instance, at a 30% target area and a 10×10 km² block size, the model in Rio de Janeiro produced an ED of around 0.33. In contrast, for the same target area value, the resolution of 5×5 km² for Florida yielded an ED close to 0.40—slightly higher than the optimal scenario—, yet costs were cut by about 80%. Similarly, other parameter settings at smaller target areas (e.g., 20%) and moderate block sizes (e.g., 10 or 20 pixels) delivered substantial cost-efficiency while maintaining acceptable ED values. These findings highlighted autoRA's versatility, allowing practitioners to tailor the balance between accuracy and cost according to specific project constraints, logistical limitations, and data requirements.

By reducing subjective expert input and introducing a reproducible, quantitative framework for RA delineation, autoRA enabled more strategic investments in soil sampling. Its capacity to preserve predictive quality while substantially lowering expenses made it a valuable tool, particularly in regions where field sampling was logistically challenging or financially constrained. Ultimately, this approach strengthened DSM workflows, fostered broader coverage in data-scarce landscapes, and supported more informed decision-making in soil resource management.

In detailed mapping at a scale of 1:100,000, it is essential to distinguish between the mapping scale and the pixel resolution of the data. The mapping scale of 1:100,000 defines the overall level of geographic generalization and is typically used for regional planning. However, when planning field campaigns or applications requiring high spatial precision, a much finer resolution (for example, a grid size of 30 meters, as provided by some digital elevation models) will yield more accurate predictions. In the case of the autoRA algorithm, when it operates on data with a finer pixel resolution, it can capture environmental heterogeneity more precisely, resulting in better predictive performance. The improvement in prediction accuracy comes from the detailed information available at a finer resolution, not from an increase in field sampling cost, but rather from higher computational demands. In other words, while the mapping scale remains at 1:100,000 for broader interpretation, employing a finer grid in autoRA allows for a more detailed representation of soil variability, which is crucial for applications like planning a field campaign.

Thus, the algorithm's performance is not degraded by using a finer resolution; on the contrary, finer resolution generally improves accuracy. The trade-off is primarily related to computational costs and processing time rather than the quality of the predictions. This distinction is important because a map generated at a 1:100,000 scale might not be sufficiently detailed for certain field applications, whereas a digital soil mapping approach using autoRA with a fine-resolution grid (such as 30 meters) would offer the necessary spatial detail to support more precise decision-making in the field.

3. CHAPTER II

AUTORA: AN ALGORITHM TO AUTOMATICALLY DELINEATE REFERENCE AREAS. A CASE STUDY TO MAP SOIL CLASSES IN BAHIA - BRAZIL

3.1 RESUMO

Este estudo em Sátiro Dias, Bahia, avaliou a eficiência do algoritmo autoRA (em inglês *Automatic Reference Areas*) para delinear áreas de referência (RAs) em um fluxo de trabalho de mapeamento digital de solos (DSM). O autoRA integrou diversas covariáveis ambientais (por exemplo, geomorfologia, geologia, modelos digitais de elevação, temperatura, precipitação, etc.) usando o Índice de Dissimilaridade de Gower para capturar a variabilidade da paisagem de forma mais abrangente. Cento e dois perfis de solo foram coletados sob via RA delineada manualmente por um especialista para mapear classes de solo, sendo esse concebido como um mapa sem erro para comparação. Testamos tamanhos de área de cobertura pelo autoRA variando de 10% a 50%, comparando-as com o delineamento manual de RA e uma abordagem de mapeamento via DSM convencional utilizando "Área Total" (TA). A heterogeneidade ambiental foi insuficientemente amostrada em coberturas mais baixas (autoRA em 10–20%), resultando em baixa precisão de classificação (0,11–0,14). Em contraste, coberturas mais extensas melhoraram significativamente o desempenho: 30% produziram uma precisão de 0,85, enquanto 40% e 50% atingiram 0,96. Notavelmente, 40% atingiram o melhor equilíbrio entre alta precisão (Kappa = 0,65) e redundância mínima, superando o delineamento manual de RA (precisão = 0,75) e correspondendo aos melhores resultados de TA. Essas descobertas ressaltam a vantagem de aplicar uma estratégia automatizada e orientada para a diversidade, como o autoRA, antes das campanhas de campo, garantindo uma amostragem representativa de gradientes ambientais críticos para melhorar os fluxos de trabalho do DSM.

Palavras-chave: Mapeamento de Classes de Solos. Mapeamento Digital de Solos. Áreas previamente mapeadas.

3.2 ABSTRACT

This study in Sátiro Dias, Bahia, evaluates the efficiency of the autoRA (automatic Reference Areas) algorithm for delineating reference areas (RAs) in a digital soil mapping (DSM) workflow. autoRA integrates multiple environmental covariates (e.g., geomorphology, geology, digital elevation models, temperature, precipitation, etc.) using the Gower's Dissimilarity Index to capture landscape variability more comprehensively. One hundred and two soil profiles were collected under a specialist's manual delineation to establish baseline mapping soil taxonomy. We tested autoRA coverages ranging from 10% to 50%, comparing them to RA manual delineation and a conventional "Total Area" (TA) approach. Environmental heterogeneity was insufficiently sampled at lower coverages (autoRA at 10–20%), resulting in poor classification Accuracy (0.11–0.14). In contrast, more extensive coverages significantly improved performance: 30% yielded an Accuracy of 0.85, while 40% and 50% reached 0.96. Notably, 40% struck the best balance between high Accuracy (Kappa=0.65) and minimal redundancy, outperforming RA manual delineation (Accuracy=0.75) and closely matching the best TA outcomes. These findings underscore the advantage of applying an automated, diversity-driven strategy like autoRA before field campaigns, ensuring representative sampling of critical environmental gradients to improve DSM workflows.

Keywords: Soil Class Mapping. Digital Soil Mapping. Previously mapped areas.

3.3 INTRODUCTION

Soil sampling is essential for characterizing soil classes and properties in environmental, agronomic, and natural resource studies (BISWAS & ZHANG, 2018; BRUS, 2014; BRUS et al, 2011; CARTER & GREGORICH, 2007; KHOMUTININ et al., 2020). One common way to structure soil sampling is to define a Reference Area (RA), which are small subregions that capture the variability of soil-forming factors and represent larger areas (ARRUDA et al., 2016; LAGACHERIE et al., 2001; LAGACHERIE et al., 1995; MALLAVAN et al., 2010). Traditionally, experts manually delineate RAs using their knowledge and covariate maps such as land use, geomorphology, and climate to establish boundaries encompassing the variability of soil properties within the broader area of interest (FERREIRA et al., 2022; MALLAVAN et al, 2010; VOLTZ et al, 1997). While this approach leverages professional experience, it can suffer from limited reproducibility, objectivity, and scalability (ARRUDA et al., 2016; FERREIRA, et al., 2023) These challenges stem from the difficulty of applying an expert's mental model and synthesizing complex environmental information in highly variable areas, potentially compromising the RA methodology.

Despite the promise of RA methodologies in optimizing soil sampling, they face significant limitations due to the lack of automated techniques for identifying the most variable areas. Studies have demonstrated various challenges, such as those of Lagacherie and Voltz (2000) highlighted the difficulty in ensuring that small RAs accurately represent more significant regions using mathematical soilscape distances, which may not always be reliable. Arruda et al. (2016) Artificial neural networks with RAs achieved an accuracy of 82%, but the approach remains preliminary and may not capture the complexity of larger areas. Lagacherie et al. (1993), Lagacherie and Voltz (2000) and Yigini and Panagos (2014) encountered issues in accurately predicting soil properties and ensuring data transferability from RAs to broader regions. Additionally, Gonçalves et al. (2021) reported only moderate accuracy when extrapolating soil maps from small to large areas, indicating that current RA methods may struggle with scalability and representativity. While RA approaches provide a structured framework for soil mapping, the absence of automated techniques to identify and delineate the most variable areas limits their effectiveness and scalability in large, environmentally complex regions.

The Total Area (TA) approach, widely used in Digital Soil Mapping (DSM) workflows, treats the entire study region as a single dataset (BOETTINGER, 2010; DORNIK et al., 2022; GRUNWALD, 2010; HORST-HEINEN et al., 2021; KHALEDIAN & MILLER, 2020; SAURETTE; HECK et al., 2024). This method requires that soil sampling covers the whole area, which can be resource-intensive and time-consuming, especially for large and complex regions. The dataset is typically divided randomly into training and validation subsets, commonly using a 70/30 split. To validate the trained models, various statistical methods are employed, including cross-validation, which repeatedly trains and tests the model on different data subsets; k-fold cross-validation, which partitions the data into k folds and iteratively trains on k-1 folds while validating on the remaining fold; and out-of-the-bag validation, often used with ensemble methods like random forests, which utilizes bootstrap samples for training and the unused samples for validation. These validation techniques are applied across multiple iterations, and the results are aggregated to assess model reproducibility and classification performance (CARVALHO et al., 2020).

In pursuit of more systematic methods, researchers in Pedometrics and DSM have developed advanced sampling schemes that leverage modern computational and sensor technologies (CASA et al., 2013; GRUNWALD; VASQUES; RIVERO, 2015; JI et al., 2019; RODRIGUES, HUGO et al., 2024). Conditioned Latin Hypercube Sampling (cLHS), for example, statistically balances soil covariates by partitioning the covariate space to ensure all

quartiles of key soil-forming factors are represented (MALONE et al., 2019; MINASNY & MCBRATNEY, 2006). However, cLHS does not rely on the RA hypothesis and instead depends on the TA approach, requiring comprehensive sampling across the entire study region. To enhance cLHS's practicality, Malone et al. (MALONE et al., 2019) addressed challenges such as optimizing sample size, relocating inaccessible sites, and prioritizing under-sampled areas by providing customized R scripts, making cLHS more adaptable to real-world field conditions. Additionally, Saurette et al. (2023) introduced divergence metrics like Kullback-Leibler and Jensen-Shannon divergences to determine optimal training sample sizes, improving the methodological framework for DSM by ensuring more accurate and efficient sample designs.

Despite these advancements, both cLHS and the methodologies proposed by Saurette et al. remain grounded in the TA approach, as they do not incorporate the RA framework. In this study, we hypothesize that an automated method for delineating RA significantly reduces subjectivity and makes the approach more effective (better accuracy of prediction models and lower cost for soil mapping). This automation enhances reproducibility and scalability, facilitating the broader adoption of RA-based methods within the DSM community.

We introduce autoRA (automatic Reference Area), a novel algorithm that locates and delineates RA based on covariate maps commonly applied in DSM workflows using Gower's Dissimilarity Index. Costa et al. (2024) utilized this index to identify the most heterogeneous regions in Rio de Janeiro, highlighting that the covariate values were notably high, indicating significant regional differences. This high variability suggests that sampling soil from these heterogeneous areas can yield a diverse dataset, which is crucial for effective digital soil modeling. Their approach underscores the importance of selecting representative regions to enhance the accuracy and reliability of predictive models in soil science.

AutoRA allows the user to generate a new RA for an area without previous mapping (legacy data) and can also be used to evaluate whether a given RA, previously delimited conventionally (manually), could be better delineated (size and shape).

The objectives of this study were to use autoRA to: (i) assess how well soil-class maps are generated via autoRA compared to those produced from a manually delineated RA; and (ii) to evaluate both RA-based methods (manually and by autoRA) against the TA approach in terms of sampling efficiency, reproducibility, and classification accuracy.

3.4 MATERIAL AND METHODS

3.4.1 The autoRA algorithm

The patented autoRA approach (BR1020240208676; trademark 937505684) for automatic delineation of RAs in soil mapping, which is already registered in Brazil and will soon be registered at the United States Patent Office, is shown in Figure 18. It illustrates the iterative flowchart whereby autoRA refines RA parameters by simultaneously assessing resolution and the proportion of the target area. It integrates a random forest-based predictive simulated theoretical surface (STS) derived from whole-area sampled points (including covariates used for Gower distance calculations). It compares that total-area approach against an STS model fitted solely with the sampled points within the RA.

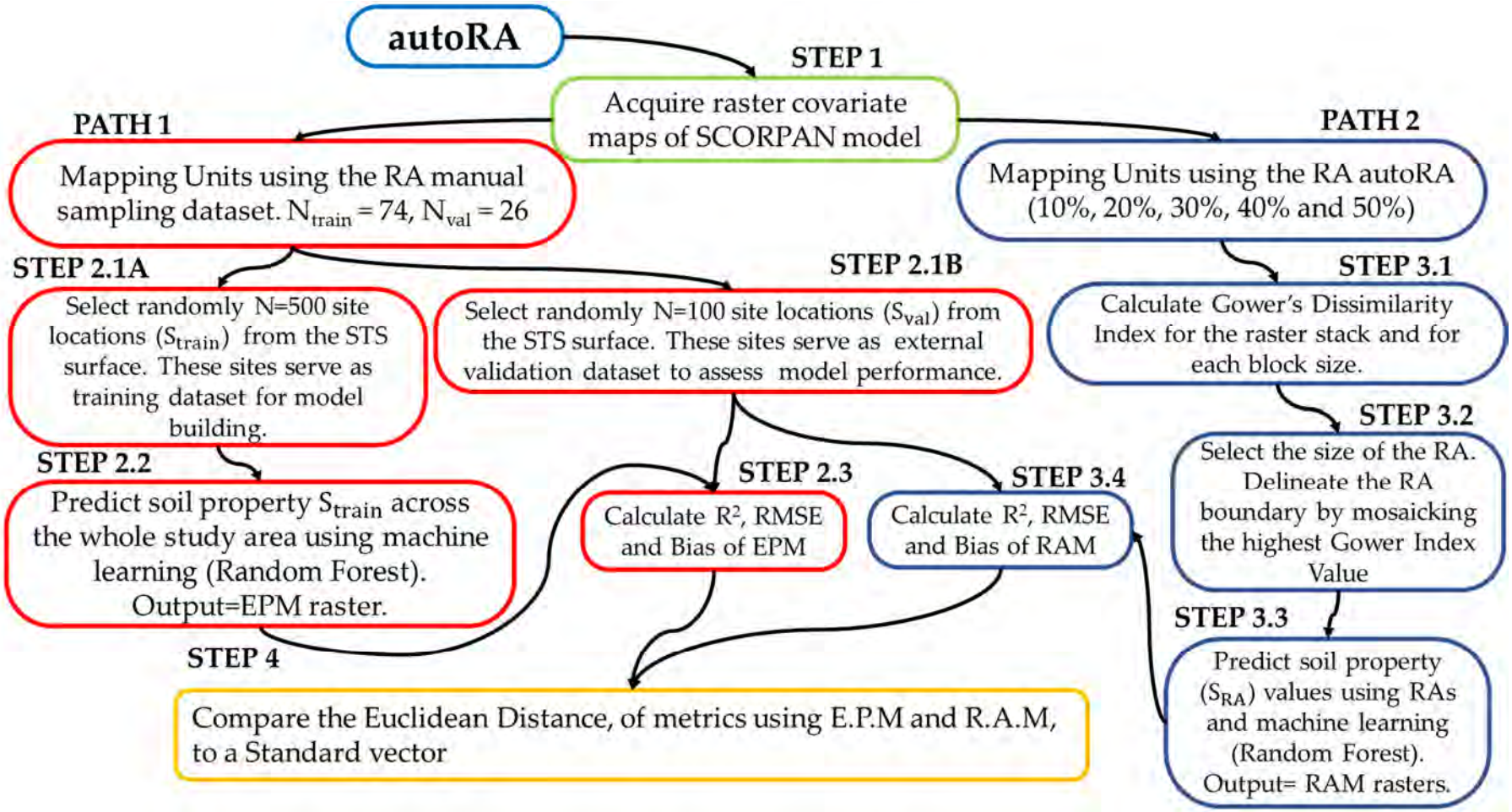


Figure 18. The general workflow of the autoRA algorithm.

In STEP 1, the user assembles the geospatial datasets $\{X_1, X_2, \ldots, X_p\}$ representing SCORPAN factors (Soil, Climate, Organisms, Relief, Parent material, Age, and spatial Neighborhood) (MCBRATNEY; MENDONÇA SANTOS; MINASNY, 2003). Each X_j can be numeric (continuous/discrete) or categorical (nominal/ordinal). The native resolution of these covariates constrains the minimal block size that can later be used in RA delineation.

The workflow proceeds along Path 1 and Path 2 in parallel. Path 1 begins with STEP 2.1A, where randomly sample n = 500 training points $\{(x_i, y_i, s_i)\}$ from an "exhaustive" simulated soil surface (S_{exh}, or STS) representing a hypothetical ground truth of the soil property of interest S_{exh}. Every pixel in the AOI has a known value s_i and used to build a Random Forest model (LIAW; WIENER, 2002) to predict S_{exh} from the covariates f_{RF}: $(X_1, \ldots, X_p) \mapsto \hat{s}_{exh}$.

In STEP 2.1B, an independent validation set $\{(x_i, y_i, s_i)\}_{j=i}^{m}$ is also randomly sampled over the entire area of interest to assess the model accuracy via the performance metrics R^2 (Equation 11), RMSE (Equation 12), and Bias (Equation 13) for observed s_j vs. predicted $\hat{s}_j$.

$$\mathbf{R^2} = \mathbf{1} - \frac{\sum_{j=1}^{m}(\hat{s}_j - s_j)^2}{\sum_{j=1}^{m}(\bar{s}_j - s_j)^2} \qquad \text{Eq. (11)}$$

Where: $\bar{s}_j = \frac{1}{m}\sum_{j=1}^{m} s_j$

$$\mathbf{RMSE} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\hat{s}_j - s_j)^2} \qquad \text{Eq. (12)}$$

$$\mathbf{Bias} = \frac{1}{m}\sum_{j=1}^{m}(\hat{s}_j - s_j)^2 \qquad \text{Eq. (13)}$$

STEP 2.2 applies f_{RF} to the entire AOI's covariate stacks to produce the "exhaustive" predictive surface S_{EPM} and STEP 2.3 serves as an ideal reference for subsequent performance comparisonsS_{EPM} finalize this "exhaustive" predictive model (EPM) as a ground-truth benchmark.

Path 2 represents the part of the algorithm that effectively creates the RAs by cropping the inputted raster maps (RA by autoRA - Figure 18). It performs the autoRA sensitivity analysis to delineate RAs using two adjustable parameters—block size (5 to 150 pixels) and RA target area (10 to 50% of the AOI, in 10% increments).

In STEP 3.1, Gower's Dissimilarity Index is computed per block for the SCORPAN covariates. For continuous Covariates, a pair of blocks i, j, the partial distance for a continuous covariate X_k is shown in Equation 14.

$$\mathbf{d_k(i,j)} = \frac{|\mathbf{X_k(i)} - \mathbf{X_k(j)}|}{\mathbf{\max X_k(i)} - \mathbf{\min X_k(i)}} \qquad \text{Eq. (14)}$$

When dealing with categorical covariates, they are transformed to be dummy-based. For a nominal X_ℓ,

$$d_{\ell(i,j)} = \begin{cases} 0, \text{if } X_\ell(i) = X_\ell(j) \\ 1, \text{otherwise} \end{cases}$$

When dealing with ordinal variables, a rank is transformed into [0, 1] and treated as continuous. Finally, the combined Gower's Dissimilarity for p covariates are given by Equation 15.

$$\mathbf{d_G(i,j)} = \frac{\sum_{k=1}^{p} \boldsymbol{\delta_k(i,j)}\ \mathbf{d_k(i,j)}}{\sum_{k=1}^{p} \boldsymbol{\delta_k(i,j)}} \qquad \text{Eq. (15)}$$

Where $\delta_k(i,j) = 1$ if both $X_k(i)$ and $X_k(j)$ are non-missing; otherwise $\delta_k(i,j) = 0$.

STEP 3.2 mosaics the highest dissimilarity values above the mean, defining RA boundaries and generating multiple RAs of varying sizes. The Block Size parameter variation aggregates blocks of pixels in sizes from 5 to 150 pixels. The pixels with high dissimilarity are mosaicked, whose mean d_G concerning the AOI average exceeds $\bar{d}_G$. These are merged to form candidate RAs. The final aggregation of high dissimilarity pixels constrains the resulting RA to 10–50% of the total AOI (in increments of 10%) by adjusting thresholds on d_G.

STEP 3.3 then samples a fixed number of points sampling within each RA draw a fixed number of training points $\{(x_r, y_r)\}$ to produce a new model fit training a Reference Area Model (RAM) $\hat{S}_{RA}$ (using RF or another learner) for predicting soil properties across the entire AOI and using just the RA sample for training. The RAM validation process uses an independent validation set to evaluate each RA's predictive surface with the same metrics (R2, RMSE, Bias).

Finally, STEP 4 compares the RAM predictions from Path 2 to the EPM (from Path 1) using R^2, RMSE, and Bias. Euclidean Distance (ED). For each RA, define its performance vector $(R^2_{RA}, RMSE_{RA}, Bias_{RA})$ using Equation 16.

$$\mathbf{ED(RA)} = \sqrt{\mathbf{(R^2_{RA} - 1)^2 + RMSE_{RA}{}^2 + Bias_{RA}{}^2}} \qquad \text{Eq. (16)}$$

The RA minimizing ED(RA) is deemed "best-performing." The EPM from Path 1 serves as a benchmark to gauge how well each RA (Path 2) extrapolates to the entire AOI.

By coupling the STS benchmark with RA-based modeling, autoRA thus provides a robust way to test different RA dimensions and ensure accurate soil property predictions.

3.4.2 Theorem: heterogeneous coverage and extrapolation

Let Ω be the set of all possible spatial units (blocks or pixels) in an Area of Interest (AOI). Each spatial unit $x \in \Omega$ is characterized by a feature vector $X_{(x)} \in R^p$, where p is the number of covariates, including both continuous and categorical covariates. Define a dissimilarity function d_G (GOWER, 1971) over Ω, and let $D(\Omega)$ be the maximum dissimilarity range found in the AOI, which is defined as the largest pairwise dissimilarity between any two spatial units in Ω, as given by Equation 17.

$$\mathcal{D}(\Omega) = \max_{x,y \in \Omega} d_G\big(X(x), X(y)\big) \quad \text{Eq. (17)}$$

Equation 17 defines $\mathcal{D}(\Omega)$ as the maximum dissimilarity, meaning the greatest observed distance between any two feature vectors in the AOI according to the Gower dissimilarity function. We aim to find a subset $\Omega^* \subseteq \Omega$, referred to as the Reference Area (RA), which "covers" a large portion of the AOI's heterogeneity. Formally, we want:

$$\forall\, x \in \Omega,\ \exists r \in \Omega^* \quad \text{such that} \quad d_G\big(X(x), X(r)\big) \leq \delta,$$

For some small $\delta > 0$.

3.4.3 autoRA's theorem

Suppose the AOI satisfies a Lipschitz-like condition (GAULD, 1974) for a soil property S, meaning there exists $L > 0$ such that (ADAMS; FOURNIER, 2003):

$$|S(x) - S(y)| \leq L \cdot d_G\big(X(x), X(y)\big), \forall\, x, y \in \Omega.$$

In this way, small changes in the covariates lead to proportionally small changes in S. If Ω^* is an RA for which:

$$\max_{x,y \in \Omega} \min_{r \in \Omega^*} d_G\big(X(x), X(r)\big) \leq \delta,$$

Then a predictive model f trained on Ω^* can extrapolate to Ω with maximum error bounded by $L\,\delta$, meaning that for every $x \in \Omega$, there exists some $y \in \Omega$ such that (FERRY; WEINBERGER, 2013):

$$\max_{x \in \Omega} |S(x) - f(x)| \leq L\,\delta.$$

This ensures that the error bound accounts for distances between all possible pairs of points in Ω, addressing the concern about the absence of y in the final bound.

A RA Ω^* that captures the most heterogeneous pixels in the covariate space ensures good predictive coverage (FALCONER; MARSH, 1992). Under mild assumptions, if the RA encloses the full range of environmental variability, a soil model (e.g., Random Forest) trained on Ω^* can extrapolate to the remainder of the AOI with limited error (bounded by $L\,\delta$). This underpins the autoRA rationale: locate a small portion of the area rich in heterogeneity so that any unvisited point in the AOI remains "close" to some training point in Ω^*.

3.4.4 Overview of the research workflow

study was conducted in Sátiro Dias, Bahia, Brazil, covering an area of approximately 901 km². The methodology comprised three main stages: 1) delineation of Reference Areas (RAs), 2) soil sampling simulation, and 3) soil classification modeling (Figure 1). Two approaches were used to delineate RAs: first, experts manually delineated an RA covering 212 km² (RA manual), and second, the autoRA algorithm was employed to identify areas of maximum dissimilarity via Gower's Dissimilarity Index (RA autoRA). By applying thresholds of 10%, 20%, 30%, 40%, and 50% of the highest dissimilarity values, multiple RA autoRA subsets were produced, each reflecting different levels of landscape variability.

In addition to these RA-based methods, a Total Area (TA) approach was implemented, treating the entire 901 km² as a single dataset, randomly split into 70% for training and 30% for validation, repeated 100 times to minimize bias (ELLILI et al., 2019; KESKIN; GRUNWALD; HARRIS, 2019; PADARIAN; MINASNY; MCBRATNEY, 2019).

For the RA manual, 74 samples inside the boundary were used to train the soil classification model, and 28 samples outside were used for external validation (102 total). For each RA autoRA, boundaries were intersected with the aggregate (training + validation) soil sample dataset, defining samples within each RA as training points and those outside as external validation.

The final soil classification stage involved creating three models: (1) RA manual-based models using the manually delineated RA samples, (2) RA autoRA-based models built from the various threshold-defined RA autoRA subsets, and (3) TA-based models using the entire dataset under the repeated 70/30 splits. Model performance was evaluated using overall accuracy and the kappa index, comparing manual and automated RA delineation alongside the aggregated TA approach, whose results were averaged over 100 iterations. An overview of the workflow applied in the present study is shown in Figure 19.

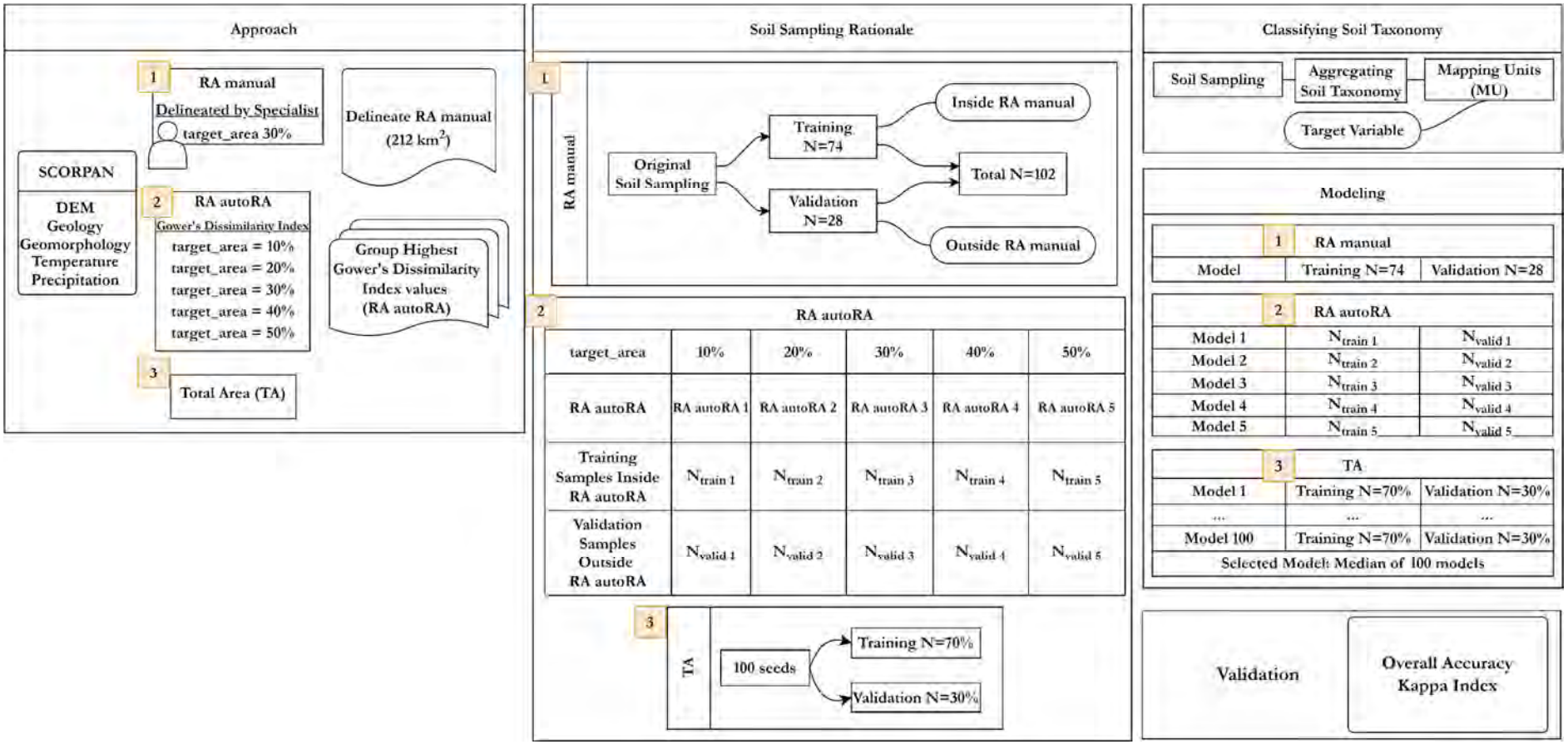


Figure 19. Flowchart of the methodology implemented in the research compares the RA manual, RA autoRA, and the Total Area.

3.4.5 Study area, data preparation, and manual reference area

The study was conducted in Sátiro Dias, Bahia, Brazil, covering a region of interest (RI) of 901 km² (Figure 20). Before initiating fieldwork, the specialist compiled spatial covariates influencing soil formation in the area. These covariates included geology, geomorphology, and pedology maps from the Brazilian Institute of Geography and Statistics (IBGE) at a 1:250,000 scale (IBGE, 2018b).

The specialist considered the climatic data from WorldClim Version 2 to enhance the environmental covariates for better comprehension. It provides high-resolution gridded climate data with monthly averages from 1970 to 2000, available at spatial resolutions ranging from 30 arc-seconds (~1 km²) to 10 arc minutes (FICK; HIJMANS, 2017) including average annual precipitation and temperature maps.

Additionally, the digital elevation model (DEM) from the Shuttle Radar Topography Mission provided by NASA was included (FARR et al., 2007) with a spatial resolution of 1 arc-second (~30 meters at the Equator). These data were critical for understanding terrain-related factors influencing soil formation and representing the SCORPAN model (MCBRATNEY, A. B; MENDONÇA SANTOS; MINASNY, 2003).

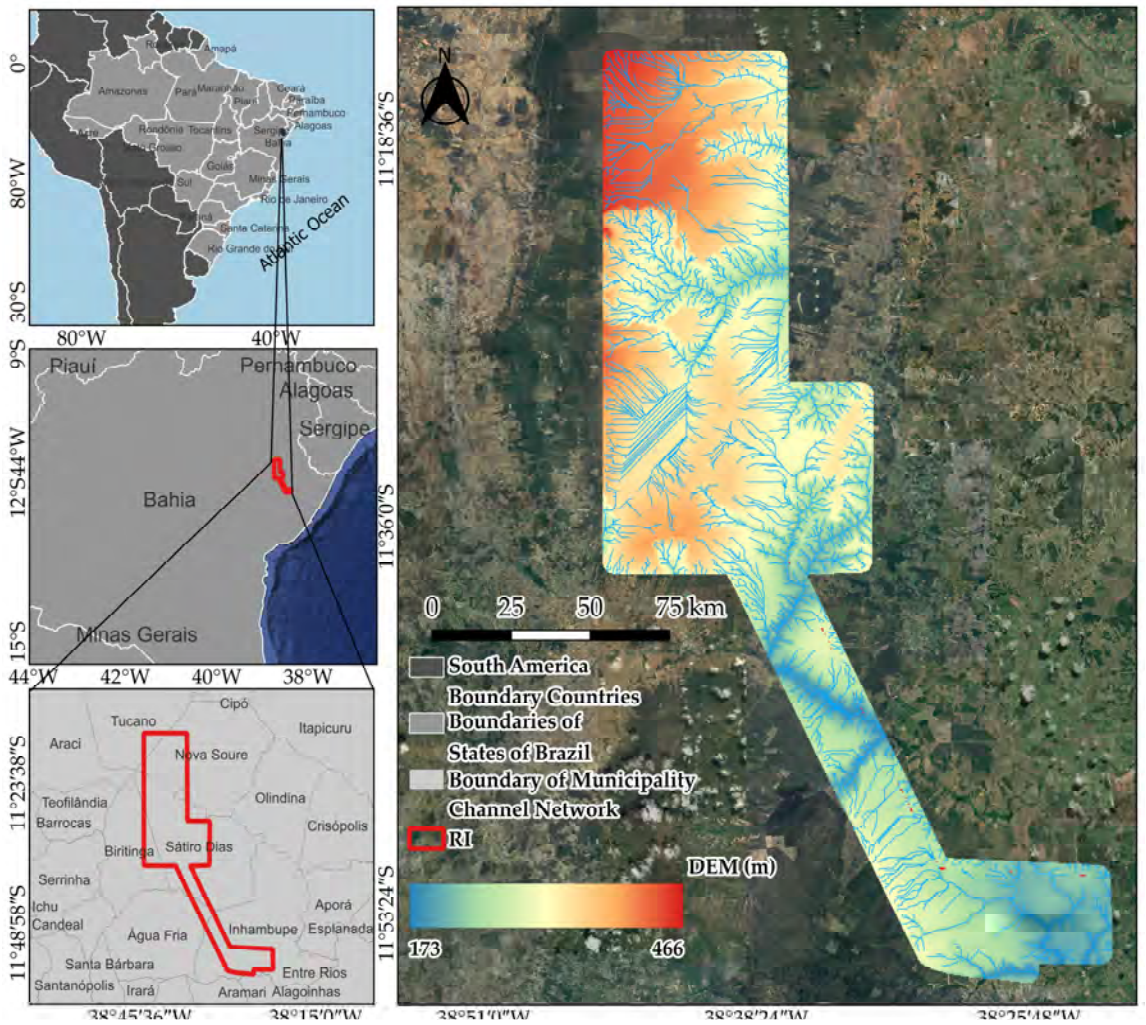


Figure 20. Location of the study area of Sátiro Dias with the RA manual. A DEM in the background to aid in understanding the physiography of the landscape.

A soil science specialist manually delineated the RA (RA manual) by visually interpreting the combined spatial patterns of the covariates (ARRUDA et al., 2016; LAGACHERIE et al., 2001; LAGACHERIE; LEGROS; BURROUGH, 1995; VOLTZ; LAGACHERIE; LOUCHART, 1997). The specialist identified regions within the region of interest (RI) that exhibited heterogeneous soil-forming conditions, ensuring that the RA manual captured the most representative combinations of geology, geomorphology, climate, and pedology. Sequentially overlaying raster layers to identify intersections of soil-forming factors considering the mental model created and refining the RA manual final shape by detecting spatial patterns and transitions indicative of distinct soil-forming environments (CARVALHO et al., 2020).

Using this integrated analysis, the specialist manually delineated the boundaries of the RA manual to capture the most representative combinations of soil-forming factors by considering the smallest regions with significant diversity. The delineation was constrained to cover no more than 30% of the RI, resulting in an RA manual spanning 212 km² (Figure 20). This approach is intended to ensure that the RA manual includes diverse environmental conditions while maintaining a manageable area for intensive soil sampling (FAVROT, 1986; FAVROT, 1981; BORNAND; FAVROT, 1998).

Following the RA manual, the specialist applied the conditioned Latin hypercube algorithm (ROUDIER, 2011) to allocate 74 sampling points within the RA manual using the environmental covariates listed (Figure 21). These points were used for detailed observation and description of complete soil profiles, forming the primary dataset for model training. Additionally, 28 points were randomly distributed outside the RA manual to serve as an external validation dataset, ensuring the accuracy of models and maps constructed during the study (Figure 21).

Figure 21 highlights how elevation, geology, and climate collectively shape soil distribution. Lower altitudes (<232 m) with higher temperatures (>23.5 °C) and lower precipitation (≤718 mm) predominantly host Orthic Entisols (RQo) and Lithic Entisols (RLd). By contrast, higher elevations (>408 m) from the Barreira Formation (geologic type), experiencing cooler temperatures (≤22.74 °C) and more significant rainfall (>1032 mm), promote Typic Udults (LVAd). Intermediate altitudes (290–408 m) support Typic Udults (LVAd) and Eutrophic Udults (PVAe) in convex and tabular reliefs, reflecting moderate climatic conditions. Overall, the Barreiras formation fosters more weathered soils, while the Marizal formation in degraded pediplanes favors less developed soils.

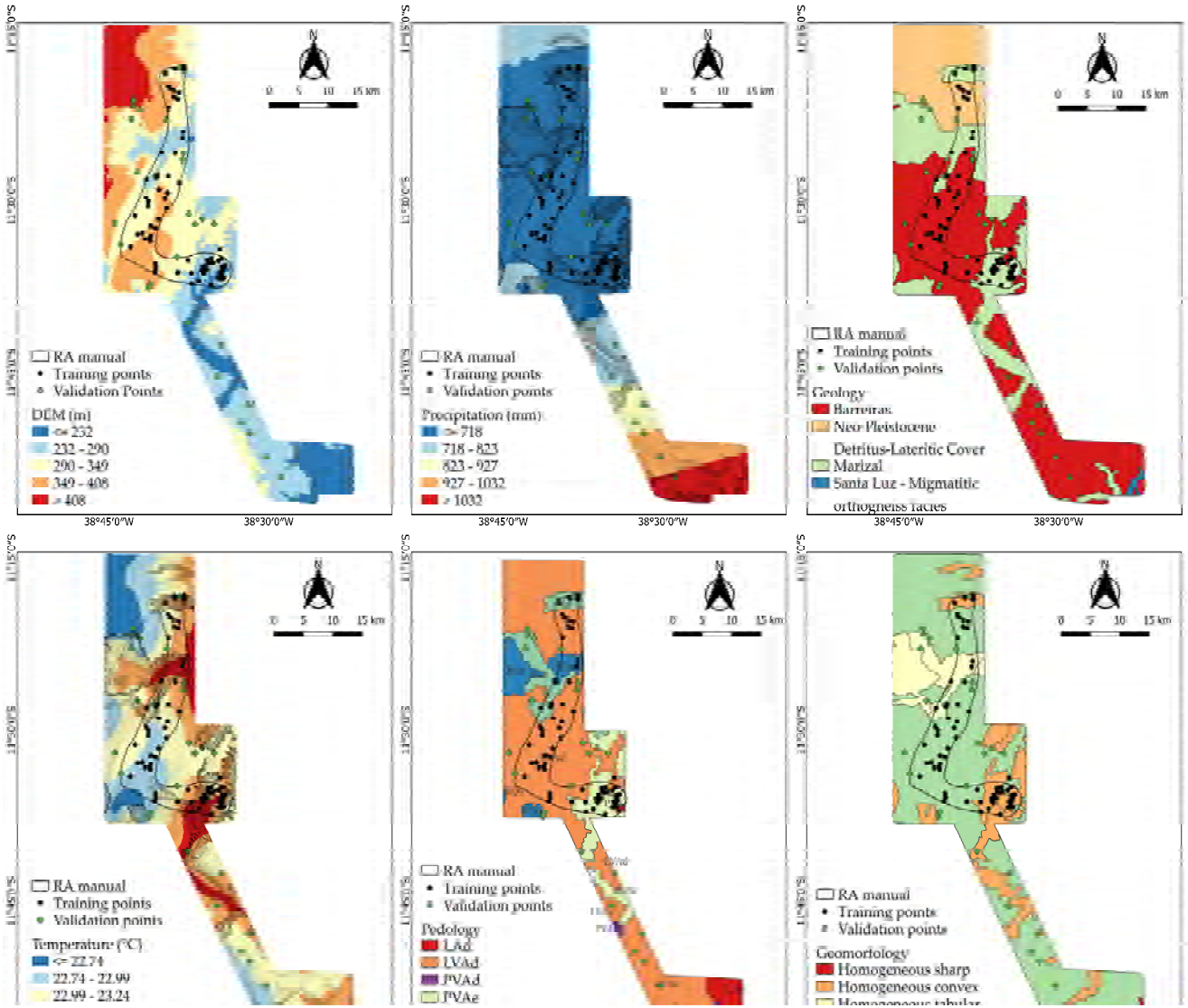


Figure 21. Maps of the covariates used to define the manual RA and automatic RA using the autoRA algorithm.

3.4.6 Characterization of mapping units

Table 1 presents the mapping units (MU) generated for the area of Sátiro Dias using the reference area manually delineated and their respective landscape characteristics. After correlating the soil types with the landscape attributes, 5 MUs were discretized.

MU1 - Plateau, flat to gently undulating relief: Soil complex composed of the main classes Quartzipsamment (RQo) and Dystrudept (CXvd), with sandy-loam texture, moderate A horizon, including Haplustox (LAd).

MU2 - Upper and middle third of the plateau slopes: Soil complex composed of the main classes Petroplinthic Paleudult (FXd) and Petroplinthic Haplustox (LAd), with sandy-loam texture, moderate A horizon, phase with epipedons, including Dystric Haplustox (LVd).

MU3 - Region of hills at lower altitudes than plateaus (Sátiro Dias central region): Simple unit dominated by the class Typic Eutrudept (CXve), with medium-clay texture and moderate A horizon, including occurrences of Argic Eutrudept (CXve), Paleudalf (LV), and Natric Haplotert (CXv) with moderate A horizons.

MU4 - Region of hills within canyons: Soil complex composed of the classes Haplustox (LAd), Dystric Haplustox (LVAd), Petroplinthic Paleudult (PAdx), Typic Eutrudept (CXve), and Quartzipsamment (RQo).

MU5 - Lowlands within canyons (North of Sátiro Dias): Simple unit occurrence of Quartzipsamment (RQo) with sandy texture and weak A horizon.

Figure 22 shows the map of soil classes (MU) in the region of Sátiro Dias. The MU1 (RQo + Lad + CXvd complex) is the one with the highest territorial expression (71%) since areas of sandy plateaus predominate. The MU with the second most significant territorial expression is MU3 (single unit CXve, 11%). This mapping unit is associated with hills at an altitude lower than the plateaus in the central region of Sátiro Dias. In this mapping unit, soils with clay of high activity and high base saturation (Ca^{++}, Mg^{++}, K^{+}, and Na^{+}) stand out.

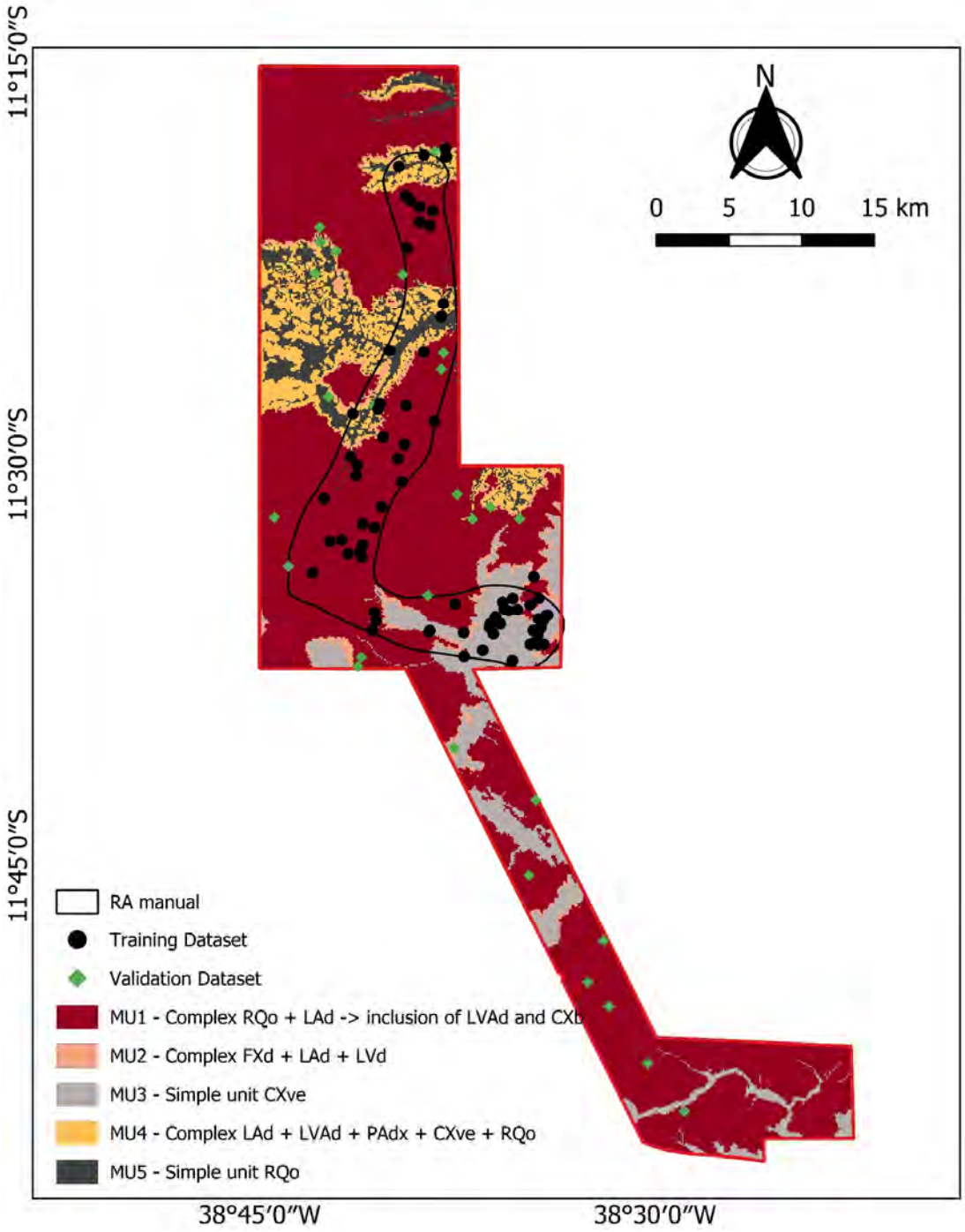


Figure 22. A map of soil classes in Sátiro Dias developed from soil samples using the RA manual approach will serve as the benchmark for comparison with the autoRA and the TA approaches.

Table 1. Soil mapping units in Sátiro Dias.

MU	Description	Environment	Area (km²)	%	Brazilian Soil Classification System	USDA Soil Taxonomy Correspondence
MU1	Complex RQo + LAd + CXbd.	Plateau, flat to gently undulating relief	636,92	71	NEOSSOLO QUARTZARÊNICO Órtico típico, LATOSSOLO AMARELO Distrófico textura média, CAMBISSOLO HÁPLICO Tb Distrófico, textura média-arenosa, A moderado	RQo: Entisols (Typic Quartzipsamments); LAd: Oxisols (Typic Hapludox); CXbd: Inceptisols (Typic Dystrudepts)
MU2	Complex FXd + LAd + LVd.	Upper and middle slopes of plateaus	28,49	3	PLINTOSSOLO HÁPLICO Distrófico petroplíntico, LATOSSOLO AMARELO Distrófico petroplíntico, A moderado, fase epipedregoso, Inclusão de LATOSSOLO VERMELHO Distrófico textura média	FXd: Inceptisols (Aquic Dystrudepts); LAd: Oxisols (Petroplinthic Haplustox); LVd: Oxisols (Typic Hapludox)
MU3	Simple unit CXve.	Hills with lower elevation than plateaus (Center of Sátiro Dias)	95,45	11	CAMBISSOLO HÁPLICO Ta Eutrófico típico, textura média-argilosa, Inclusões: LUVISSOLO HÁPLICO Pálico típico, VERTISSOLO HÁPLICO Sódico	CXve: Inceptisols (Typic Eutrudepts); Luvisolo: Alfisols (Typic Haplustalfs); Vertissolo: Vertisols (Typic Haplusterts)
MU4	Complex LAd + LVAd + PAdx + CXve + RQo.	Hill regions within canyons	78,04	8	LATOSSOLO AMARELO Distrófico textura média, LATOSSOLO VERMELHO-AMARELO Distrófico textura média, ARGISSOLO AMARELO Distrófico petroplíntico, CAMBISSOLO HÁPLICO Ta Eutrófico típico, textura média-argilosa, NEOSSOLO QUARTZARÊNICO Órtico típico	LAd: Oxisols (Typic Haplustox); LVAd: Oxisols (Typic Kandiudox); PAdx: Alfisols (Plinthic Kandiustalfs); CXve: Inceptisols (Typic Eutrudepts); RQo: Entisols (Typic Quartzipsamments)
MU5	Simple unit RQo.	Lowlands within canyons (North of Sátiro Dias)	61,82	7	NEOSSOLO QUARTZARÊNICO Órtico típico, textura muito arenosa, A fraco	RQo: Entisols (Typic Quartzipsamments)
TOTAL	-	-	900,72	100	-	-

3.4.7 Soil sampling regrouping

The original soil sampling dataset comprised 102 points, with 74 designated for training and 28 for external validation, collected both within and outside the RA manual, respectively, and was used to simulate the dataset for the RA autoRA. To assess the efficacy of RA autoRA delineations, the sampling points were reassigned based on their spatial relation to each RA autoRA subset. Points intersected within the RA autoRA boundaries for each threshold (10%, 20%, etc.) were designated as part of the training set for that specific RA autoRA scenario. In contrast, points located outside the RA autoRA boundaries were reserved as the external validation set.

3.4.8 Spatial prediction using the reference area and the total area dataset

The mapping procedure for the MU was performed for the RA manual by training a Random Forest (RF) classification algorithm using the package of the same name (LIAW; WIENER, 2002) present in the R software (R CORE TEAM, 2024) using the training dataset with 74 points. For the RA autoRA mapping procedure, the training dataset varied in terms of the number of profiles in a way that this number was a direct reflection of the intersection between the boundaries of the delineated RA autoRA and the whole soil profile dataset reclassified with inner soil samples in each RA autoRA target area size reclassified as training dataset while the remaining reserved as external validation dataset. For the TA approach, the 102 soil samples were split into training and validation using the 70% and 30% ratio, respectively. To reduce the randomization effect during the sampling process, it was repeated 100 times, and, consequently, 100 RF were adjusted.

The RF's fit parameters of the model were maintained by the default defined by the package's authors, in which the number of trees was 500. The minimum amount of data in each terminal node parameter has been set to the default of five for each terminal node. Regarding the number of variables used in each tree, for classification problems, the default value is one-third of the total predictor variables.

3.4.9 Accuracy of the mapping unit maps

The following quality measures were used to assess the quality of the predicted MU maps: Overall accuracy (OA), Kappa coefficient of agreement, User's Accuracy (UA), and Producer's Accuracy (PA). All of them were based on the confusion matrix (BRUS; KEMPEN; HEUVELINK, 2011) and are calculated as the proportion of the samples or soil types correctly predicted over the total number of validation locations (reference field data). The OA was given by Equation 18 (BRUS; KEMPEN; HEUVELINK, 2011).

$$\mathbf{OA} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{c}} \mathbf{E_{ij}}}{\mathbf{n}} \qquad \text{Eq. (18)}$$

In which E is the confusion or error matrix of dimensions c x c; and n is the number of samples (observations). In literature, overall accuracy is also called overall purity, map purity, global accuracy, and general accuracy (BRUS; KEMPEN; HEUVELINK, 2011). UA is given by Equation 19.

$$\mathbf{UA} = \frac{\mathbf{E_u}}{\mathbf{E_{iu}}} \qquad \text{Eq. (19)}$$

Which E_{iu} denotes the number of points mapped as the mapping unit u, that is, the sum of the rows in the confusion matrix, and E_u are the classes correctly classified in unit u, the main diagonal of the confusion matrix. The complement of UA (1- UA) is referred to as the error of commission (inclusion), that is, the error ruled by including pixels from other classes in the class in question. In the literature, other synonyms are also used for User's Accuracy, such as map unit purity (BRUS; KEMPEN; HEUVELINK, 2011), which is about predicted classes (map) (COSTA et al., 2021). The PA is given by Equation 20.

$$\mathbf{PA} = \frac{\mathbf{E_u}}{\mathbf{E_{ju}}} \qquad \text{Eq. (20)}$$

In which E_{ju} denotes the number of points mapped as the mapping unit u, that is, the sum of the columns in the confusion matrix, and E_u are the classes correctly classified in unit u, the main diagonal of the confusion matrix. The complement of PA (1- PA) is referred to as the omission errors (exclusion) when a pixel ceases to be classified correctly in that mapping unit and is incorrectly classified as another unit. In the literature, other synonyms are also used for Producer's Accuracy, such as class representation on terrestrial truth (reference field data) (COSTA et al., 2021). The Kappa Index is given by Equation 21.

$$\hat{\mathbf{k}} = \frac{\mathbf{n}\sum_{\mathbf{i=1}}^{\mathbf{c}}\mathbf{E_{ij}} - \sum_{\mathbf{i=1}}^{\mathbf{c}}\mathbf{E_i E_j}}{\mathbf{n^2} - \sum_{\mathbf{i=1}}^{\mathbf{c}}\mathbf{E_i E_j}} \qquad \text{Eq. (21)}$$

Where c is the number of classes on the matrix, E_{ij} values on the row i and column i, E_i total on the row i and E_i total on columnj, and n the number of samples (observations).

Finally, as different soil maps of the same study site using RA manual, RA autoRA, and TA approaches were compared, the indexes WPAI (Weighted Producer Accuracy Index) and WUAI (Weighted User Accuracy Index) were also computed (Equations 22 and 23, respectively).

Generating the WUAI and WPAI indices aims to give a global view of each map's user and producer accuracy. Thus, as in each map, the mapping units have different territorial expressions. Both indexes are weighted averages of user and producer accuracy. The weighting is done by multiplying this OA by the area of each mapping unit (MU) divided by the total area of the map (A).

The WUAI and WPAI indices allow us to know how the types of errors are distributed (commission or omission, respectively) in each map (give an overview of these errors for a specific map). Thus, after comparing the OA and the kappa index, before entering into the detailed evaluation of the types of errors per mapping unit (which is conventionally done), we used the WUAI and WPAI indices to compare the relevance of commission and omission errors on each map (5 generated by RA autoRA, 1 by TA and 1 by RA manual generated by RF). The index values range from 0 to 1, with 0 lacking OA and 1 maximum OA.

$$\mathbf{WPAI} = \frac{\sum_{\mathbf{j}}^{\boldsymbol{n}} \frac{\mathbf{E_u}}{\mathbf{E_{ju}}} * \boldsymbol{a_u}}{\mathbf{A}} \qquad \text{Eq. (22)}$$

In which E_{ju} denotes the number of points mapped as the mapping unit u that is, the sum of the columns in the confusion matrix and E_u are the classes correctly classified in that

unit u, the main diagonal of the confusion matrix, a_u is the surface area of the mapped unit u, and A is the total surface area of the map.

$$\mathbf{WUAI} = \frac{\sum_i^n \frac{\mathbf{E_u}}{\mathbf{E_{iu}}} * \boldsymbol{a_u}}{\mathbf{A}} \qquad \text{Eq. (23)}$$

In which E_{iu} denotes the number of points mapped as the mapping unit u that is, the sum of the rows in the confusion matrix and E_u are the classes correctly classified in that unit u, the main diagonal of the confusion matrix, a_u is the surface area of the mapped unit u, and A is the total surface area of the map.

3.5. RESULTS AND DISCUSSION

3.5.1 Soil landscape relationship and spatial distribution

The benchmark for this study is the Mapping Unit (MU) map created during the original field campaign project, which serves as a reference for detailing the manually delineated Reference Areas (RA manual) and associated soil profiles. To provide an in-depth understanding, the study area has been divided into seven sections (Sections 1 through 7), each highlighting key soil profiles and their characteristics. These sections include detailed photos and descriptions of representative profiles to illustrate the spatial variability and transitions within the region.

The first subregion (Section 1), as illustrated in Figure 23, covers the northern part of the RI, encompassing profiles PE006, PE007, and PE1008. Profile PE006, situated in an elevated and well-drained area, has sandy textures, whereas PE1008, located in flatter terrain, shows organic matter accumulation in the upper layers due to reduced drainage (CHENG et al., 2011). These differences align with the variability in geological and pedological covariates, emphasizing the critical influence of parent material and topography on soil properties in the region.

Additionally, profiles PE0006, PE0007, and PE1008 are representative soil types of transition characteristics between canyon and plateau regions in Nova Soure (Neighbor Municipality of Sátiro Dias). These profiles exhibit a medium to high degree of development on gently to strongly undulating slopes, occupying the middle and upper thirds of the landscape. A notable feature in this geological context (Marizal formation) is the frequent presence of surface gravel, ranging from slightly gravel to extremely gravel. In the subsurface, the soils display yellowish, reddish-yellow, and/or yellowish-red hues, resembling plateau soils but differing in the absence of surface gravel and exhibiting deeper, more developed profiles.

PE0006 may show mottling in the C horizons, linked to the parent material, reflecting the soil-rock interface's transitional nature (CLOTHIER; POLLOK; SCOTTER, 1978; OWENS; RUTLEDGE, 2005). In the lower part of this region, extremely sandy, deep, and sharply drained soils such as PE1008 are found (Figure 23). Despite its darker color, the soil is very sandy and has relatively homogeneous horizons between them. The horizons of these soils have a weak aggregation and predominantly loose dry and wet consistency.

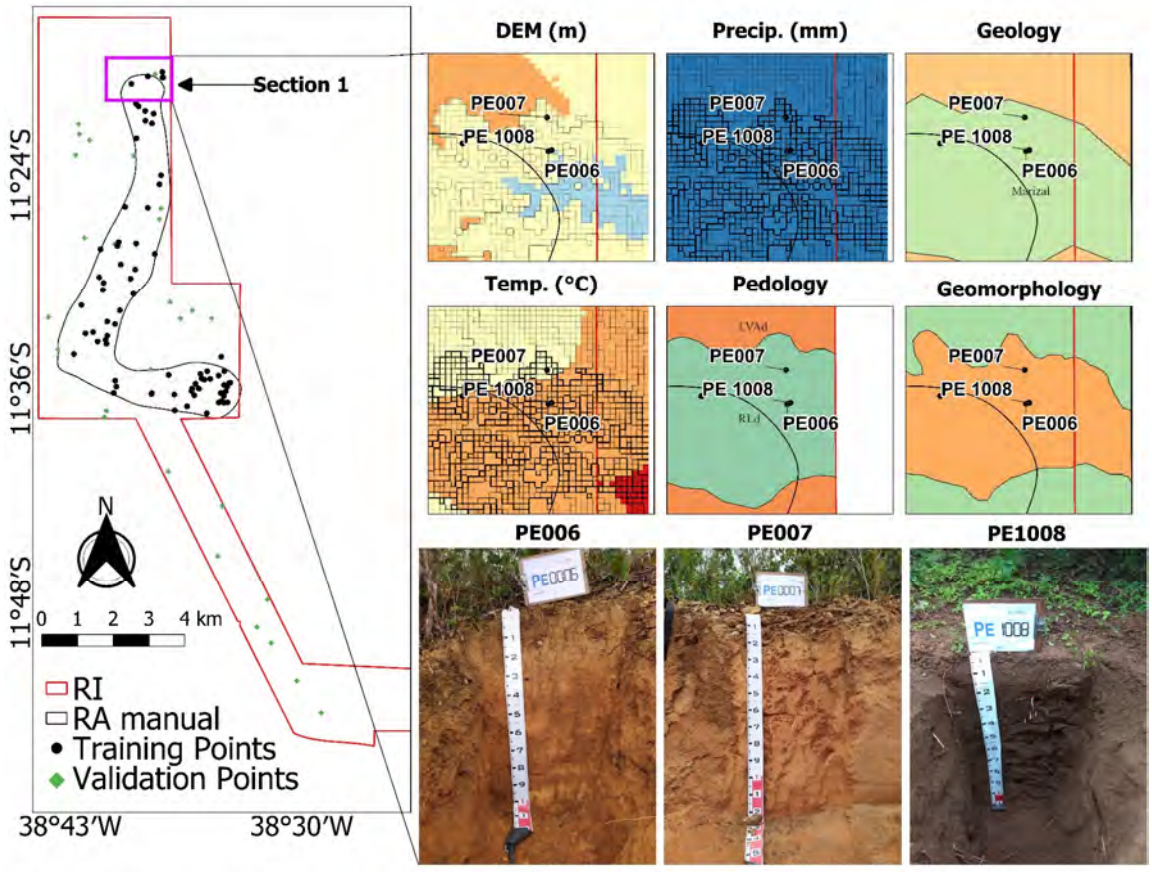


Figure 23. Section 1 for explaining the covariates and the soil type described.

Figure 24 represents Section 2, which features soil profiles PE1003 and PE1004 and the external validation point V2. Situated on a typical plateau, V2 represents the characteristic soil of upland regions, shaped by higher elevations and distinct geomorphological features. These soils are generally deep, well-drained, and exhibit minimal surface stoniness, indicative of stable environmental conditions(HORST-HEINEN et al., 2021).

In contrast, PE1004 is located in a flat lowland area and is characterized by sandy, weathered, drained soils with a lighter color and no rocky material. The covariates, such as precipitation and geology, highlight the influence of hydrological and geomorphic processes in these low-land terrains (CEDDIA et al., 2015). PE1003 occupies a transitional zone between the plateau and lowland regions. While it shares some characteristics with plateau soils, its moderate elevation and geomorphological transitions give it unique soil attributes reflective of its intermediary position.

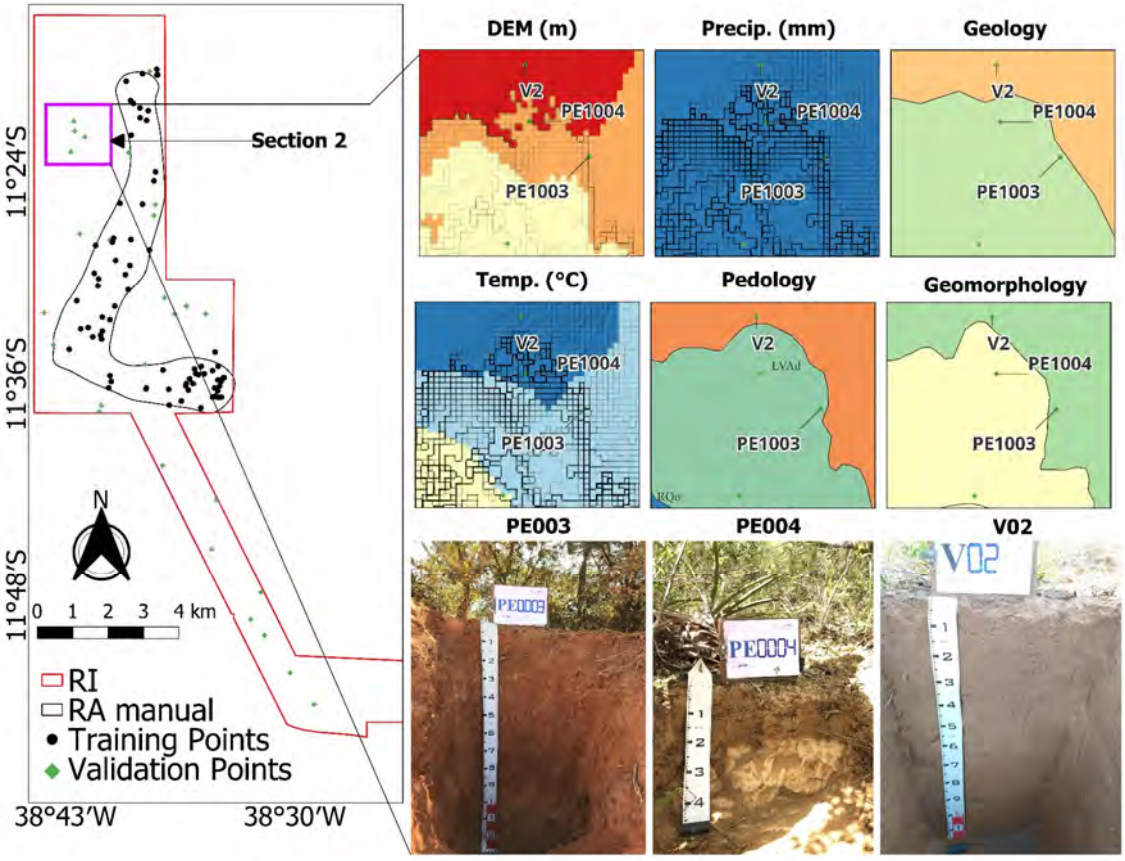


Figure 24. Section 2 for explaining the covariates and the soil type described.

Section 3, depicted in Figure 25, highlights the central portion of the region of interest (RI), focusing on profiles P806, PE808, and PE1007. Similar to the patterns observed in the northern area (Figure 25), this central region distinguishes between lowland and plateau soils.

Profile P806 represents the lowland soils characterized by sandy, deep, and excessively drained conditions. In contrast, plateau soils, exemplified by profile PE808, occur in flat terrain free of stones and rocks. These soils are deep, well-developed, and well-drained, with sandy textures and subsurface colors ranging from brown to light brown, reflecting stable geomorphic conditions (DA SILVA FREITAS; CAVALCANTI; NETO, 2024; SILVA et al., 2025).

A notable feature in this section is the presence of mottling in profile PE1007, similar to that observed in PE006. This characteristic is attributed to the influence of the soil's parent material and distinguishes it from the upper-third transitions and topsoil layers, where mottling is absent.

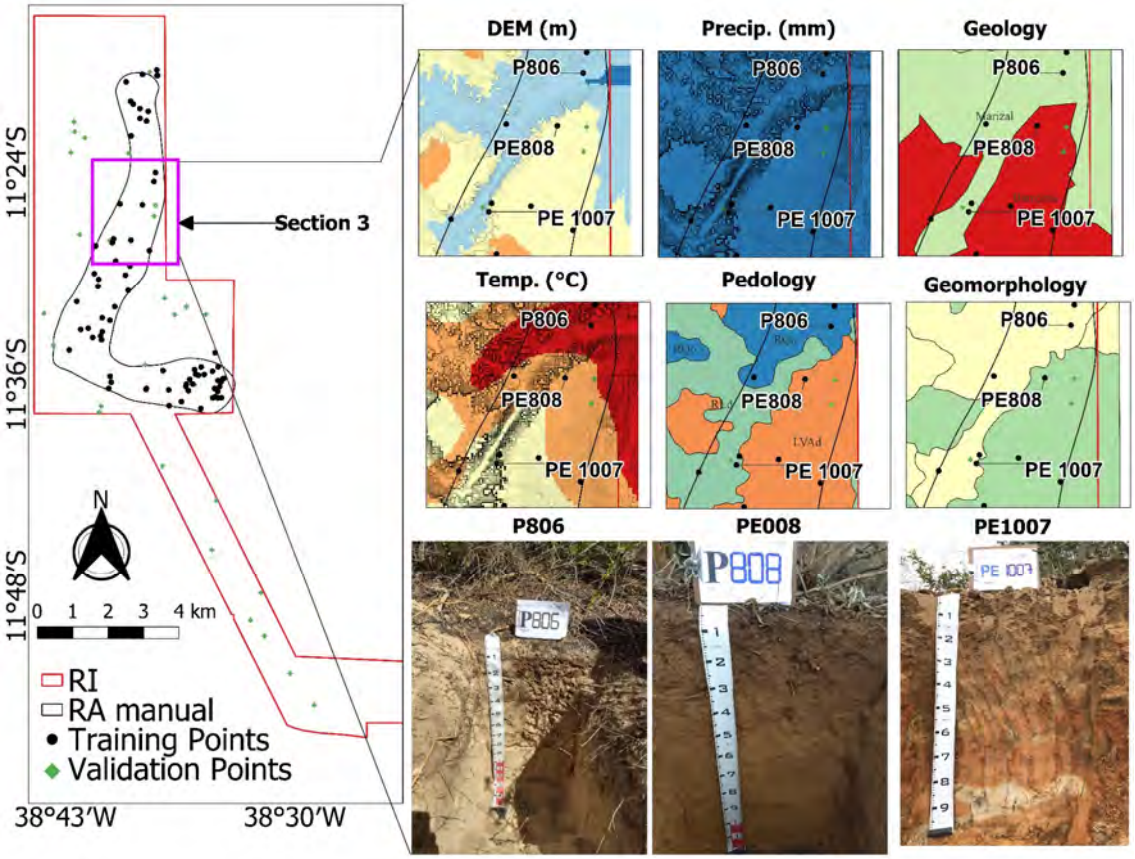


Figure 25. Section 3 for explaining the covariates and the soil type described.

Section 4, representing the southern part of the RI, is illustrated in Figure 26 and features profiles P826, P834, and P853, demonstrating notable geological and geomorphological variability. Profile P853, situated on rocky substrata, has shallow horizons influenced by lithic contact, while P826 exhibits deeper profiles with significant clay accumulation in the subsurface. These variations align with earlier pedology maps, underscoring the critical role of geological features in driving soil variability and emphasizing the importance of accounting for geomorphological heterogeneity in soil classification and management strategies (JUNIOR; NASCIMENTO, 2022).

The central portion of Section 4 focuses on plateau soils near the municipalities of Sátiro Dias and Biritinga. These soils, derived predominantly from Marizal Formation geological material, are deep, well-drained, and well-developed. They occur in non-stony, non-rocky terrain under native forest cover and in texture from sandy to medium. Subsurface colors exhibit a range of 5YR to 7.5YR hues, including yellow-brown, brown-yellow, brown, red-yellow, and reddish-brown tones.

Texturally, sandy loam dominates, but clay-silty loam and clayey loam are also present, highlighting the region's complexity and variation. This diversity, shaped by the interaction of parent material, geomorphological setting, and vegetation cover, illustrates the area's nuanced soil-landscape relationships.

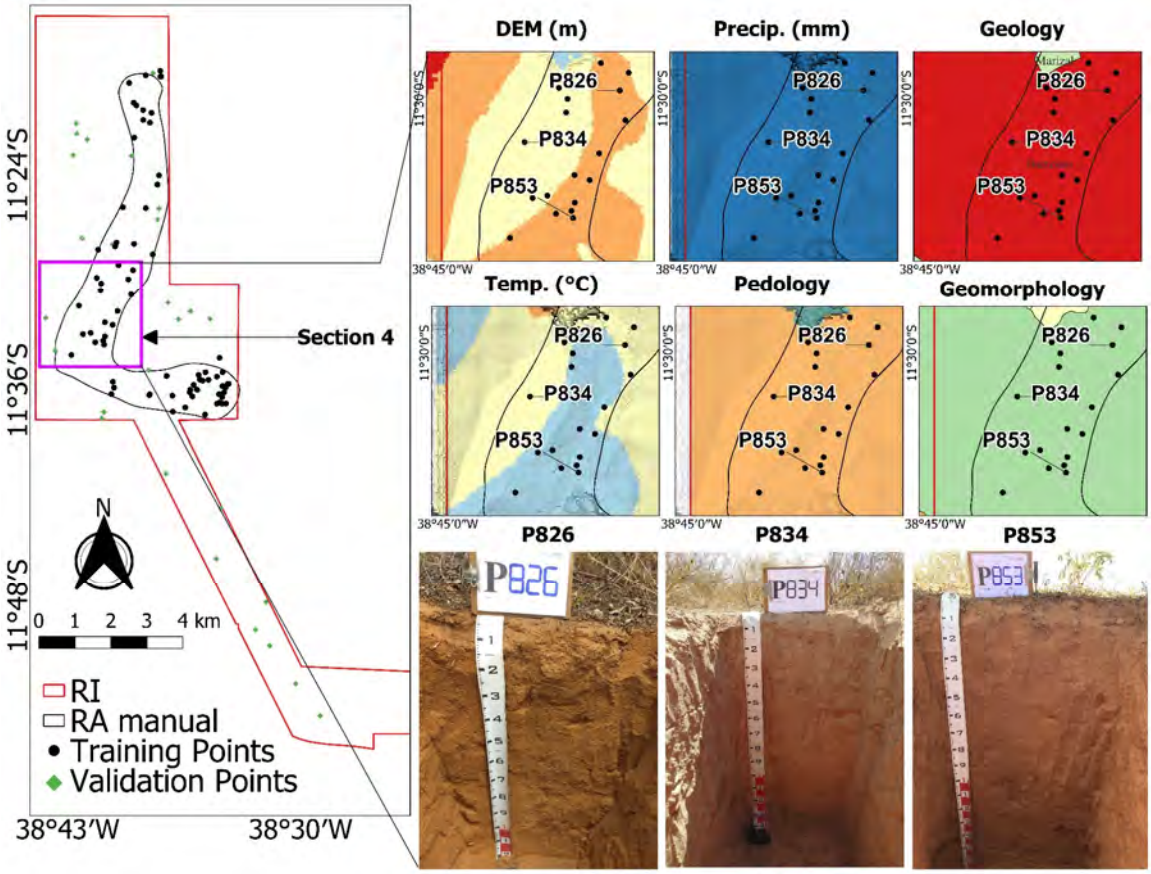


Figure 26. Section 4 explains the covariates and the soil type described.

Section 5, summarized in Figure 27, focuses on the southeastern corner of the RI, highlighting profiles V11, V12, and PE003. This region captures significant climatic and geomorphological transitions, as reflected in the covariate data. Profiles V11 and V12 demonstrate distinct drainage characteristics shaped by local topography, while pedology and geology maps reveal variations in soil-forming processes, emphasizing the critical role of integrating topographic and climatic data for precise soil delineation (GONÇALVES et al., 2022).

In the eastern plateau region (Figure 27), profiles generally follow the patterns described for plateau soils. Profile V11 is characterized by a brown color and loam to silty loam texture, while PE003 displays a red hue with a clay-silty loam texture. Both profiles are deep, well-developed, and sharply drained, with minimal horizon differentiation and no surface stoniness. The redder coloration of PE003, compared to P868, likely reflects differences in parent material. This profile is located at the transition zone from the plateau, resembling PE1003, to an area where two canyons begin to diverge.

Profile V12, situated in a transition zone between the plateau and the canyon, is located on the upper third of a slope with undulating relief. The soil in this position is stony, well-drained, and shallower than similar transitions observed in the northern region. Unlike north counterparts, the soils here are more influenced by erosive processes, suggesting variability in source material despite the underlying geology being associated with the Marizal Formation (MOMOLI; COOPER, 2016). Additionally, the soil exhibits mottling due to its parent material, with a brownish-yellow background color indicative of its transitional nature.

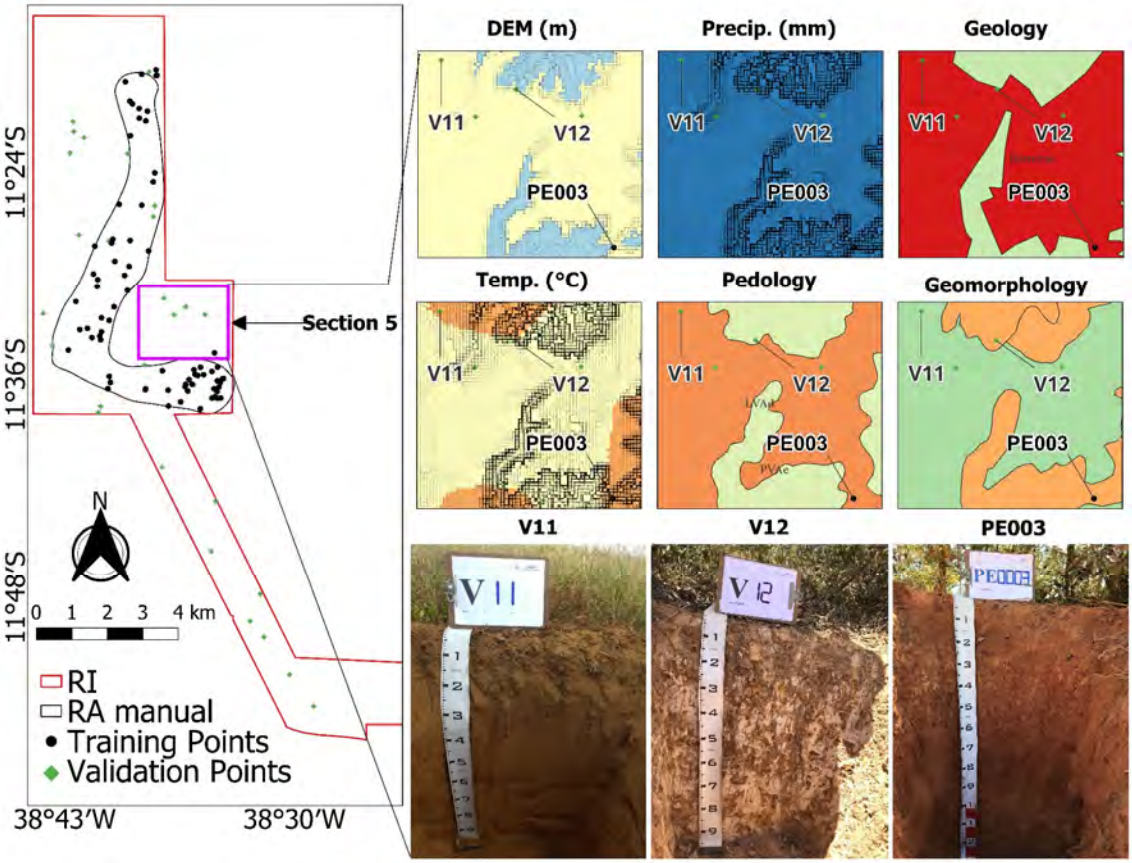


Figure 27. Section 5 for explaining the covariates and the soil type described.

Section 6, illustrated in Figure 28, represents the western region of interest (RI), highlighting profiles P875, P888, and V19. Profile P875 is characterized by reddish soils rich in iron oxides, reflecting its formation from specific geological materials, while profile P888 displays lighter horizons indicative of pronounced leaching and lower mineral content. These profiles underscore the influence of geological and climatic covariates on soil mineralogy, offering valuable insights for refined soil classification and management strategies (VAN WESTEN; CASTELLANOS; KURIAKOSE, 2008).

In the southwestern region of Sátiro Dias, profiles from the extreme south of Section 6, such as V19, share similarities with previously described plateau soils. These profiles are typically brown with sandy loam textures, exhibiting characteristics consistent with stable geomorphic conditions under native forest cover. In contrast, profiles P875 and P888 have loam textures with distinct red-yellow and reddish-brown hues, respectively, reflecting variability in parent material and soil-forming processes.

The soils in this section are predominantly well to excessively drained, with textures ranging from loam to sandy loam. They are characterized by aggregates of weak development and small size, transitioning from granular structures at the surface to subangular blocks in the subsurface. A notable feature of plateau soils in this region is the presence of shallow horizons with a low degree of development, often classified as weakly developed (FERNANDES et al., 2024). This variability highlights the interaction of parent material, drainage, and geomorphic processes in shaping soil characteristics across the region.

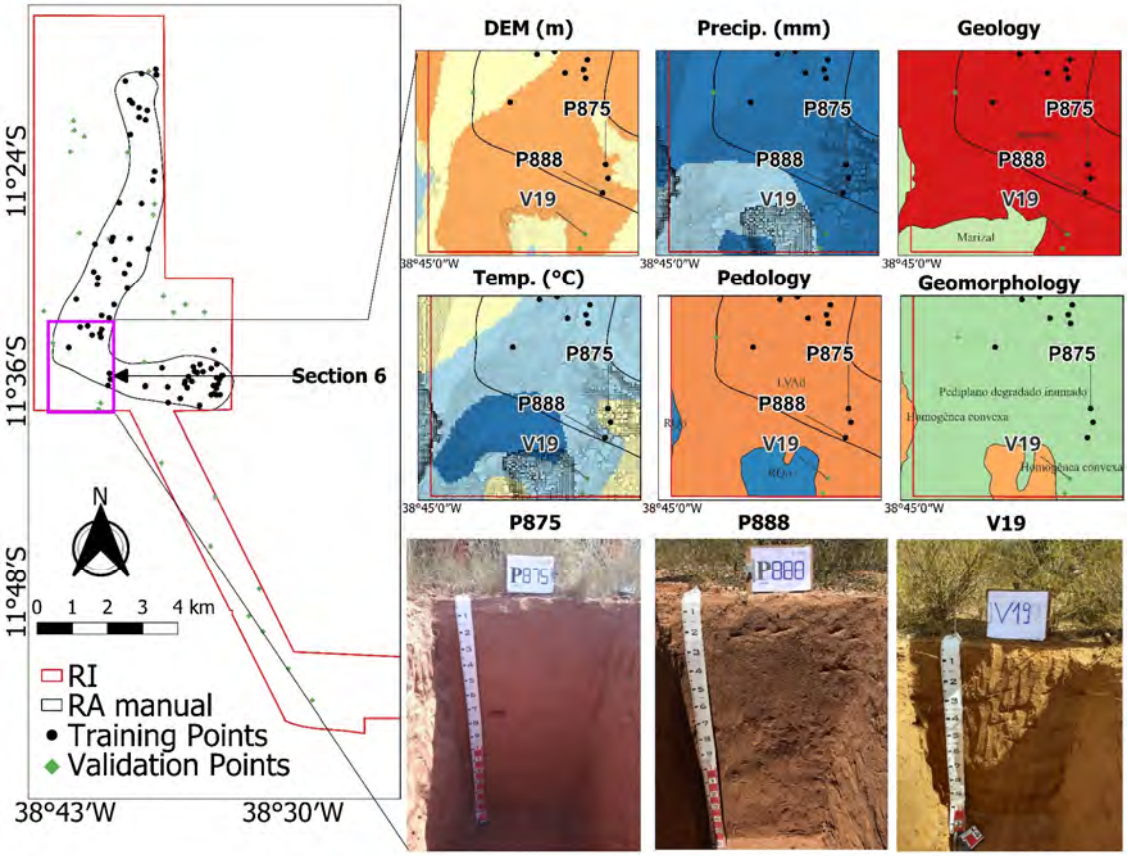


Figure 28. Section 6 explains the covariates and the soil type described.

Section 7, illustrated in Figure 29, highlights the southwestern corner of the region of interest (RI) and features profiles P868, PE005, and PE1006. Profile P868 is notable for its sharp transitions between horizons, indicative of high variability in soil-forming factors. At the same time, PE005 and PE1006 exemplify the influence of geomorphological patterns and precipitation gradients on soil development.

Profiles PE005 and PE1006 showcase significant differences despite their proximity. PE005 is characterized by a reddish color with dark red mottling, influenced by rock fragments rich in ferromagnesian minerals (PEREIRA; ANJOS, 1999). This young soil, situated on the upper third of a hill with undulating relief, exhibits abundant stoniness and a clay-silty loam texture with a high content of primary minerals, reflecting its less-developed state and strong lithological influence.

In contrast, PE1006, located in a flatter landscape nearby, represents a deep, well-developed soil with uniform horizons. Unlike PE005, it lacks stoniness or rockiness and aligns with the typical characteristics of plateau soils, illustrating the pronounced impact of landscape position on soil properties. These profiles highlight the intricate interplay between geomorphology, parent material, and drainage in shaping soil variability within this region.

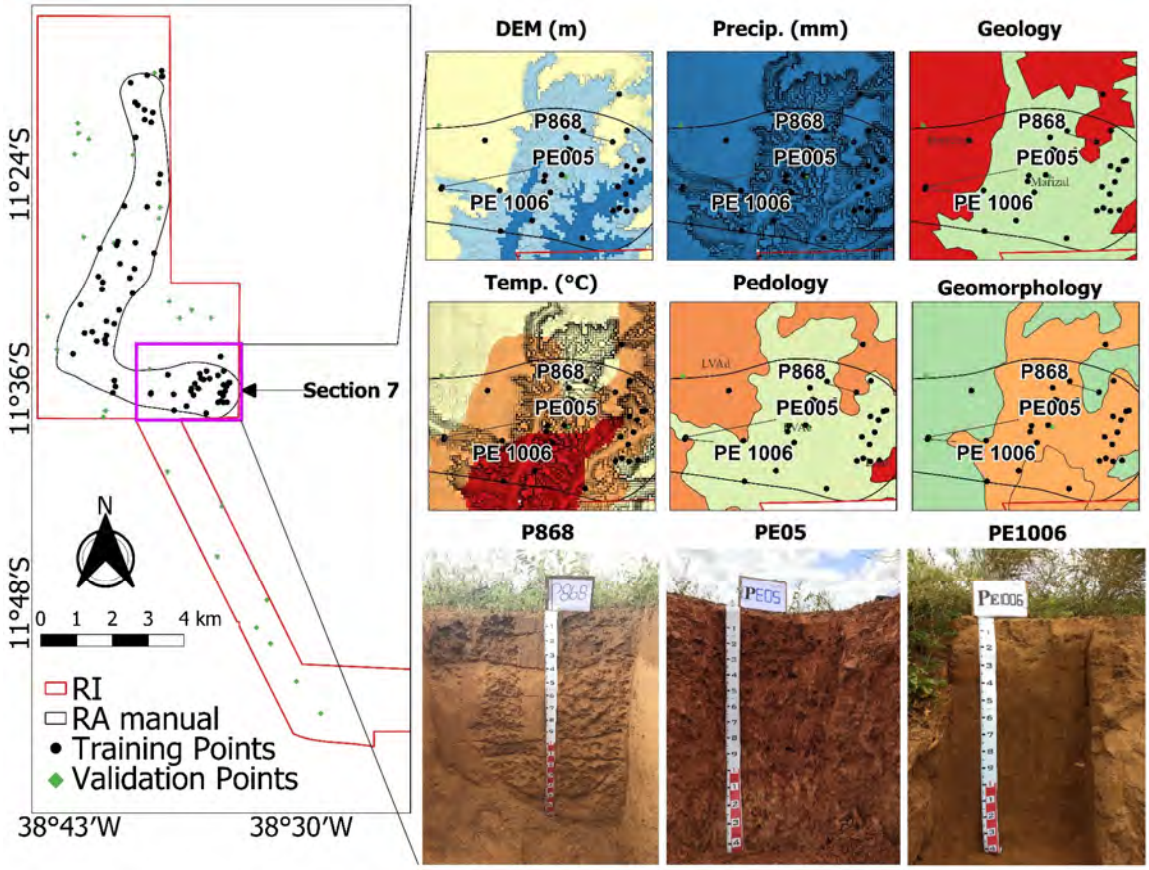


Figure 29. Section 7 for explaining the covariates and the soil type described.

3.5.2 The gower dissimilarity index map

The Gower Dissimilarity Index map (Figure 30) represents the spatial variation in dissimilarity across the RI, with values categorized into five distinct classes: ≤ 0.24 (blue), 0.24–0.27 (green), 0.27–0.32 (yellow), 0.32–0.45 (orange), and > 0.45 (red). This map provides a spatially explicit visualization of similar or dissimilar areas based on the multiple SCORPAN covariates used to calculate Gower's Dissimilarity Index.

Areas of high dissimilarity (>0.45) often feature heterogeneous attributes, typically where dominant and minor classes intersect or where pedological and geomorphological complexity abounds (e.g., Lithic Entisols, Eutrophic Udults with smooth relief, and transitions between "Tabular Homogeneous" and "Degraded Pediplane"). In contrast, low values of Gower's Dissimilarity Index (≤0.24) denote homogeneity regarding the covariates, often dominated by consistent classes such as Typic Udults (Pedology) or Tabular Homogeneous (Geomorphology).

Intermediate values of Gower's Dissimilarity Index (0.24–0.45) indicate transitions between homogeneous and heterogeneous zones (COSTA et al., 2024). For example, geological classes transition from "Barreiras" to "Santa Luz" in the southern part of Figure 30. The DEM and precipitation map also show opposing patterns, with maximum values in the northwest and minimum values in the south, as illustrated in Figure 30.

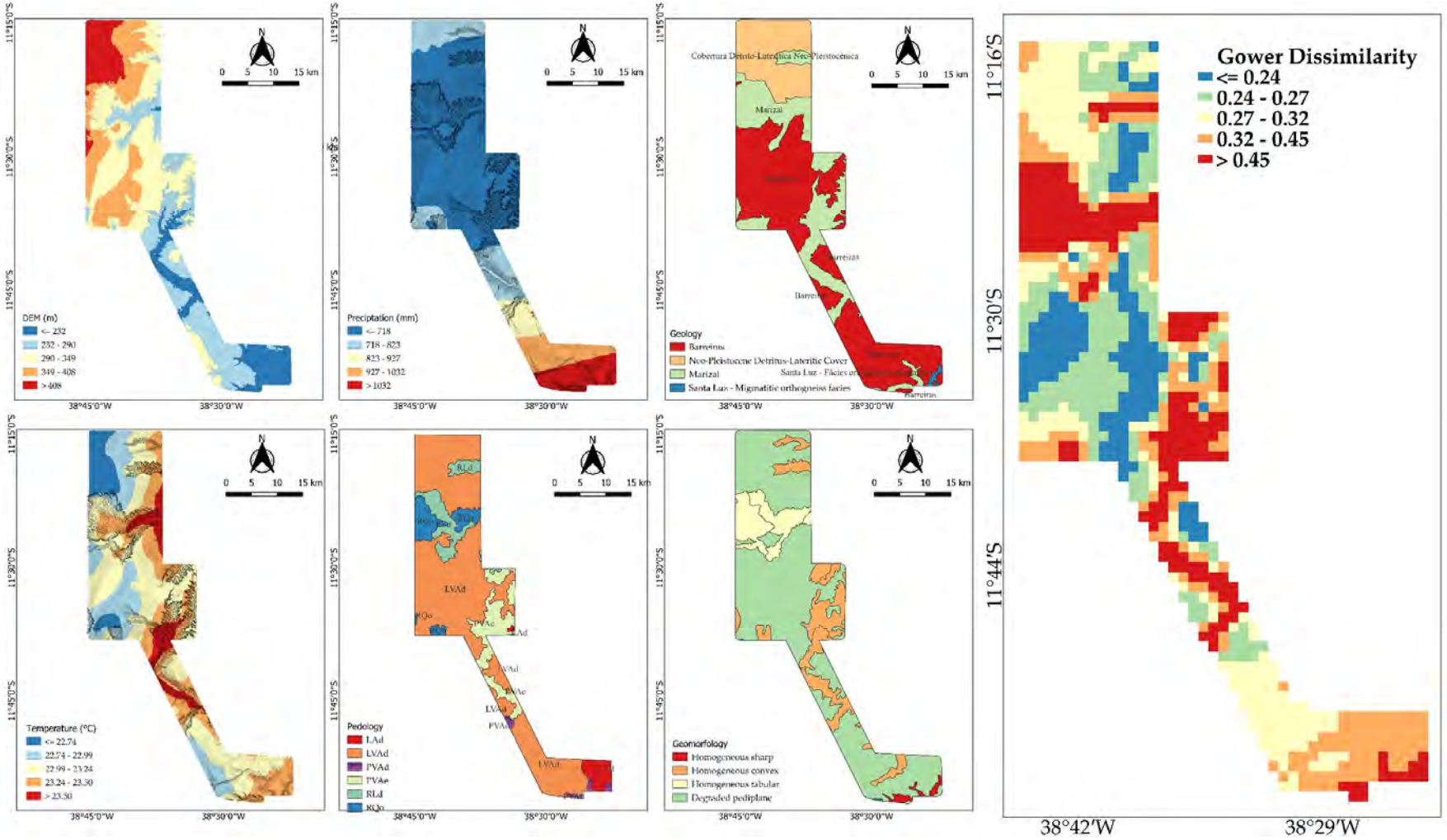


Figure 30. The covariates used in the gower's dissimilarity index map calculus.

3.5.3 Reference area delineation using autoRA and training and validation datasets

Figure 31 shows the spatial distribution of the training and validation datasets within the RA autoRA at varying target area percentages of the RI (10%, 20%, 30%, 40%, and 50%). The validation datasets (green points in Figure 31) remain distributed outside the autoRA-defined RA, serving as an external validation dataset to evaluate the performance of models constructed using the training data within each RA autoRA delineation.

At 10% coverage, the autoRA delineation captures 22 training points within the RA autoRA 10%, primarily focusing on areas with the highest Gower's Dissimilarity Index values. The remaining 80 points fell outside the RA autoRA 10%, serving as validation data. As the RA autoRA expands to 20% and 30%, the number of training points increases to 38 and 43, respectively, while the validation points decrease correspondingly. This trend continues with larger RA autoRA target area sizes, reaching 51 training points and 51 validation points at 40% coverage and 53 training points and 49 validation points at 50% coverage. The manually delineated RA manual includes 74 training points within its boundaries, leaving 28 points for validation.

Figure 31 highlights the autoRA methodology's capacity to adapt the RA boundaries to different levels of coverage while preserving the representativeness of soil-forming factors. Including points within high Gower's Dissimilarity Index values regions in smaller RAs autoRA target area sizes demonstrate the algorithm's focus on maximizing variability. Conversely, larger RAs enable a broader representation of the RI, which may benefit applications requiring wide spatial coverage.

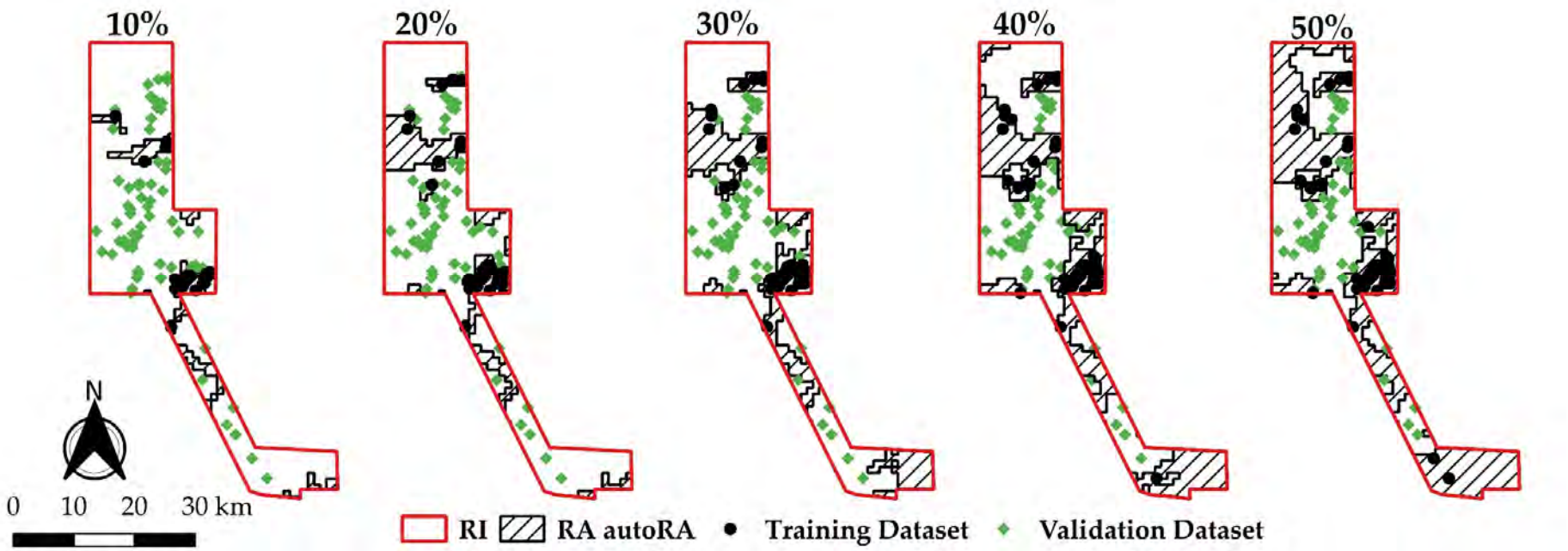


Figure 31. Reference Areas delineated by autoRA start from 10% of target area coverage concerning the Region of Interest (RI) with increments of 10% until 50%. The training and validation datasets were reclassified based on the intersection of the outlined RAs autoRA, with the training dataset, considered the inner points, and the external validation dataset, the validation points.

Figures 32 and 33 compares the pixel distributions of the study's covariates—Pedology, Geomorphology, Geology, Temperature, DEM, and Precipitation—using RI (employed by the TA to capture overall area variability), manual RA, and autoRA at reduction levels of 10%, 20%, 30%, 40%, and 50%. AutoRA at 30–40% closely matches RI across all covariates: in Pedology, LVAd and PVAd each represent ~35–40%, similar to RI's ~40%, whereas manual RA over-represents LAd1 (~40%) and under-represents LVAd and PVAd (~25–30%). For Geomorphology, autoRA at 30–50% balances BDP (~60–70%), HC (~30%), and HT (~10–15%) in alignment with RI, while manual RA over-represents HC (~35%) and under-represents BDP (~50%).

In Geology, autoRA at 30–40% mirrors RI with category B (~40%), FO (~25%), and M (~35%), compared to manual RA, which over-represents category B (~70%) and under-represents FO (~10%) and M (~20%). Temperature is accurately captured by autoRA at 30–50%, maintaining the dominant 23°C range (~60%) and adjacent transitions at 22°C (~20–25%) and 24°C (~15–20%), unlike manual RA, which under-represents the 23°C range (~40%) and fails to capture adjacent temperatures.

For DEM, autoRA at 30–50% balances elevations across peaks (~250–300 m) and lower areas (~0–200 m), closely mirroring RI, whereas manual RA smooths the distribution and under-represents higher elevations. In Precipitation, autoRA at 30–40% closely matches RI by distributing low (~2–3%), moderate (~1.5–2%), and high (~0.5%) precipitation ranges. At the same time, manual RA oversimplifies the distribution, over-representing low precipitation (~4%) and under-representing moderate (~1%) and high (<0.5%) ranges.

Overall, autoRA at 30–40% provides the most balanced and accurate representation compared to RI, whereas manual RA often over-represents dominant classes or ranges, reducing variability. Lower autoRA thresholds (10–20%) significantly diverge from RI, and a 50% reduction introduces minor over-representations in some variables, such as lower elevations in DEM and LAd1 in Pedology.

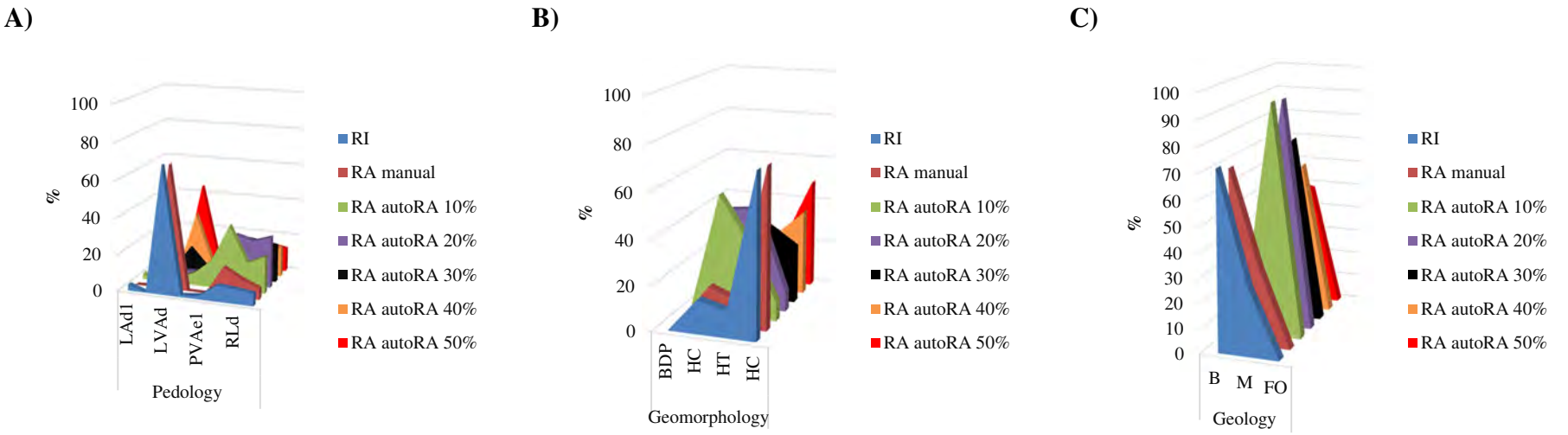


Figure 32. Graph comparing the capabilities of the autoRA and retrieving the most heterogeneous information of the covariates by representing the pixel frequency at each class/value. **A)** Pedology: LAd1, Typic Udults; LAd2, Typic Udults (clayed texture); LVAd, Typic Udults with mixed hematite and goethite; PVAd, Typic Ultisol; PVAe1, Eutrophic Udults on a smooth relief; PVAe2, Eutrophic Udults on a wave relief; RLd, Lithic Entisols. RQo, Orthic Entisol. **B)** Geomorphology: BDP, Buried Degraded Pediplain; HC, Homogeneous Convex; HT, Homogeneous Tabular; HS, Homogeneous Sharp. **C)** Gelogy: B, Barreiras Group; M, Marizal Group; FO, Orthogneiss-Migmatite Facies. (to be continued…).

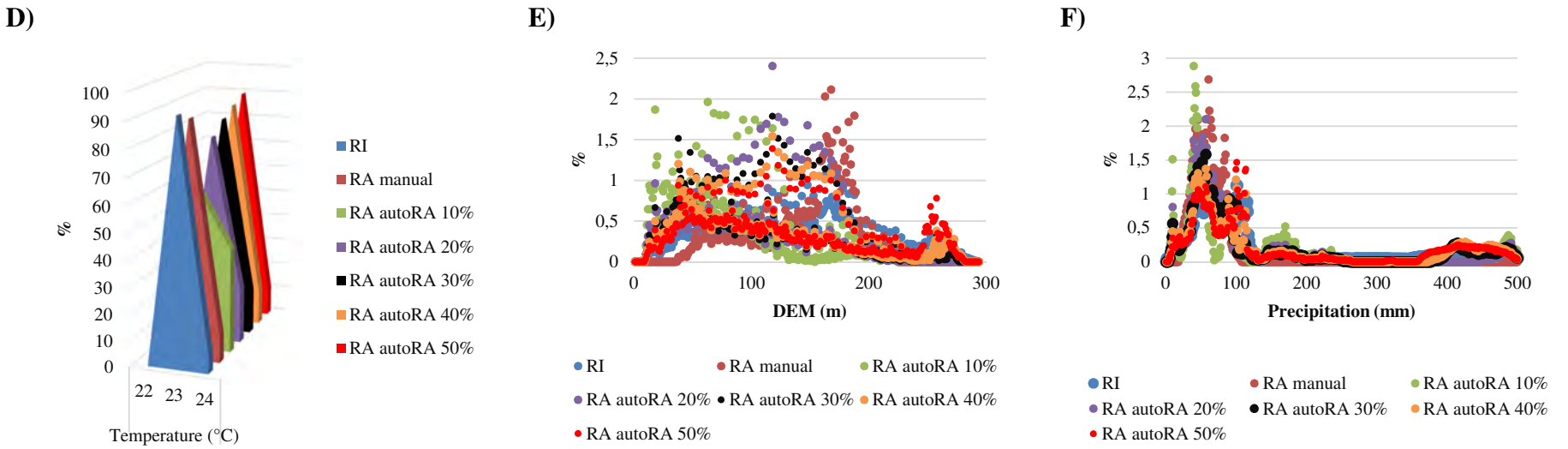


Figure 33. Graph comparing the capabilities of the autoRA and retrieving the most heterogeneous information of the covariates by representing the pixel frequency at each class/value (Continuation). **D)** Year Average Temperature; **E)** DEM, Digital elevation model; **F)** Year Average Precipitation.

3.5.4 Soil maps and performance

Figure 34 displays the MU maps for Sátiro Dias created by three approaches—RA manual (benchmark), autoRA, and TA—each with respective accuracy and agreement measures (Table 2). The RA manual approach, intentionally adopted during the sampling design survey, achieved 0.75 accuracies and a Kappa of 0.50, thus establishing a baseline for comparison with the other methods, autoRA and TA.

The autoRA method performed poorly at 10% and 20% target areas, with accuracy dropping to around 0.14–0.11 and minimal Kappa (0.06–0.01). These low scores highlight that insufficient spatial coverage (only 10–20% of the territory) fails to capture environmental variability, leading to weak model training. However, at 30% coverage, the autoRA-based model displayed a considerable leap in accuracy (0.85) and a moderate Kappa (0.42), suggesting that a broader spatial representation bolsters predictive performance by sampling more heterogeneous zones. When coverage reached 40% and 50%, the model attained accuracies of 0.96 (Kappa=0.65 for 40% and 0.49 for 50%), surpassing the RA manual and nearly matching the top performance of the TA approach.

Figure 34. Variation of the Overall Accuracy and Index Kappa results for the 100 sampling seeds splitting to mapping units using the total area (TA) approach.

Table 2. External validation metrics for each mapping unit map obtained from the different groupings of training data in association with the reference area approach (RA manual and autoRA) method and total area (TA).

	N Training	N Validation	Accuracy	Kappa
RA manual	74	28	0,75	0,50
RA autoRA 10%	22	72	0,14	0,06
RA autoRA 20%	38	64	0,11	0,01
RA autoRA 30%	43	59	0,85	0,42
RA autoRA 40%	51	51	0,96	0,65
RA autoRA 50%	53	49	0,96	0,49
TA	74	28	0,84	0,74

The TA method allocates training/validation sets randomly across the entire domain, yielding an overall accuracy of 0,84 and a Kappa of 0,74, with best splits achieving 0,93 and 0,89, respectively (Figure 35). This indicates that randomly sampling the TA —and repeating splits to stabilize estimates—can generate highly reliable models (CARVALHO et al., 2020; MCBRATNEY; MENDONÇA SANTOS; MINASNY, 2003).

Nonetheless, autoRA at 30%, 40%, and 50% successfully narrowed the gap. This indicates that a targeted and oriented sampling design considering spatially diverse units can outperform a static RA manual design and challenge the comprehensive TA approach. In particular, 40%autoRA stands out for balancing coverage and modeling success, suggesting that automated, diversity-driven approaches can reduce fieldwork while preserving (and sometimes surpassing) accuracy at this proportional RI-reduced area.

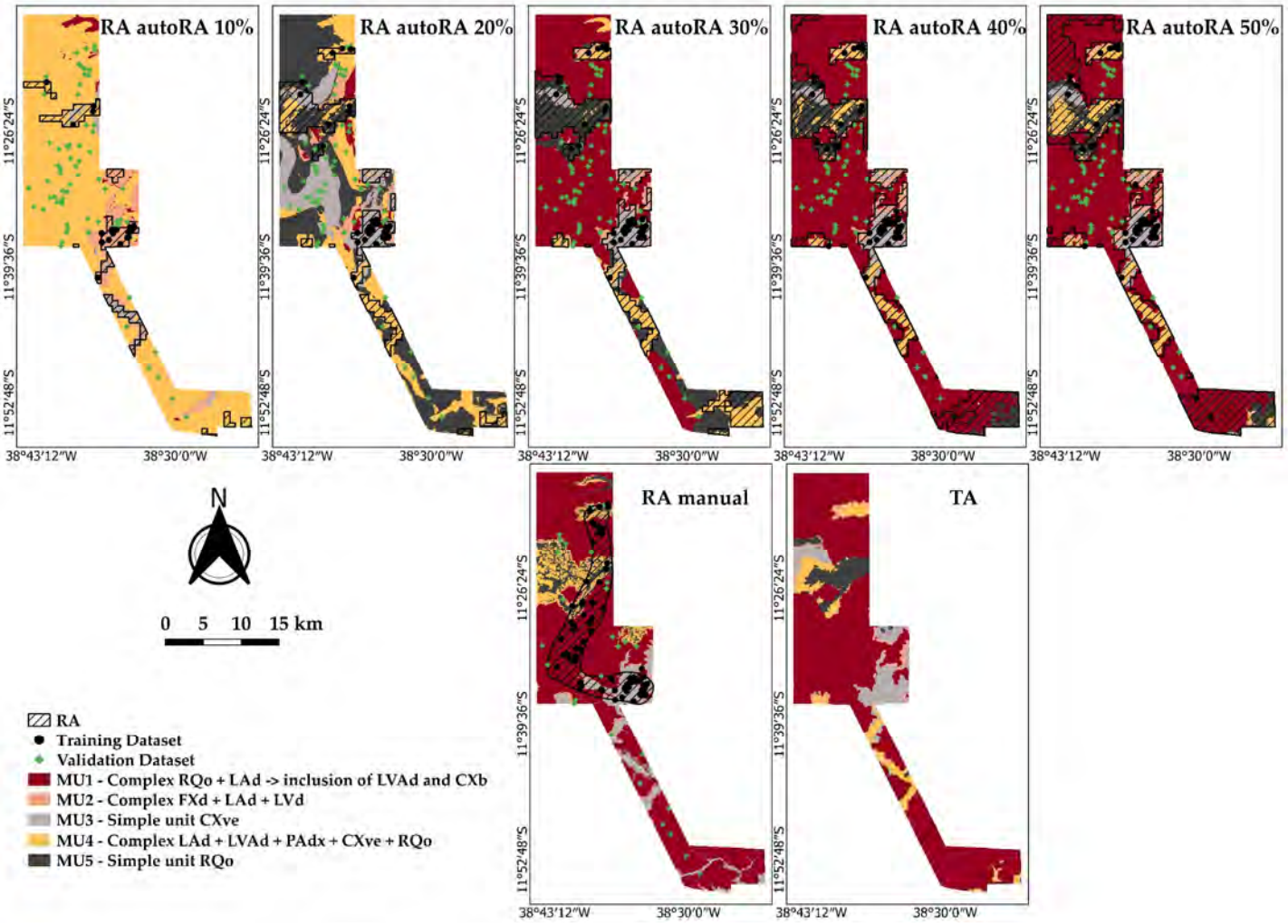


Figure 35. Mapping Units from the dataset located within the manually delineated reference area (RA). Overlap: RA boundaries dashed, training points in black circles, and external validation points in green lozenges; TA, Total Area.

Figure 36 shows the MU frequencies across the training datasets (RA manual, RA autoRA-target areas, TA), revealing each approach's strengths and shortcomings. The TA dataset, typical in DSM workflows, relied heavily on MU1 and underrepresented MU2, MU4, and MU5, risking poor model generalization for those classes(BARUCK et al., 2016; NEYESTANI et al., 2021; ODGERS; MCBRATNEY; CARRÉ, 2018). Meanwhile, the RA manual—guided by expert judgment—also concentrates on MU1, indicating potential subjective bias.

In contrast, RA autoRA at higher percentages (40% and 50%) balances MU representation, notably improving coverage of MU4 and MU5 and reducing MU1's dominance, thus enhancing reproducibility and scalability. The validation dataset confirmed that RA autoRA consistently encompasses a broader range of units (especially MU3, MU4, MU5) than the TA or RA manual methods, underscoring its capacity to produce more comprehensive datasets during the training soil sampling design.

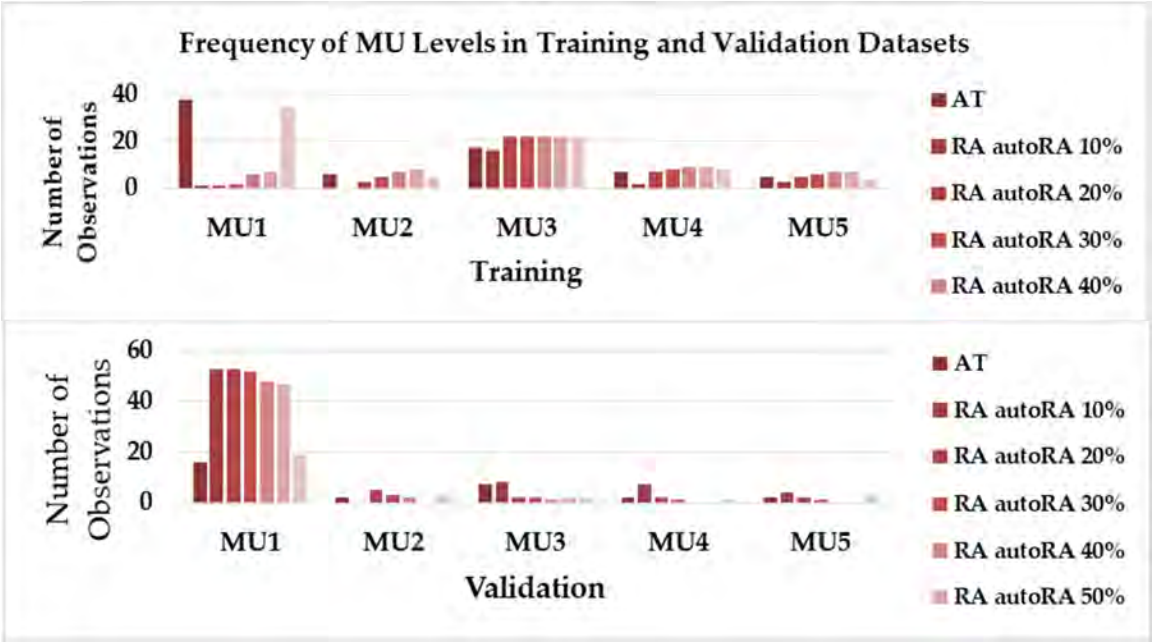


Figure 346. Frequency of MUs class on each dataset for training and validation dataset.

The confusion matrix (Table 3) highlights a significant concentration of MU1, showing that autoRA's optimized sampling distribution provides superior representation of soil variability compared to the RA manual and TA methods. This is particularly evident when mitigating the overrepresentation of dominant units. When autoRA coverage is low (10–20%), it fails to capture sufficient heterogeneity, resulting in weaker model performance and higher misclassification rates (DEVINE; STEENWERTH; O'GEEN, 2021; ODGERS; MCBRATNEY; CARRÉ, 2018).

In contrast, increasing autoRA coverage to 30–50% leads to a more even distribution of samples across MUs, enhancing overall model performance by capturing more significant soil heterogeneity and reducing classification errors, especially in underrepresented mapping units. Although the TA method encompasses the full range of RI variability, its overall accuracy does not consistently surpass that of autoRA, suggesting that a selective, diversity-driven approach can be advantageous (ZHANG et al., 2023).

Evaluating RA manual, autoRA, and TA methods reveals MU representation and accuracy differences using the Weighted Producer Accuracy Index (WPAI), Weighted User Accuracy Index (WUAI), and area extension metrics. Under the RA manual approach, MU1 achieves a high User Accuracy (0,95) and perfect Producer's Accuracy, indicating reliable classification in a significant, homogeneous region. The RA manual also included MU3, MU4, and MU5 profiles in the external validation dataset, as shown in the confusion matrix. Specifically, MU3 and MU4 had moderate representation, whereas MU5 exhibited low classification accuracy due to limited validation samples (User's Accuracy of 0,5 and 0, respectively). The WPAI for MU1 (0,71) further validates the RA manual's effectiveness in capturing large, consistent MUs while exposing its limitations in addressing variability within smaller classes (HENRYS; MONDAIN-MONVAL; JARVIS, 2024).

Conversely, autoRA's performance across different sample proportions demonstrates its capability to account for more diverse MU variability. At 10% coverage, high User Accuracy (0,95) for MU3 and MU5 indicates successful detection of dominant variability, supported by a WPAI of 0,76 for MU1. However, the low Producer's Accuracy for MU5 (0,05) highlights the difficulty in fully representing smaller or more variable units with sparse sampling. Increasing coverage to 20% improves Producer's Accuracy for MU1 (0,95) and introduces additional MUs (e.g., MU2, MU5) into the validation set, as indicated by a WUAI of 0,59 for MU1(COSTA et al., 2021). At 30% coverage, MU1's User's Accuracy (0,92) and WPAI (0,59) reflect enhanced classification stability, while coverage levels of 40–50% achieve near-perfect performance for MU1. This underscores autoRA's increasing reliability when the training set encompasses sufficient variability (GOMES et al., 2023; HUANG et al., 2017).

Table 3. Confusion Matrix of RA manual, RA autoRA, and Total Area (TA).

RA manual										
Reference	MU1	MU2	MU3	MU4	MU5	Total	User's Accuracy	Area (km²)	WPAI	WUAI
MU1	19	0	0	0	1	20	0,95	636,96		
MU2	0	1	0	0	0	1	1	28,6		
MU3	0	0	1	1	0	2	0,5	95,29	0,71	0,00
MU4	0	2	1	*	2	5	---	78,1		
MU5	0	0	0	0	*	0	---	62,12		
Total	19	3	2	1	3	28	---			
Producer's Accuracy	1	0,33	0,5	0	0	---	---			
RA autoRA 10%										
Reference	MU1	MU2	MU3	MU4	MU5	Total	User's Accuracy	Area (km²)	WPAI	WUAI
MU1	*	0	0	0	53	53	*	14,89		
MU2	0	*	0	0	0	0	*	96,42		
MU3	0	0	7	0	1	8	0,95	29,79	0,76	0,00
MU4	1	0	2	*	4	7	*	746,44		
MU5	1	0	0	0	3	4	0,95	0		
Total	2	0	9	0	61	72	---			
Producer's Accuracy	*	*	0,78	*	0,05	---	---			
RA autoRA 20%										
Reference	MU1	MU2	MU3	MU4	MU5	Total	User's Accuracy	Area (km²)	WPAI	WUAI
MU1	4	4	18	13	14	53	0,08	26,55		
MU2	1	*	2	1	1	5	*	33,58		
MU3	0	2	*	0	0	2	*	193,41	0,03	0,59
MU4	0	0	0	1	1	2	0,5	228,04		
MU5	0	0	0	0	2	2	1	405,96		
Total	5	6	20	15	18	64	---			
Producer's Accuracy	0,95	*	*	0,07	0,11	---	---			

To be continued…

Continuation of **Table 3.**

RA autoRA 30%										
Reference	MU1	MU2	MU3	MU4	MU5	Total	User's Accuracy	Area (km²)	WPAI	WUAI
MU1	48	0	0	0	4	52	0,92	506,48		
MU2	1	1	0	0	1	3	0,33	33,18		
MU3	0	2	*	0	0	2	*	60,05	0,59	0,46
MU4	0	0	0	*	1	1	*	96,09		
MU5	0	0	0	0	1	1	1	191,74		
Total	49	3	0	0	7	59	---			
Producer's Accuracy	0,98	0,33	*	*	0,14	---	---			
RA autoRA 40%										
Reference	MU1	MU2	MU3	MU4	MU5	Total	User's Accuracy	Area (km²)	WPAI	WUAI
MU1	48	0	0	0	0	48	1	637,68		
MU2	0	1	0	0	0	1	1	39,57		
MU3	0	2	*	0	0	2	*	58,23	0,76	0,22
MU4	0	0	0	*	0	0	*	56,95		
MU5	0	0	0	0	*	0	*	95,11		
Total	48	3	0	0	0	49	---			
Producer's Accuracy	1	0,33	*	*	*	---	---			
RA autoRA 50%										
Reference	MU1	MU2	MU3	MU4	MU5	Total	User's Accuracy	Area (km²)	WPAI	WUAI
MU1	47	0	0	0	0	47	1	631,64		
MU2	0	*	0	0	0	0	*	43,36		
MU3	0	2	*	0	0	2	*	58,95	0,76	0,00
MU4	0	0	0	*	0	0	*	78,35		
MU5	0	0	0	0	*	0	*	75,24		
Total	47	2	0	0	0	49	---			
Producer's Accuracy	1	*	*	*	*	---	---			
Total Area										
	MU1	MU2	MU3	MU4	MU5	Total	User's Accuracy	Area (km²)	WPAI	WUAI
MU1	16	0	0	0	0	16	1	652,88		
MU2	0	*	0	0	0	0	*	27,19		
MU3	0	2	7	0	1	10	0,70	80,9	0,84	0,00
MU4	0	0	0	1	1	2	0,5	69,96		
MU5	0	0	0	1	*	1	*	56,61		
Total	16	2	7	2	2	29	---			
Producer's Accuracy	1	*	1	0,5	*	---	---			

* There is no reference class in the validation dataset.

Figure 37 illustrates that among the tested scenarios, the MU distribution under autoRA with 40% coverage is the most balanced. In the RA manual method, MU3, MU4, and MU5 were well represented in the validation dataset, ensuring a balanced external validation. However, under autoRA 40% coverage, these MUs lack validation points because their profiles were included in the training dataset. As a result, the predicted STS map using autoRA 40% coverage achieves high accuracy for MU3, MU4, and MU5 due to their presence in the training set. Despite the absence of validation points for these MUs—stemming from the original sampling strategy being developed at a different time—the overall accuracy of the predicted STS map remains high at 0,96.

Specifically, for MU5, profiles V9 and PE1004, initially designated for validation, were included in the training. Similarly, for MU4, profile V20 was selected for training, and MU3, profiles V18 and V3 were also allocated to the training set. Additionally, for MU2, validation profiles V1 and PE1003 were included in the training set by autoRA 40%. This imbalance highlights autoRA's emphasis on selecting profiles with higher dissimilarity, which shifted specific profiles from validation to training. Meanwhile, the highly sampled MU1 and MU2 were retained in the validation dataset due to their lower dissimilarity scores.

This distribution indicates that while the RA manual method ensured the representation of smaller classes like MU3, MU4, and MU5 in the validation dataset, autoRA prioritized diversity in training, resulting in different trade-offs. The TA method provides a valuable baseline: for MU1,

a User's Accuracy of 1 and a WPAI of 0,84 confirm strong agreement across its extensive distribution (652,88 km²). However, smaller or more variable classes, such as MU3 (User's Accuracy of 0,7), remain challenging to capture consistently across the domain. Overall, while RA manual effectively characterizes large, consistent units, autoRA excels in addressing variability and shows improved performance as sampling proportions increase (ABDULRAHEEM et al., 2023; ZHANG et al., 2022). The WPAI and WUAI indices illustrate how sampling strategies and inherent heterogeneity influence accuracy and reliability. Ultimately, autoRA is distinguished by its ability to accommodate high variability, steadily enhancing performance as the proportion of training samples increases.

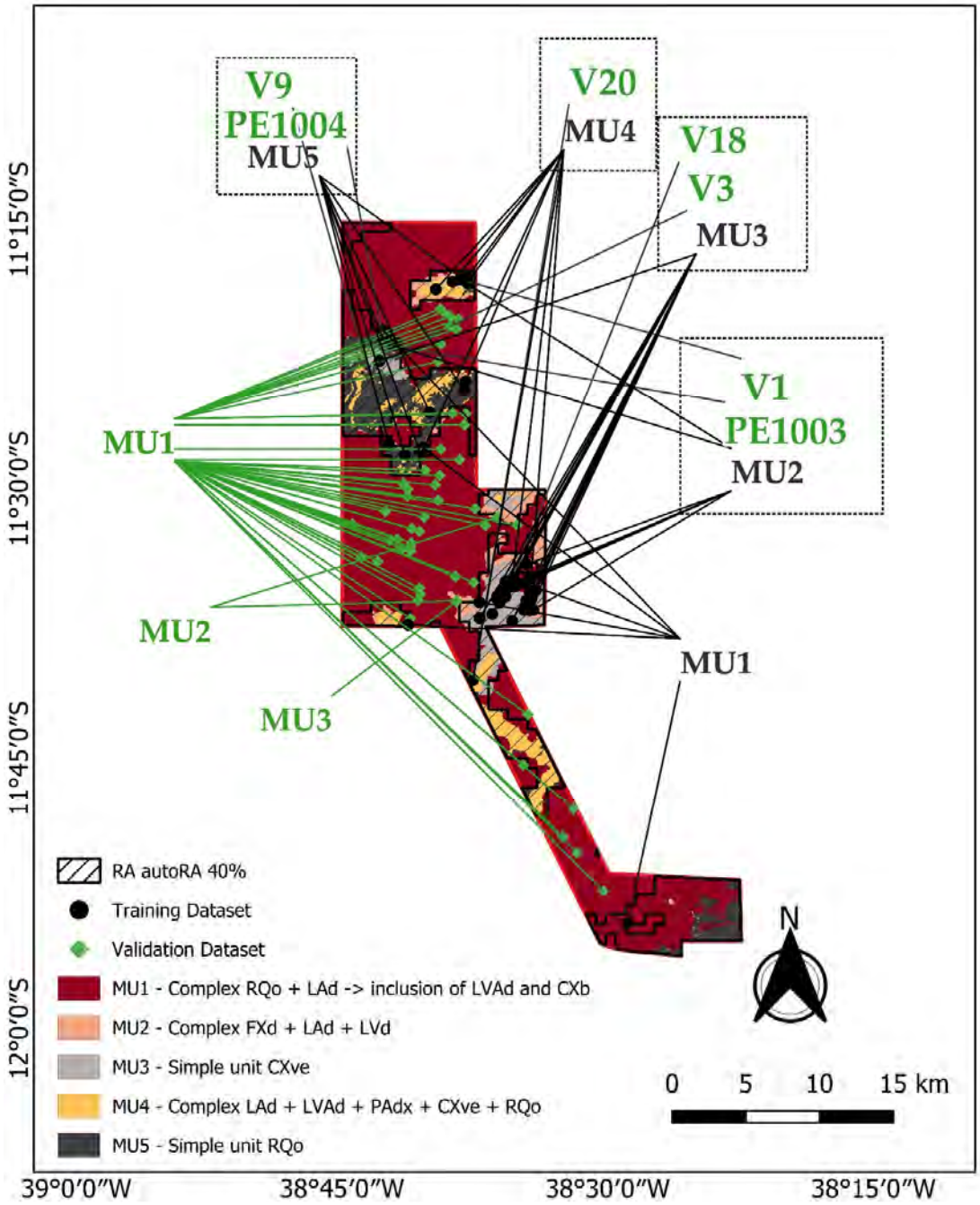


Figure 37. RA autoRA at 40% alongside the corresponding MU map and detailed validation profiles, illustrating their inclusion in the training dataset.

3.6 CONCLUSIONS

The results underscored the pivotal role of delineating an adequately sized reference area (RA) and ensuring sufficient training points for accurately mapping soil Mapping Units (MU). Applying autoRA based on the highest similarity index values proved key to capturing the full range of environmental variability. It enhanced the agreement between generated MU maps and the benchmark MU map. As the RA increased from 10% to 50%—and training points grew accordingly—external validation metrics (Overall Accuracy, Index Kappa) rose markedly, confirming that broader RA coverage strengthened model performance.

Lower coverage at 10% and 20% yielded insufficient performance (accuracies of 0.14 and 0.11), demonstrating that minimal training data failed to capture environmental complexity. Moving to 30% triggered a step change (accuracy=0.85, Kappa=0.42), while 40% and 50% produced even higher accuracy (0.96), with 40% offering the best balance (Kappa=0.65). Although the 50% map was also accurate, redundancy in the extensive RA reduced agreement (Kappa=0.49). Meanwhile, manual RA mapping, although moderately effective (accuracy=0.75, Kappa=0.50), struggled with complex transitions, evidencing the merits of data-driven automation by autoRA.

Comparisons with the TA approach—where points spanned the entire study region—indicated that TA could deliver strong results but did not consistently surpass autoRA at 40%. While TA benefited from broader coverage of the study area, it oversampled certain features and failed to capture key environmental gradients systematically. In contrast, autoRA strategically pinpointed diverse zones, avoided redundancy, and enhanced modeling performance.

Thus, autoRA emerged as a tool for pre-campaign planning, allowing users to delineate small RAs that encapsulated essential environmental variability and oriented a sampling design toward high-diversity areas regarding soil-forming factors. Specifically, a well-calibrated RA of around 40% proved highly effective, balancing resource demands compared to conventional splitting DSM workflows. These findings fully addressed our initial hypothesis, demonstrating that an automated method for delineating RA significantly reduced subjectivity and enhanced effectiveness—yielding improved predictive accuracy and lower costs—while bolstering reproducibility and scalability for RA-based soil mapping within the Pedometrics community.

4. CHAPTER III

AUTORA: AN AUTOMATIC REFERENCE AREA ALGORITHM FOR OPTIMIZE SOIL SAMPLING AND MAPPING

4.1 RESUMO

A aplicação da abordagem autoRA (em inglês Automatic Reference Area) na Margem Equatorial do Brasil mostrou-se uma estratégia altamente eficaz para otimizar metodologias de amostragem de solo que equilibram a precisão do modelo com a eficiência de custos. O ponto central para esse sucesso foi a incorporação de testes de sensibilidade facilitados por 100 iterações algébricas da Superfície Teórica Simulada (STS). Essas iterações foram projetadas para refletir cenários adimensionais associados às propriedades derivadas do modelo de solo SCORPAN (Solo, Clima, Organismos, Relevo, Material de Origem, Localização e Tempo). Ao realizar 100 simulações STS e calcular a média dos resultados, o autoRA garantiu uma representação robusta da variabilidade ambiental. Por meio da metodologia de mapeamento via Área Total (TA), um tamanho amostral de n=800 foi identificado como ótimo, alcançando alto coeficiente de determinação (R2=0,97), baixo raiz quadrática do erro médio (RMSE=0,34), mínimo viés (Bias) (0,02) e Distância Euclidiana de 0,01 em relação ao modelo de referência TA. Essa configuração reduzindo as despesas em US$ 200.000 em comparação com um tamanho de amostra maior de 1.000 em TA. A validação do Modelo de Área de Referência (RAM) com um tamanho amostral reduzido de 600 pontos demonstrou mais flexibilidade e eficiência do autoRA com a menor métrica ED via autoRA com 40% da área total recoberta, alcançando R2=0,87, RMSE=0,68, Bias=0,08 e ED = 0,21. Essa validação destacou a capacidade do autoRA de concentrar os esforços de amostragem nas regiões mais heterogêneas, conforme determinado pelo Índice de Dissimilaridade de Gower. Notavelmente, as métricas apresentadas para o Modelo Predito Exaustivo (EPM) e o RAM foram derivadas de validações externas de Random Forest (RF) dos modelos previstos pelo STS. Cada modelo de RF foi avaliado utilizando validação externa com conjunto de 300 amostras em TA, garantindo que as avaliações de desempenho fossem confiáveis e comparáveis em diferentes estratégias de amostragem. Consequentemente, o autoRA se destacou como uma ferramenta poderosa para aumentar a eficiência e a precisão dos estudos ambientais, facilitando a tomada de decisões sustentáveis e informadas na ciência do solo e disciplinas relacionadas.

Palavras-chave: Amostragem de solo. Mapeamento Digital de Solos. Otimização de amostra.

4.2. ABSTRACT

Applying the autoRA (Automated Reference Area) approach within the Equatorial Margin of Brazil proved to be a highly effective strategy for optimizing soil sampling methodologies that balanced model accuracy with cost efficiency. Central to this success was incorporating sensitivity tests facilitated by 100 algebraic iterations of the Simulated Theoretical Surface (STS). These iterations were meticulously designed to reflect feasible, dimensionless covariate scenarios associated with SCORPAN (Soil, Climate, Organisms, Relief, Parent Material, Location, and Time) soil-derived properties. By performing 100 STS simulations and averaging the results, autoRA ensured a robust representation of environmental variability, which is critical for accurate predictive modeling. Through the Total Area (TA) methodology, a sample size of n=800 was identified as optimal, achieving a high coefficient of determination (R^2=0.97), low Root Mean Square Error (RMSE=0.34), minimal Bias (0.02), and an impressive Euclidean Distance (ED=0.01) relative to the benchmark model. This configuration delivered exceptional model performance and resulted in substantial cost savings, reducing expenditure by $200,000 compared to a larger sample size of 1,000. The spatial distribution analysis confirmed that the model effectively captured regional environmental heterogeneity, which is essential for accurate predictive insights in complex terrains. Furthermore, the autoRA methodology's ability to utilize the same map covariates in both pre-fieldwork planning and post-fieldwork modeling stages underscored its utility as a comprehensive tool for project planning. By aligning the sampling strategy with the covariates used in digital soil mapping, autoRA allowed users to accurately dimension critical project parameters, including budget allocation, survey labor days, and acceptable levels of accuracy. The validation of the Reference Area Model (RAM) with a reduced sample size of 600 points further demonstrated autoRA's flexibility and efficiency with the lowest ED metric among the STS maps achieved using autoRA via the target area of 40% of the total area, underscoring the effectiveness of this configuration in capturing regional environmental heterogeneity. Achieving R^2=0.87, RMSE=0.68, Bias=0.08, and ED=0.21, the RAM maintained substantial model accuracy while significantly reducing field effort and associated costs. This validation highlighted autoRA's capacity to focus sampling efforts on the most heterogeneous regions, as determined by Gower's Dissimilarity Index, thereby optimizing resource utilization without compromising the integrity of the predictive model. Notably, the metrics presented for both the Exhaustive Predicted Model (EPM) and the RAM were derived from Random Forest (RF) external validations of the STS-predicted models. Each RF model was rigorously tested against a consistent external validation set of 300 samples, ensuring that performance assessments were reliable and comparable across different sampling strategies. This systematic validation approach underscored autoRA's robustness in handling varying target area sizes and proportions of Gower's Dissimilarity Index, affirming its applicability across diverse environmental contexts. Its ability to integrate sensitivity testing through simulated theoretical surface iterations, maximize entropy by capturing spatial variability, and align pre-planning with post-fieldwork covariate usage ensured that sampling projects were economically viable and scientifically sound. Consequently, autoRA stood out as an invaluable tool for enhancing the efficiency and accuracy of environmental studies, facilitating sustainable and informed decision-making in soil science and related disciplines.

Keywords: Soil Sampling. Digital Soil Mapping. Sample Optimization.

4.3 INTRODUCTION

Soil sampling is a critical component of soil science, providing essential data for understanding soil properties and dynamics (CARTER; GREGORICH, 2007; GRUIJTER, 2006). However, obtaining representative soil samples across large, data-limited regions presents significant challenges. These challenges are particularly pronounced in areas with limited or no existing soil samples or detailed soil maps (GRUNWALD 2010; HARTEMINK; MCBRATNEY; MENDONÇA-SANTOS, 2008), such as the Equatorial Margin of Brazil that encompasses the States of Amapá, Pará, Maranhão, Piauí, Ceará, and Rio Grande do Norte (TERUIYA et al., 2008). These regions face unique logistical, financial, and environmental constraints, making developing efficient and cost-effective soil sampling strategies vital (FURTADO; PONTE, 2013; GUIMARAES FILHO; BORBA, 2020).

The soil sampling process has undergone significant innovation over the years, with various strategies emerging to address representativeness and cost-effectiveness. Approaches like Latin Hypercube Sampling (cLHS), conditioned to environmental covariates, aim to maximize coverage of the feature space (BRUNGARD; BOETTINGER, 2010; MA et al., 2020; MINASNY; MCBRATNEY, 2006). More recently, Malone et al. (2019) proposed strategies to improve the cLHS application in the Pedometrics including optimizing the sample size, re-locating sites when an original site is deemed inaccessible, and accounting for existing sample data so that under-sampled areas can be prioritized for sampling. Yet, in the approach of the cLHS, Saurette et al. (2024a) We present a way to retain the minimum sample needed for the cLHS based on histograms of the predictor variables, using the Freedman-Diaconis rule to determine optimal bin width.

Mallavan et al. (2010) introduced the Homosoil method to address soil information gaps by identifying donor and recipient sites based on climatic, topographic, and lithological covariates. Their approach computes a similarity index via Gower's distance to pinpoint regions with comparable soil-forming factors. It hypothesizes that soil mapping rules from donor areas can be transferred to poorly characterized sites. However, the hierarchical weighting of climate, geology, and topography in calculating the Gower index may prove restrictive in settings with scarce data, as it can mask locally significant covariates and limit overall model flexibility.

Building upon this foundation, Nenkam et al. (2023, 2022) introduce refinements that make the Homosoil framework more robust for regions lacking comprehensive data. They use equal weighting of soil-forming factors in calculating distances, cluster environmental covariates to boost computational efficiency, and incorporate Mahalanobis distance with percentile-based thresholds for defining homosoils. Moreover, they emphasize adding local soil samples in the calibration dataset to enhance model accuracy, drawing attention to the need for updated, harmonized soil data in poorly mapped areas. Despite these methodological advances, both Mallavan's Homosoil framework and Nenkam's refined strategies emphasize the need for additional soil samples to improve model robustness. However, they offer limited guidance on systematically collecting new data in regions with scarce existing datasets.

The concept of Reference Areas (RA) effectively addresses soil sampling challenges by identifying smaller, representative zones within a more extensive study region. Instead of sampling the Total Area (TA), researchers focus on these RAs, which encapsulate the key environmental and soil characteristics of the broader landscape (ARRUDA et al., 2016; LAGACHERIE et al., 2001; LAGACHERIE; VOLTZ, 2000; VOLTZ; LAGACHERIE; LOUCHART, 1997; YIGINI; PANAGOS, 2014). Soil sampling within RA allows for the development of models that accurately predict soil properties across the entire region (FERREIRA et al., 2022; FERREIRA, 2023).

However, the traditional methods for delineating an RA have relied heavily on experts' manual delineation, introducing subjectivity and limiting reproducibility (FAVROT, 1986; FAVROT, 1981; M. BORNAND; FAVROT, 1998). This limitation, compared to the modern artificial intelligence algorithm implemented in the field of Digital Soil Mapping (DSM), required an update on the process of creating the RA in a more automatic way (DE CARVALHO JUNIOR et al., 2024; GRUNWALD 2021; WADOUX et al., 2021).

Automating the delineation of RA reduced reliance on subjective expert judgment and offered a reproducible solution for large, data-scarce regions (JEAN-MARC ROBBEZ-MASSON, 1994; LAGACHERIE; LEGROS; BURROUGH, 1995). To that end, we introduced autoRA (Automatic Reference Area), an algorithm that selected a smaller yet representative fraction of a Total Area (TA) using Gower's Dissimilarity Index. Specifically, pixels in the TA were ranked by high or low dissimilarity and grouped in varying proportions (e.g., 100/0, 90/10, 80/20, down to 10/90, 0/100), each forming a candidate RA. This yielded multiple subsets that differed in overall area and internal heterogeneity, which aimed to capture enough environmental variability to maintain predictive accuracy.

The algorithm's performance was assessed with a Simulated Theoretical Surface (STS), a dimensionless n-interaction algebraic surface formed by randomly iterating basic arithmetic operations (addition, subtraction, multiplication) and power functions (1st and 2nd degree) on the same covariates that informed the Gower's Dissimilarity Index. A final average from the n-interactions was taken as the STS. A full-coverage (TA-based) sampling first established a "gold-standard" STS model fitted via Random Forest. Then, each RA scenario was sampled at multiple densities smaller than the TA-based design. New Random Forest models were fitted to these RA samples, each repeating the seed multiple times to minimize randomness in the model outcomes. The resultant STS predictions were validated against an independent set of points spread across the TA to determine how effectively RA-based sampling extrapolated the broader region.

Such an approach was well suited to the Equatorial Margin of Brazil—an area of approximately 242,000 km² spanning the States of Amapá, Pará, and Maranhão—where difficult access, dense vegetation, and limited soil data constrained extensive field campaigns (GUIMARAES FILHO; BORBA, 2020; TERUIYA et al., 2008). The present study hypothesizes that by systematically adjusting RA size, proportions of high/low dissimilarity values, and repeated sampling for Random Forest modeling, autoRA identified the minimal portion of the TA that yielded predictive performance comparable to full-coverage sampling.

The objective of this research was: i) to present autoRA as a reproducible, data-driven procedure for RA delineation, and ii) to determine the slightest sampling effort necessary to achieve an STS predictive accuracy that rivaled the exhaustive TA approach, as demonstrated in the Equatorial Margin of Brazil.

4.4. MATERIAL AND METHODS

4.4.1 Study area

The study concentrated on the Equatorial Margin of Brazil, encompassing the states of Pará, Amapá, and Maranhão and spanning an extensive area of approximately 242,000 km² (Figure 38). This region is strategically significant due to its remarkable ecological diversity and considerable economic potential, particularly in expanding oil and gas exploration activities (ABREU, 1949; PEREIRA, et al., 2022). Despite its importance, there was a notable deficiency in detailed soil maps and comprehensive soil data for this area (MARQUES et al., 2019). To address this gap, the current study employed the autoRA algorithm to develop a strategic and optimized soil sampling plan. By targeting the most heterogeneous regions within the Equatorial Margin, the autoRA algorithm ensured efficient utilization of limited survey resources while maximizing the representativeness and accuracy of the collected soil data.

The Equatorial Margin was delineated using a circumscription that extended 200 km inland from the coastline, capturing the terrestrial areas most directly influenced by coastal and offshore activities (FRANCINI-FILHO et al., 2018). This delineation ensured the inclusion of zones susceptible to environmental impacts, which made it a critical focus for environmental monitoring and planning (MOURA, L. et al., 2016). Major urban centers such as Macapá, Belém, and São Luís were situated within this boundary, highlighting the socioeconomic and environmental significance of the area (CARNEIRO et al., 2022; SOARES et al., 2021).

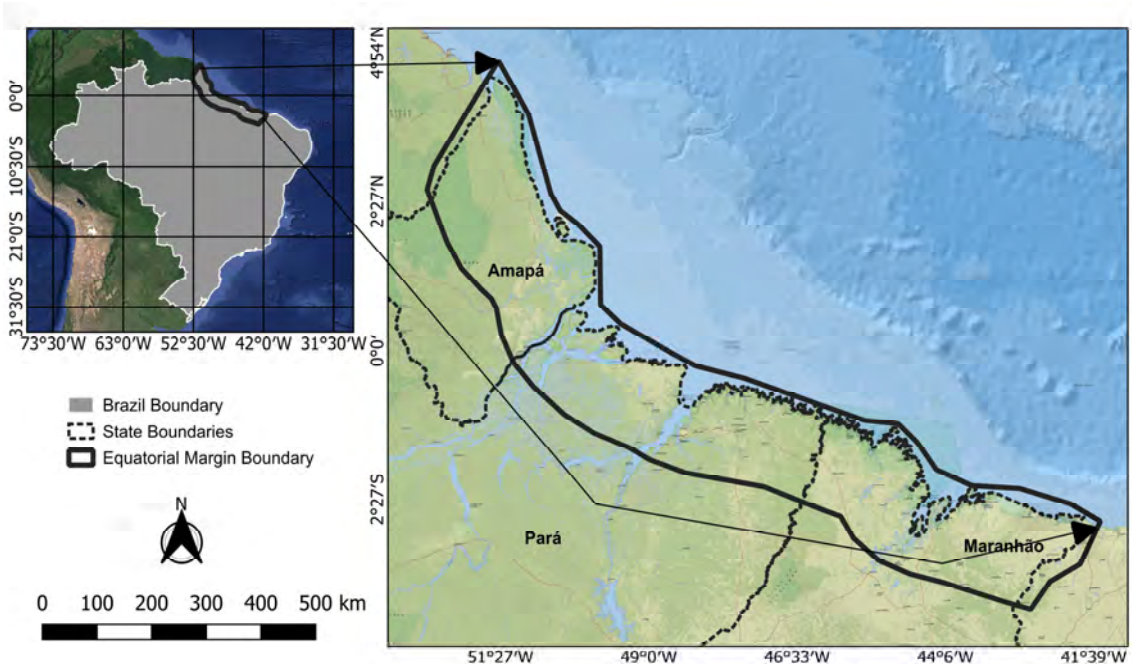


Figure 358. Location of equatorial margin of Brazil.

4.4.2 Data

a) Environmental covariates

Figure 39 provided a comprehensive overview of the environmental covariates in the study region, encompassing both categorical and numerical variables. The categorical variables—geology, geomorphology, and pedology—contributed to the region's significant soil diversity, driven by the dynamic interplay of fluvial, marine, and terrestrial processes (GONÇALVES et al., 2022). Predominant soil types included Oxisols, Ultisols, and hydromorphic soils such as Aquic Entisols and Aquic Ultisols. These soils reflected geomorphological and hydrological influences, with salic, sodic, and arenic characteristics at their horizons enhancing their heterogeneity (BARROS et al., 2018; BIJOS; DA SILVA; MUNHOZ, 2023; IBGE, 2018b; LOPES; MARIANO-NETO; AMORIM, 2016; MOREIRA; SIQUEIRA; BRUSSAARD, 2006).

Geomorphologically, as shown in Figure 39B, the region was shaped by fluvial, coastal, and tectonic processes (MAIA; BEZERRA, 2020). Coastal lowlands featured mangroves, estuaries, and tidal flats, which were influenced by tidal amplitudes up to 7 meters (COSTA et al., 2020). Extensive floodplains arose from seasonal hydrological variability and sediment deposition by the Amazon River and its tributaries, while upland residual plateaus represented stable, weathered terrains (FREIRE et al., 2017).

Geological data from IBGE (2018) at a 1:250,000 scale indicated that the area was underlain by sedimentary and crystalline formations, including the Gurupi, Parnaíba, and São Luís groups (Figure 39A), which significantly influenced soil Properties (DE ALKMIM, 2015). The Amazon Basin within the region consisted of Tertiary and Quaternary sands, silts, and clays transported by the Amazon River (ROSSETTI, 2006; SCARPELLI; HORIKAVA, 2018). The Parnaíba Basin in Maranhão was composed of Paleozoic sedimentary rocks such as sandstone and shale, while coastal areas were dominated by Holocene marine and fluvial sediments (ADRIANO et al., 2015; SCARPELLI; HORIKAVA, 2017).

Numerical variables included climate and topography, crucial in shaping the environmental landscape. Climate data from WorldClim (FICK; HIJMANS, 2017) at a 1 km resolution showed an equatorial (Af) climate transitioning to monsoonal (Am) near Maranhão (BOVOLO et al., 2012; TEEGAVARAPU; SHARMA, 2021). Annual precipitation ranged from 2,000 mm to 3,000 mm (Figure 39E), with a wet season from December to May. Temperatures were stable (Figure 39F), averaging 24°C–27°C, and relative humidity often exceeded 80% (DANTAS et al., 2022). Topography, represented by a 90-meter resolution DEM from TOPODATA (DE MORISSON VALERIANO; DE FÁTIMA ROSSETTI, 2012), varied in altitude from 2.7 m to 420 m, influencing soil and vegetation distribution (Figure 39D).

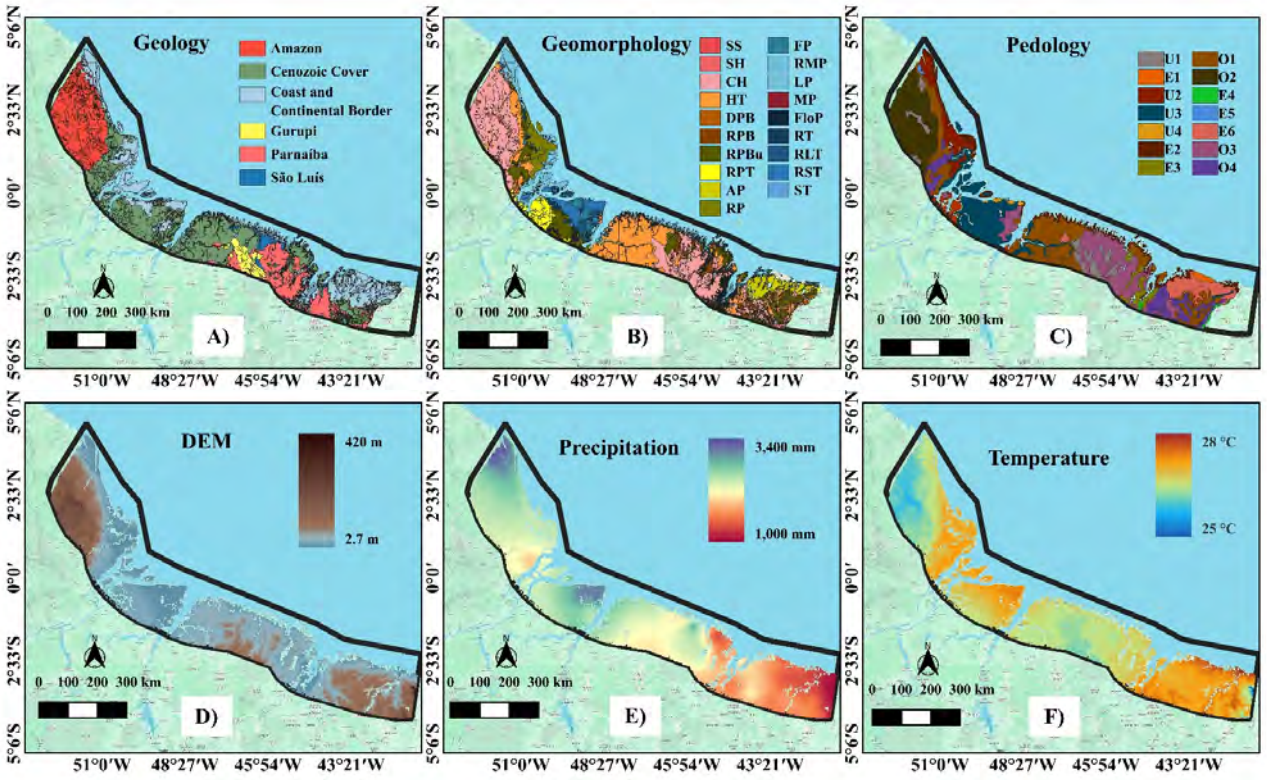


Figure 369. A) Geology map for the Equatorial Margin area at scale 1:250,000; B) Geomorphology map at scale 1:250,000; C) Pedology map at scale 1:250,000; D) DEM with 90 m resolution; E) Precipitation map at 1 km resolution; F) Temperature map at 1 km spatial resolution. U1, Argissolos Vermelho-Amarelos Distróficos; E1, Dunas; U2, Gleissolos Háplicos Ta Distróficos; U3, Gleissolos Háplicos Ta Eutróficos; U4, Gleissolos Háplicos Tb Distróficos; E2, Gleissolos Sálicos Órticos; E3, Gleissolos Sálicos Sódicos; U5, Gleissolos Tiomórficos Órticos; O1, Latossolos Amarelos Distróficos; O2, Latossolos Vermelho-Amarelos Distróficos; E4, Neossolos Flúvicos Tb Distróficos; E5, Neossolos Litólicos Distróficos; E6, Neossolos Quartzarênicos Órticos; O3, Plintossolos Háplicos Distróficos; O4, Plintossolos Pétricos Concrecionários; Sharp structural, SS; Sharp homogeneous, SH; Convex homogeneous, CH; Homogeneous tabular, HT; Degraded pediplane buried,DPB; Retouched pediplane bare, RPB; Retouched pediplane buried, RPBu; River plain and terrace, RPT; Aeolian plain,AP; River plain, RP; Fluviolacustrine plain, FP; River-marine plain, RMP; Lake plain, LP; Marine plain, MP; Flood plan, FloP; River terrace, RT; River-lake terrace, RLT; River-sea terrace, RST; Sea terrace,ST.

Figure 40 illustrates the distribution of land use and land cover within the study area, distinguishing between human-managed and environmental categories. Yellow areas indicated regions used for agricultural activities—such as cultivating bulbs, roots, and tubers—alongside large-scale animal farming and plant extraction in forested zones. In contrast, green areas represented conservation units with complete protection and indigenous lands, highlighting forest preservation and limited human intervention. This spatial distribution reflected the coexistence of productive land use and forest conservation efforts, embodying the region's diverse socio-environmental dynamics (IBGE, 2018b). Due to the area's ecological sensitivity and complexity, Figure 41 was presented separately. It shows that conservation areas (green zones) were predominantly located away from roads.

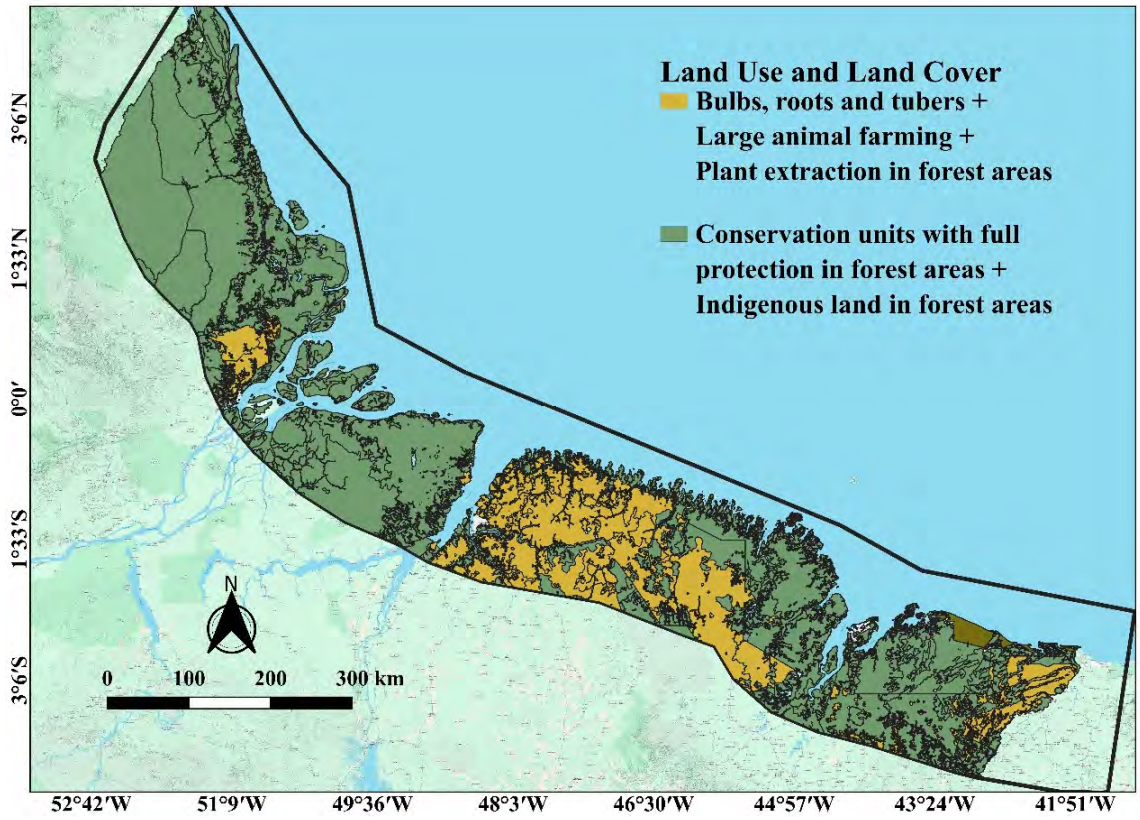


Figure 40. Land use and cover map for the Equatorial Margin area. Scale 1:250,000.

b) Access routes

This study used road network data from IBGE (2018) to establish a 10 km buffer around existing roads, restricting the analysis to pixels within this zone (Figure 41). This buffer was crucial in the Equatorial Margin, where infrastructure limitations could significantly impact accessibility and the feasibility of fieldwork. Figure 41 displayed the access routes, with black lines representing car-accessible roads and green areas indicating the 10 km buffer. By focusing on the sampling strategy within this accessible zone, the approach ensured logistical feasibility and optimized field operations. This method aligned with established soil sampling practices, where accessibility and practicality were essential considerations (CEDDIA et al., 2015; FERREIRA et al., 2022).

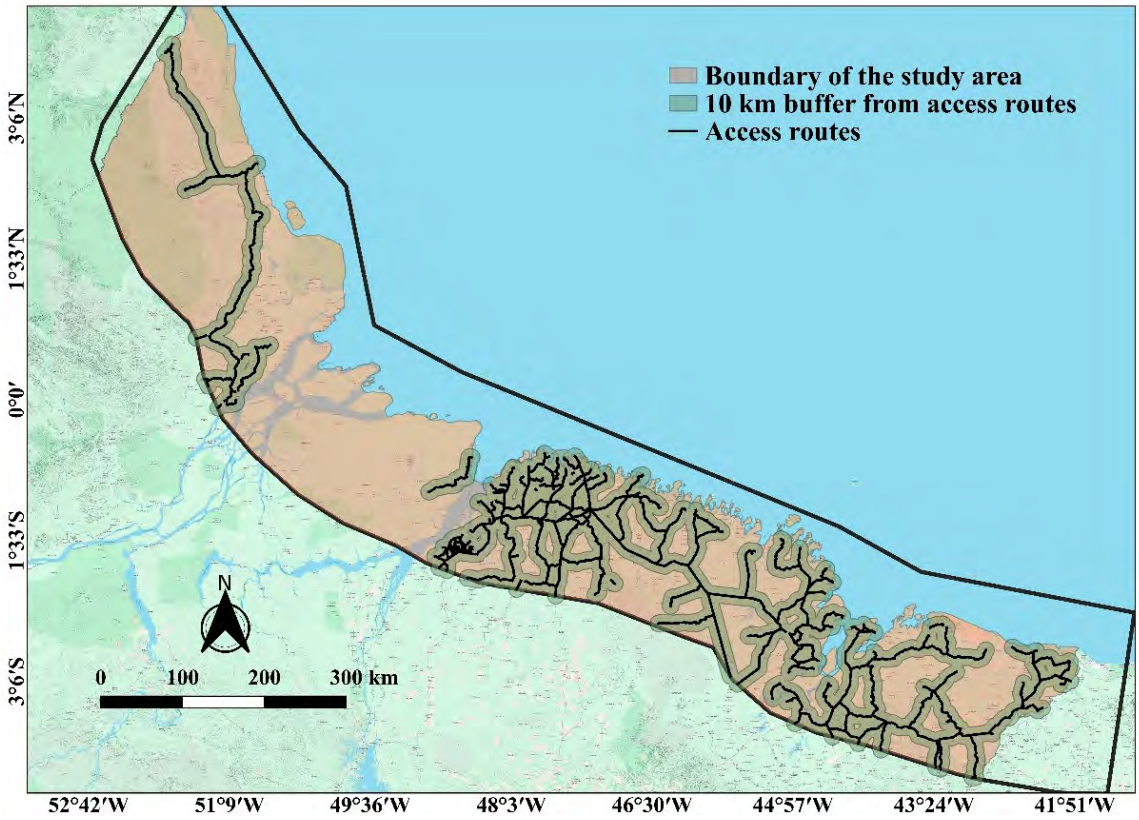


Figure 41. Area of interest with overlapping layers representing the roads (access routes) and a 10 km buffer as margin for walking.

4.4.3 The autoRA rationale

The autoRA algorithm captured the essential variability required for training robust predictive models in spatial analyses. This rationale, as exhibited in Figure 42, elucidated the comprehensive methodology integrated within autoRA, encompassing simulated surface generation, input data processing, calculation of Gower's Dissimilarity Index, delineation of Reference Areas (RAs), sample definition and validation, and a rigorous theoretical framework grounded in information theory and statistical principles. This integrated approach ensured that autoRA identified representative sample sets within highly variable areas and optimized model performance and extrapolation capabilities.

autoRA

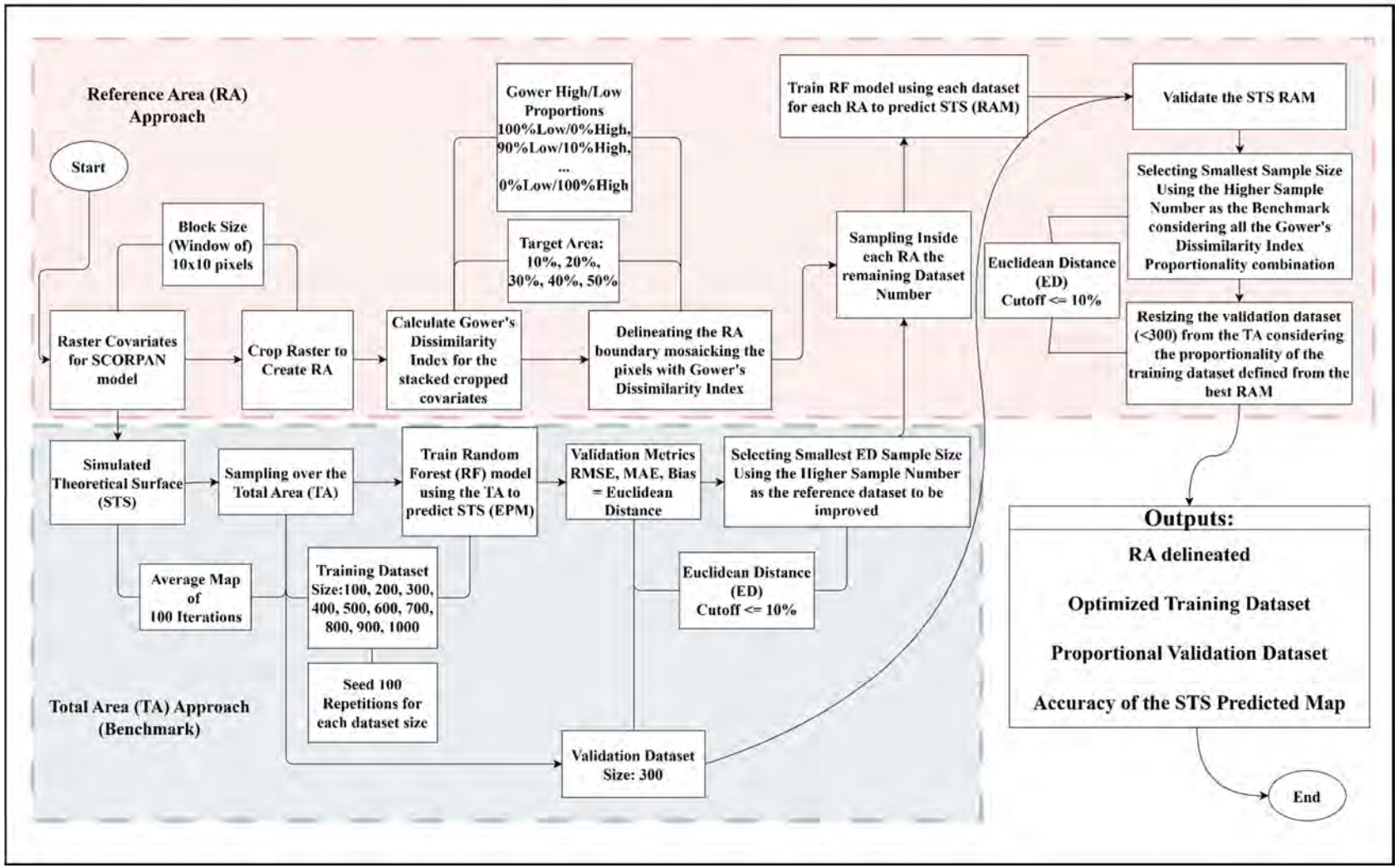


Figure 42. Flowchart of the autoRA rationale.

4.4.4 Simulated theoretical surface

The autoRA methodology's initial step involves generating a Simulated Theoretical Surface (STS), which serves as a model training and validation benchmark. This process utilizes all stacked covariates selected by the user, ensuring that the simulated surface accurately reflects the underlying spatial variability of the study area.

Categorical variables within the covariate set $C = \{C_1, C_2, \ldots, C_K\}$ are transformed into dummy variables through one-hot encoding. Specifically, each categorical variable C_i with m_i unique categories are converted into $m_i - 1$ dummy variables:

$$\mathbf{C_i} \rightarrow \{\mathbf{C_{i1}}, \mathbf{C_{i2}}, \ldots, \mathbf{C_{i(m_i-1)}}\} \quad (1) \tag{1}$$

Where each dummy variable C_{ij} is defined as:

$$\mathbf{C_{ij}} = \begin{cases} \mathbf{1} & \text{if } \mathbf{C_i} = \text{Category } \mathbf{j} \\ \mathbf{0} & \text{otherwise} \end{cases} \quad \forall \mathbf{j} \in \{\mathbf{1}, \mathbf{2}, \ldots, \mathbf{m_i - 1}\} \tag{2}$$

These dummy variables are incorporated into the covariate matrix X, alongside the original numerical covariates $X_{numerical}$, resulting in:

$$\mathbf{X} = \{\mathbf{X}_{\text{numerical}}, \mathbf{X}_{\text{dummy}}\} \tag{3}$$

Each numerical raster layer x_i is normalized to the range $[0,1]$ using:

$$\mathbf{x_i'} = \frac{\mathbf{x_i} - \mathbf{min}(\mathbf{x_i})}{\mathbf{max}(\mathbf{x_i}) - \mathbf{min}(\mathbf{x_i})} \tag{4}$$

The normalized numerical layers and dummy variables are then stacked into a transformed raster T:

$$\mathbf{T} = \{\mathbf{x_1'}, \mathbf{x_2'}, \ldots, \mathbf{d_{ij}}, \mathbf{x_m'}\} \quad (5) \tag{5}$$

Predictors T are randomly shuffled and transformed by applying exponents p selected from $\{1, 2\}$:

$$\mathbf{t_i^{(p_i)}} \quad \text{where} \quad \mathbf{p_i} \in \{\mathbf{1}, \mathbf{2}\} \tag{6}$$

These transformed predictors are then combined using randomly selected operators $\oplus \in \{+, -, \times\}$:

$$\mathbf{R} = \mathbf{t_1^{(p_1)}} \oplus_1 \mathbf{t_2^{(p_2)}} \oplus_2 \cdots \oplus_{k-1} \mathbf{t_{k-1}^{(p_k)}} \tag{7}$$

The raw response surface R is subsequently scaled to the range $[0,100]$:

$$\mathbf{R'} = \frac{\mathbf{R} - \mathbf{min(R)}}{\mathbf{max(R)} - \mathbf{min(R)}} \times \mathbf{100} \quad (8)$$

The STS is established by averaging 100 such iterations:

$$\text{STS} = \frac{\mathbf{1}}{\mathbf{100}} \sum_{\mathbf{k=1}}^{\mathbf{100}} \mathbf{M_k} \quad (9)$$

This ensemble approach ensures that the STS robustly represents the spatial variability inherent in the covariate stack.

4.4.5 Input data processing and Gower's dissimilarity index calculation

Following the STS generation, autoRA processed the input data, which consisted of raster-formatted covariates representing essential predictors for the area of interest and soil formation processes. The updated version of autoRA allowed the insertion of road shapes, enabling the calculation of Gower's Dissimilarity Index within a user-defined buffer mask.

A block size of 10×10 pixels were applied to the raster covariates $C_{SCORPAN} = \{X_1, X_2, \ldots, X_n\}$ to aggregate the data:

$$\mathbf{R}_{\text{block}} = \text{window}(\mathbf{C}_{\text{SCORPAN}}, \mathbf{10 \times 10}) \quad (10) \quad (10)$$

Gower's Dissimilarity Index $D_{Gower}(x, y)$ was calculated for each aggregated block to quantify the dissimilarity between pixel pairs and accommodated both numerical and categorical variables:

$$\mathbf{D}_{\text{Gower}}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_i} \cdot \mathbf{d_i(x, y)}}{\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{w_i}} \quad (11)$$

Where:

$$\mathbf{d_i(x, y)} = \begin{cases} \mid \mathbf{x_i} - \mathbf{y_i} \mid & \text{for numerical variables} \\ \begin{cases} \mathbf{0} & \text{if } \mathbf{x_i} = \mathbf{y_i} \\ \mathbf{1} & \text{if } \mathbf{x_i} \neq \mathbf{y_i} \end{cases} & \text{for categorical variables} \end{cases} \quad (12)$$

Here, w_i denoted the weight of covariate i, and x_i, y_i were the values of the covariate i for samples x and y, respectively.

4.4.6 Reference area delineation

After calculating Gower's Dissimilarity Index, autoRA delineated RAs by systematically grouping pixels based on their dissimilarity values to form candidate RAs that satisfied specific target area proportions. The delineation process integrated a rigorous theoretical framework grounded in information theory and statistical learning principles to maximize entropy and variability within the RA, thereby enhancing the model's extrapolation capabilities.

a) Entropy maximization and variability representation

To underpin the RA selection process, we presented a theoretical framework demonstrating how autoRA aligned with information theory principles, particularly entropy maximization, to ensure high variability essential for practical extrapolation in predictive modeling.

b) Theorem: Optimal reference area and sample size selection in autoRA

We considered a spatial domain $\mathcal{A} = \{x_1, x_2, \dots, x_N\}$ that comprised N pixels, each characterized by a set of covariates $C = \{C_1, C_2, \dots, C_n\}$. We defined Gower's Dissimilarity Index $D_{Gower}(x_i, x_j)$ for any pair of pixels $x_i, x_j \in \mathcal{A}$. We let $T_p \in (0,1)$ denote the target area proportion, and $p_H \in [0,1]$ represent the proportion of high dissimilarity pixels within the RA.

The autoRA algorithm delineated a candidate $RA_{candidate}(T_p, p_H)$ by selecting the top $N_H = \lfloor p_H \times N_{RA} \rfloor$ pixels with the highest D_{Gower} values and the subsequent $N_L = N_{RA} - N_H$ pixels with lower D_{Gower} values, where $N_{RA} = \lfloor T_p \times N \rfloor$.

For each candidate RA, the algorithm varied the sample size n to determine the minimal number of samples n^* that satisfied a predefined Euclidean Distance (ED) threshold δ.

Formally, for each $RA_{candidate}(T_p, p_H)$, we defined:

$$\mathbf{S_{RA}(T_p, p_H, n) = \{x_i \in RA_{candidate}(T_p, p_H) \mid i \leq n\}} \quad (13)$$

We trained a Random Forest (RF) model using $S_{RA}(T_p, p_H, n)$ and evaluated its performance on an external validation dataset V, which yielded normalized metrics $RMSE_{norm}$, $Bias_{norm}$, and R^2_{norm}. We computed the Euclidean Distance (ED) as:

$$\mathbf{ED(T_p, p_H, n) = \sqrt{(RMSE_{norm})^2 + (Bias_{norm})^2 + (R^2_{norm})^2}} \quad (14)$$

Since the objective is to determine the minimal sample size n^* such that the error remains within a predefined threshold δ, we established the following conditions:

$$\mathbf{ED(T_p, p_H, n^*) \leq \delta} \quad (15)$$

Where δis the predefined ED threshold (e.g., $\delta = 0.2$). This constraint guarantees that the selected RA candidate minimizes the error, ensuring that the sample size is as small as possible while maintaining model performance within an acceptable range.

We now prove the existence of an optimal Reference Area RA_{best} and corresponding sample size n_{best}.The function $ED(T_p, p_H, n^*)$ is well-defined for all feasible values of n and, by construction, is bounded below by zero. The minimization criterion follows from:

$$\mathbf{RA_{best}} = \mathbf{arg}\min_{(\mathbf{T_p},\mathbf{p_H})} \mathbf{ED}(\mathbf{T_p}, \mathbf{p_H}, \mathbf{n}^*) \tag{16}$$

$$\mathbf{n_{best}} = \mathbf{n}^* \quad \mathbf{for\ RA_{best}} \tag{17}$$

Since $ED(T_p, p_H, n^*)$ is defined over a finite set of candidate RAs and is continuous with respect to n in a discrete domain, there exists at least one optimal pair (T_p, p_H) that minimizes $ED(T_p, p_H, n^*)$. This follows directly from the Extreme Value Theorem (ADAMS; FOURNIER, 2003), which ensures that a minimum must exist for any continuous function over a compact domain.

c) Maximizing entropy within reference area

In information theory, entropy H quantified uncertainty or variability within a dataset. For a discrete random variable X with probability mass function $p(x)$, entropy was defined as:

$$\mathbf{H(X)} = -\sum_{\mathbf{x}} \mathbf{p}(\mathbf{x})\mathbf{log}\mathbf{p}(\mathbf{x}) \tag{18}$$

We considered the distribution of Gower's Dissimilarity Index$D_{Gower}(x_i, x_j)$ within the RA. By selecting the top N_H pixels with the highest dissimilarity and the subsequent N_L pixels with lower dissimilarity, autoRA effectively maximized the entropy of the selected sample set S_{RA}.

Let S_{RA} be the sample set within the RA. The entropy of S_{RA} concerning D_{Gower} is:

$$\mathbf{H}(\mathbf{D_{Gower}}) = -\sum_{\mathbf{d}} \mathbf{p}(\mathbf{d})\mathbf{log}\mathbf{p}(\mathbf{d}) \tag{19}$$

Where $p(d)$ was the probability of mass function of dissimilarity values d within the RA.

By selecting pixels that spanned the entire range of D_{Gower} values autoRA ensured a high entropy $H(D_{Gower})$, indicating maximal variability and information content within the sample set.

According to statistical learning theory, a training set that captures high variability within the feature space leads to models with lower generalization errors. Specifically, maximizing the diversity of training samples reduces the model's variance, enhancing its ability to generalize to unseen data.

By maximizing entropy $H(D_{Gower})$, autoRA ensured that the selected sample set S_{RA} encompassed many feature combinations, minimizing overfitting and reducing generalization errors.

The generalization error E_{gen} could be bounded by the empirical error E_{emp} and the complexity of the hypothesis space $\mathcal{H}$:

$$\mathbf{E_{gen} \leq E_{emp} + \sqrt{\frac{C \cdot H(D_{Gower})}{n^*}}} \quad (20)$$

Where C was a constant, $H(D_{Gower})$ was the entropy of the sample set, and n^* is the sample size. Maximizing $H(D_{Gower})$ while minimizing n^* ensured that E_{gen} was minimized.

d) Defining the reference area

The reference area is thus a candidate subset of the total area, defined by specific proportions of high and low Gower's Dissimilarity pixels. For each target area ratio $T_p \in \{0.1,0.2,0.3,0.4,0.5\}$, the algorithm iterates through the defined (p_H, p_H) combinations (with $p_L = 1 - p_H$) to create multiple candidate RAs:

$$\mathbf{RA_{candidate}(T_p, p_H) = \bigcup_{i=1}^{N_H} x_i \cup \bigcup_{j=N_H+1}^{N_{RA}} x_j} \quad (21)$$

Where $N_{RA} = \lfloor T_p * N \rceil$ is the total number of pixels in the RA, $N_H = \lfloor p_H * N_{RA} \rfloor$, is the number of high-dissimilarity pixels, and the pixels x_i are ordered in descending order by their D_{Gower} values.

e) Optimal sample size determination

The ED metric $ED(T_p, p_H, n)$ integrated multiple performance indicators into a single measure of prediction error. By setting a threshold δ, autoRA ensured that the sample size n^* was sufficient to achieve the desired level of model accuracy.

The selection of n^* was analogous to determining the sample size required to estimate population parameters within a specified confidence interval and margin of error. By iteratively increasing n until $ED(T_p, p_H, n^*) \leq \delta$, autoRA ensured that the model had sufficient data to capture the underlying spatial variability while avoiding unnecessary sampling.

The selection of the optimal RA, denoted as RA_{best}, involved evaluating all possible combinations of T_p and p_H. The goal was to minimize the ED metric across these combinations, effectively searching for the RA configuration that offered the best trade-off between sample size and model accuracy.

This selection process can be formulated as a mathematical optimization problem. The objective function seeks to minimize $ED(T_p, p_H, n^*)$, subject to the constraint $ED(T_p, p_H, n^*) \leq \delta$.

In this formulation, $T_p \in (0,1)$ represents the target area proportion, while $p_H \in [0,1]$ denotes the proportion of high-dissimilarity pixels within the RA. The function $ED(T_p, p_H, n)$ quantifies the total error of the predictive model trained on an RA candidate, and the threshold δ defines the maximum allowable error to maintain an acceptable level of model accuracy.

Since the optimization problem involves a finite number of possible configurations for T_p and p_H, the solution is obtained through a discrete search over a constrained space. The existence of an optimal solution is guaranteed because $ED(T_p, p_H, n)$ is well-defined and the search space is

bounded. By solving this optimization problem, autoRA identifies the RA configuration that maximizes entropy while ensuring that the model performance remains within the acceptable threshold δ, providing a structured and rigorous approach to selecting the most effective RA.

f) Entropy maximization and variability representation

In this section, we delved deeper into the statistical underpinnings of the autoRA algorithm, primarily focusing on how entropy maximization and Gower's Dissimilarity Index contributed to capturing spatial variability essential for robust predictive modeling.

Entropy H quantified uncertainty or variability within a dataset. In the context of spatial data, higher entropy implied more significant variability among the covariates, which was crucial for training models that generalized well.

For a discrete distribution of Gower's Dissimilarity Index values within the RA, entropy was defined as:

$$\mathbf{H(D_{Gower}) = -\sum_{d} p\,(d) log p(d)} \tag{22}$$

Where $p(d)$ was the probability mass function of dissimilarity values d within the RA. The autoRA algorithm aimed to maximize $H(D_{Gower})$ by selecting pixels that spanned the entire range of D_{Gower} values. This ensured that the training sample captured the full spectrum of spatial variability, enhancing the model's ability to extrapolate to unseen areas.

$$\mathbf{\max_{RA} H(D_{Gower}) \quad subject\ to \quad |\ RA\ | = N_{RA}} \tag{23}$$

By solving this optimization problem, autoRA selected the RA that offered maximal variability quantified by entropy.

g) Gower's Dissimilarity Index

Gower's Dissimilarity Index $D_{Gower}(x_i, x_j)$ was a metric that quantified the dissimilarity between two pixels based on a combination of numerical and categorical covariates. It was defined as:

$$\mathbf{D_{Gower}(x_i, x_j) = \frac{\sum_{k=1}^{n} w_k\, d_k(x_i, x_j)}{\sum_{k=1}^{n} w_k}} \tag{24}$$

Where:

w_k was the weight assigned to covariate C_k.

$d_k(x_i, x_j)$ was the dissimilarity for covariate C_k, defined as:

$$\mathbf{d_k(x_i, x_j)} = \begin{cases} \dfrac{|\, \mathbf{C_k(x_i) - C_k(x_j)} \,|}{\mathbf{range(C_k)}} & \textbf{if } \mathbf{C_k} \textbf{ was numerical} \\ \begin{cases} \mathbf{0} & \textbf{if } \mathbf{C_k(x_i) = C_k(x_j)} \\ \mathbf{1} & \textbf{otherwise} \end{cases} & \textbf{if } \mathbf{C_k} \textbf{ was categorical} \end{cases} \quad (25)$$

By computing $D_{Gower}(x_i, x_j)$ for all pixel pairs, autoRA quantified the dissimilarity across the spatial domain. High D_{Gower} values indicated regions of high variability, while low values signified homogeneity. This metric allowed autoRA to systematically select samples that captured the essential spatial variability necessary for accurate model training and extrapolation.

Selecting pixels with a wide range of D_{Gower} values ensure that the training set S_{RA} encompasses diverse covariate profiles, thereby maximizing the entropy $H(D_{Gower})$ Moreover, reducing the model's reliance on homogenous samples.

Predictive models often face challenges when extrapolating to areas with covariate values not well-represented in the training data. High variability within the training set mitigated this issue by providing the model with a comprehensive understanding of the covariate space.

From a statistical standpoint, a training set with high entropy $H(D_{Gower})$ ensured that the model was exposed to a wide range of covariate combinations, reducing the likelihood of encountering unseen scenarios during prediction.

The generalization error E_{gen} of a model was inversely related to the entropy of the training set:

$$\mathbf{E_{gen}} \propto \frac{\mathbf{1}}{\mathbf{H(D_{Gower})}} \quad \textbf{(26)} \quad (26)$$

Thus, maximizing $H(D_{Gower})$ directly contributed to minimizing E_{gen}, enhancing the model's extrapolation capabilities.

4.4.7 Model training and validation

With the benchmark sample size n^* established from the RA selection process, the RA approach defines and validates samples within each candidate Reference Area.

For each candidate $RA_{candidate}(T_p, p_H)$, an equal number of soil samples n^* are distributed using the Total Area (TA) sampling approach:

$$\mathbf{S_{RA}(T_p, p_H)} = \{\mathbf{x_i} \in RA_{candidate}\mathbf{(T_p, p_H)} \mid \mathbf{i \leq n^*}\} \quad (27)$$

For each candidate RA, a Random Forest (RF) model is trained using the RA-specific dataset $X_{train,\, RA}$ and $y_{train,\, RA}$:

$$RF_{RA}\mathbf{(T_p, p_H)} = \mathbf{argmin_{\theta} L(\hat{y}, y)} \quad (28)$$

Each RF model trained on a candidate RA is validated using the same external and fixed validation dataset V, exclusively sampled from the TA. The validation metrics (RMSE, Bias, R^2) are normalized and combined into the Euclidean Distance (ED):

$$\mathrm{ED}_{\mathrm{RA}}\left(\mathbf{T_p}, \mathbf{p_H}\right) = \sqrt{\left(\mathrm{RMSE}_{\mathrm{ext,\ norm}}\right)^2 + \left(\mathrm{Bias}_{\mathrm{ext,\ norm}}\right)^2 + \left(\mathbf{R}^2_{\mathrm{ext,\ norm}}\right)^2} \tag{29}$$

4.4.8 Selecting the best reference area

The optimal RA is determined by selecting the candidate RA with the smallest $\mathrm{ED}_{\mathrm{RA}}\left(\mathrm{T_p}, \mathrm{p_H}\right)$:

$$\mathrm{RA}_{\mathrm{best}} = \mathbf{arg}\,\mathbf{min}_{\left(\mathbf{T_p}, \mathbf{p_H}\right)} \mathrm{ED}_{\mathrm{RA}}\left(\mathbf{T_p}, \mathbf{p_H}\right) \tag{30}$$

This ensures that the selected RA meets the target area ratio size and utilizes the minimum sample size necessary to achieve optimal model performance.

4.4.9 Cost evaluation

The cost evaluation component of the autoRA methodology quantifies the financial implications associated with sample collection and model validation. This section delineates the mathematical framework for calculating the total cost based on sample size proportions and logistical factors.

4.4.10 Proportional allocation of training and validation samples

The proportion of samples allocated to training and validation datasets is defined by:

$$\mathbf{p} = \{\mathbf{p}_{\mathrm{train}}, \mathbf{p}_{\mathrm{valid}}\} \quad \text{where} \quad \mathbf{p}_{\mathrm{train}} + \mathbf{p}_{\mathrm{valid}} = \mathbf{1} \tag{31}$$

In this study, the standard approach in digital soil mapping was employed with $\mathrm{p}_{\mathrm{train}} = 0.7$ and $\mathrm{p}_{\mathrm{valid}} = 0.3$. However, the methodology allows for other proportional allocations, such as 50% training and 50% validation or 60% training and 40% validation.

4.4.11 Calculation of road length within reference areas

The total length of roads within each RA is computed to account for logistical costs associated with sample collection:

$$\mathbf{L}_{\mathrm{train}} = \mathrm{compute\_road\_length}(\mathrm{roads}, \mathrm{RA}_{\mathrm{train}}) \tag{32}$$

Where $\mathrm{L}_{\mathrm{train}}$ is the road length within the training area. roads represents the road network data. $\mathrm{RA}_{\mathrm{train}}$ is the delineated Reference Area allocated for training.

4.4.12 Determination of observation counts and associated costs

The number of observations for training and validation datasets is calculated based on the sample size n^* and the defined proportions p:

$$\mathbf{O}_{\text{train}} = \mathbf{n}^* \times \mathbf{p}_{\text{train}} \tag{33}$$

$$\mathbf{O}_{\mathbf{valid}} = \mathbf{n}^* \times \mathbf{p}_{\mathbf{valid}} \tag{34}$$

The total number of observations is the sum of training and validation observations:

$$\mathbf{O}_{\text{total}} = \mathbf{O}_{\text{train}} + \mathbf{O}_{\text{valid}} \tag{33}$$

Given a cost $\mathbf{C}$ per observation, the training, validation, and total costs are computed as follows:

$$\mathbf{C}_{\text{train}} = \mathbf{O}_{\text{train}} * \mathbf{C} \tag{36}$$

$$\mathbf{C}_{\text{valid}} = \mathbf{O}_{\mathbf{valid}} * \mathbf{C} \tag{37}$$

$$\mathbf{C}_{\mathbf{total}} = \mathbf{C}_{\mathbf{train}} + \mathbf{C}_{\mathbf{valid}} \tag{38}$$

In this study, a fixed cost of $1000 per observation was stipulated:

$$\mathbf{C} = \$\mathbf{1000} \tag{39}$$

Thus, the total cost associated with sampling and validation is:

$$\mathbf{C}_{\text{total}} = (\mathbf{n}^* \times \mathbf{p}_{\text{train}} \times \mathbf{1000}) + (\mathbf{n}^* \times \mathbf{p}_{\text{valid}} \times \mathbf{1000}) \tag{40}$$

4.4.13 autoRA shiny application

The autoRA function optimizes soil sampling strategies by delineating an optimal RA and determining the best sample size. It is written in R and relies on several libraries (such as terra, randomForest, Metrics, clhs, gower, smoothr, progress, and shiny) to process spatial data, generate simulated surfaces, and compute model performance metrics.

The function's arguments allow the user to configure various aspects of the analysis. In the autoRA Shiny app, these arguments are populated through user inputs (Figure 42). The following is a detailed explanation of each argument and how it is linked to the interface:

1. r (Raster Stack of Predictors) – This primary raster object contains all predictor variables used in the analysis. Users upload one or more raster files (formats such as .tif, .grd, or .nc) via the “Upload Raster Stack (Predictors)” file input. After doing so, the user is required to declare which raster is type “categorical". In this way, maps such as Pedology, Geology, Land Use and Land Cover, and Geomorphology, for instance, will be pre-treated as dummies.

2. Input Target Map (Optional) – The user can input a target map to serve as a benchmark. It will be sampled and mapped using the total area and the RA approach. If it is used, the simulated theoretical surface is not calculated.

3. roads (Optional Roads Shapefile) – An optional spatial dataset representing roads. It creates a buffer that may affect the sampling and validation processes. Users upload the necessary files for a roads shapefile (including .shp, .dbf, .sbn, .sbx, .shx, and .prj) via the “Upload Roads Shapefile (Optional)” file input.

4. buffer_size (Buffer Size in Meters) – A numeric value (default 1000) that sets the width of the buffer (in meters) around the roads. This value is specified using a numeric input labeled “Buffer Size (meters):” with a default value of 1000.

5. block_size (Vector of Block Sizes) – A vector of candidate block sizes (e.g., 5, 10, 20, 30, 40, 50) is used when partitioning the study area into blocks. Block sizes are chosen via a checkbox group input, where the user can select from the provided options; by default, 10 is selected.

6. target_area (Vector of Target Area Percentages) – A vector defining candidate percentages of the total area to be used as the reference area (e.g., 10%, 20%, 30%, 40%, 50%). Users select target area percentages using a checkbox group input. All percentages (10, 20, 30, 40, 50) are selected by default.

7. size_blocks_to_remove – An optional argument (default is NULL) that can be used to specify block sizes to remove from consideration. This parameter is not directly exposed in the UI and uses its default value unless modified in code.

8. sampling_pattern (Sampling Pattern) – A character string specifying the sampling pattern (options include “random”, “regular”, “stratified”, “nonaligned”, “hexagonal”, “clustered”, “Fibonacci”, “clhs”). The default is “clhs” (Conditional Latin Hypercube Sampling). The sampling pattern is chosen using a select input, with “clhs” set as the default option.

9. sample_sizes (Vector of Sample Sizes) – A numeric vector of candidate sample sizes to test (the function’s default is a sequence from 100 to 500 in steps of 100). In the app, users can provide a more flexible set of values. This is entered as a comma-separated string in a text input labeled “Sample Sizes (comma-separated):” (e.g., "100,200,300,400,500,1000,10000").

10. cutoff (Percentage Improvement Cutoff) – A numeric threshold (default 10) that defines the maximum acceptable percentage difference in the Euclidean Distance between configurations when selecting the optimal sample size. Specified via a numeric input labeled “Percentage Improvement Cutoff (%)” with a default value of 10.

11. seed.rep (Number of Repetitions for Sampling) – The number of times (default 10 in the function, though the UI may pass 100) the sampling process is repeated to average performance metrics and ensure stability. The user enters this number through a numeric input labeled “Number of Repetitions for Sampling:”.

12. sim_surface_rep (Number of Simulated Surface Repetitions) – The number of simulated surfaces (default 3) generated when no user-provided surface is available. These repetitions are averaged to create a robust response surface. Provided using a numeric input labeled “Number of Simulated Surface Repetitions:” with a default value of 3.

13. price_per_observation (Price per Observation) – A numeric value (default 1000) representing the cost of each observation is used in calculating the overall cost of the sampling strategy. Entered via a numeric input labeled “Price per Observation:” with a default of 1000.

14. proportion_RA_train_validation_sampling (Training/Validation Proportion) – A vector with named elements (default c(train = 70, valid = 30)) sets the proportion of the total sample allocated to training versus validation. The training proportion is set through a numeric input labeled “Training Proportion (%)” (default 70). The validation proportion is implicitly the complement (30%).

15. sample_validation_with_buffer (Validation with Buffer Flag) – A boolean (default FALSE) indicating whether the validation sample should be restricted to areas within the road buffer. Controlled by a checkbox input labeled “Sample Validation with Buffer” (unchecked by default).

16. num_validation_datasets (Number of Validation Points) – A fixed number (default 300) indicating how many points are used for the initial validation step. Set using a numeric input labeled “Number of Validation Points for Step 1:” with a default value of 300.

17. gower_high_low (Gower’s Dissimilarity Proportions) – A list specifying the proportions for high Gower values (default is a sequence from 0 to 100 in 10% increments) and, if not provided, calculates the corresponding low proportions as the complement to 100. Users select these values using a picker input labeled “Gower High Proportions (%)” that allows multiple selections. All options (0%, 10%, …, 100%) are selected by default.

18. user_simulated_surface (Optional User-Provided Simulated Surface) – A single-layer raster that the user can provide as a simulated response surface. AutoRA uses this instead of generating a new simulated surface if it is supplied. Uploaded through a file input labeled “Upload User Simulated Surface (Optional)” that accepts file types such as .tif, .grd, or .nc.

The autoRA Shiny application (Figure 43) divides the user interface into a header, sidebar, and main panel. The sidebar contains all the input controls that gather the parameters described above. For example:

In Figure 44, we present some interesting results. A - Custom Plot Selection Panel: This panel allows users to customize the visualization of reference area results. Users can select block sizes, target area percentages, and Gower high proportions from a dropdown menu. Additionally,

checkboxes enable the display of training and validation points, roads, and the best reference area on the map. B - Reference Map Tab: This section displays the Reference Map, highlighting the selected reference area for soil sampling based on the optimized parameters. The color gradient represents different levels of environmental heterogeneity as captured by the autoRA algorithm. C - Metric Plots Tab: This section contains multiple plots summarizing key performance metrics, such as R^2, RMSE, Bias, and Euclidean Distance for different sample sizes. These plots allow users to analyze the impact of varying sample sizes and target areas on the accuracy and cost-effectiveness of the reference area delineation.

D - Results Table Tab: This tab contains all computed results from the autoRA analysis. The table includes metrics for different configurations, such as block size, target area, sample size, model accuracy (R^2, RMSE, Bias), and total cost. This structured data allows users to compare different parameter settings and determine the optimal configuration. E - Simulated Surfaces Tab: This section visualizes the Simulated Theoretical Surfaces (STS) used in the sensitivity analysis. The simulated surfaces allow for the evaluation of different environmental conditions and their effects on the accuracy of the reference area delineation. The maps shown here are averaged representations of the simulated surfaces generated during the process. F - Gower Map for Best Combination: This tab presents Gower's Dissimilarity Index Map, showing the spatial variability of the selected reference area. The color variations represent different degrees of dissimilarity across the study area. This information is crucial for identifying heterogeneous zones that contribute to robust reference area selection.

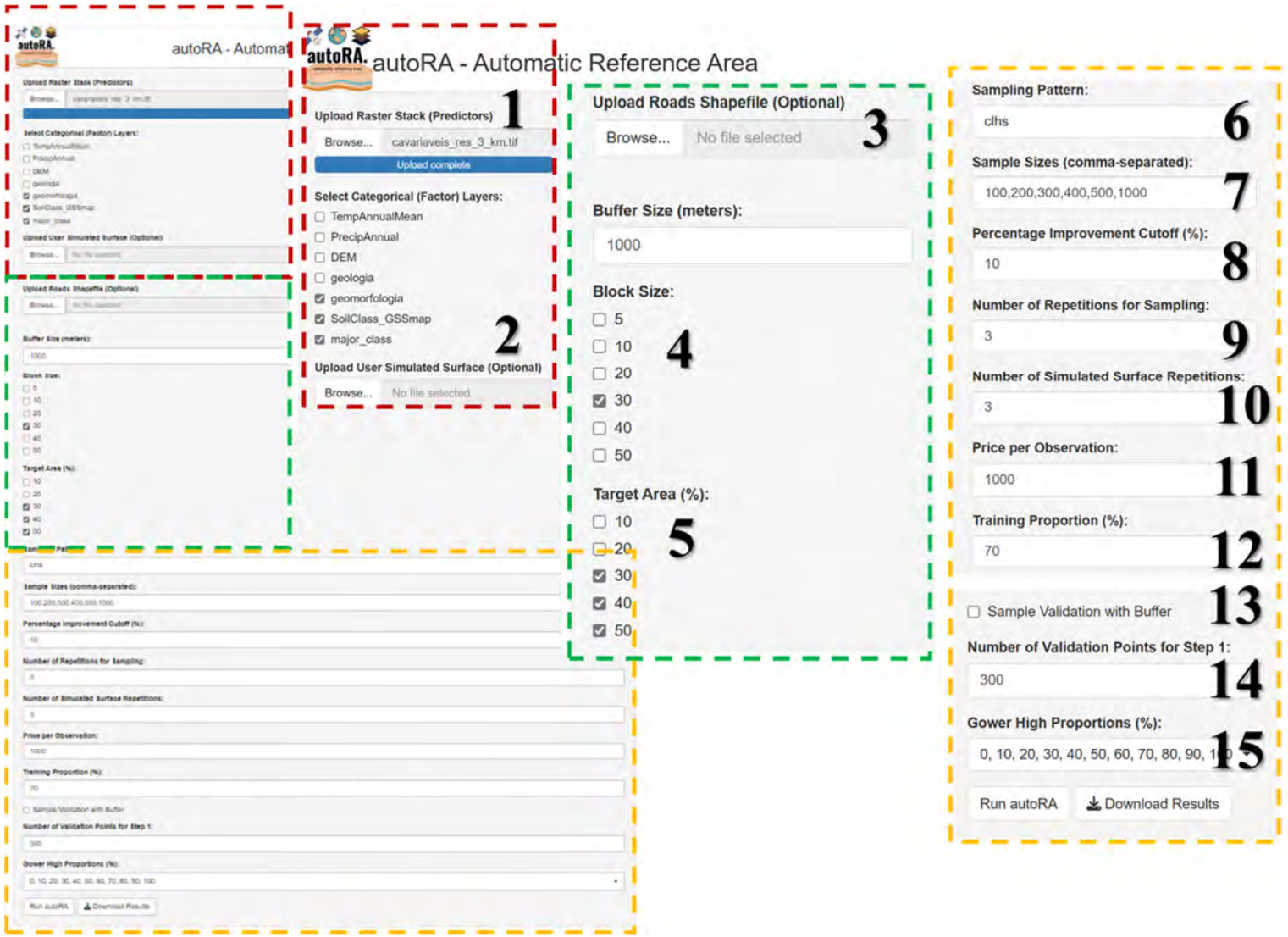


Figure 373. Scheme of the autoRA application.

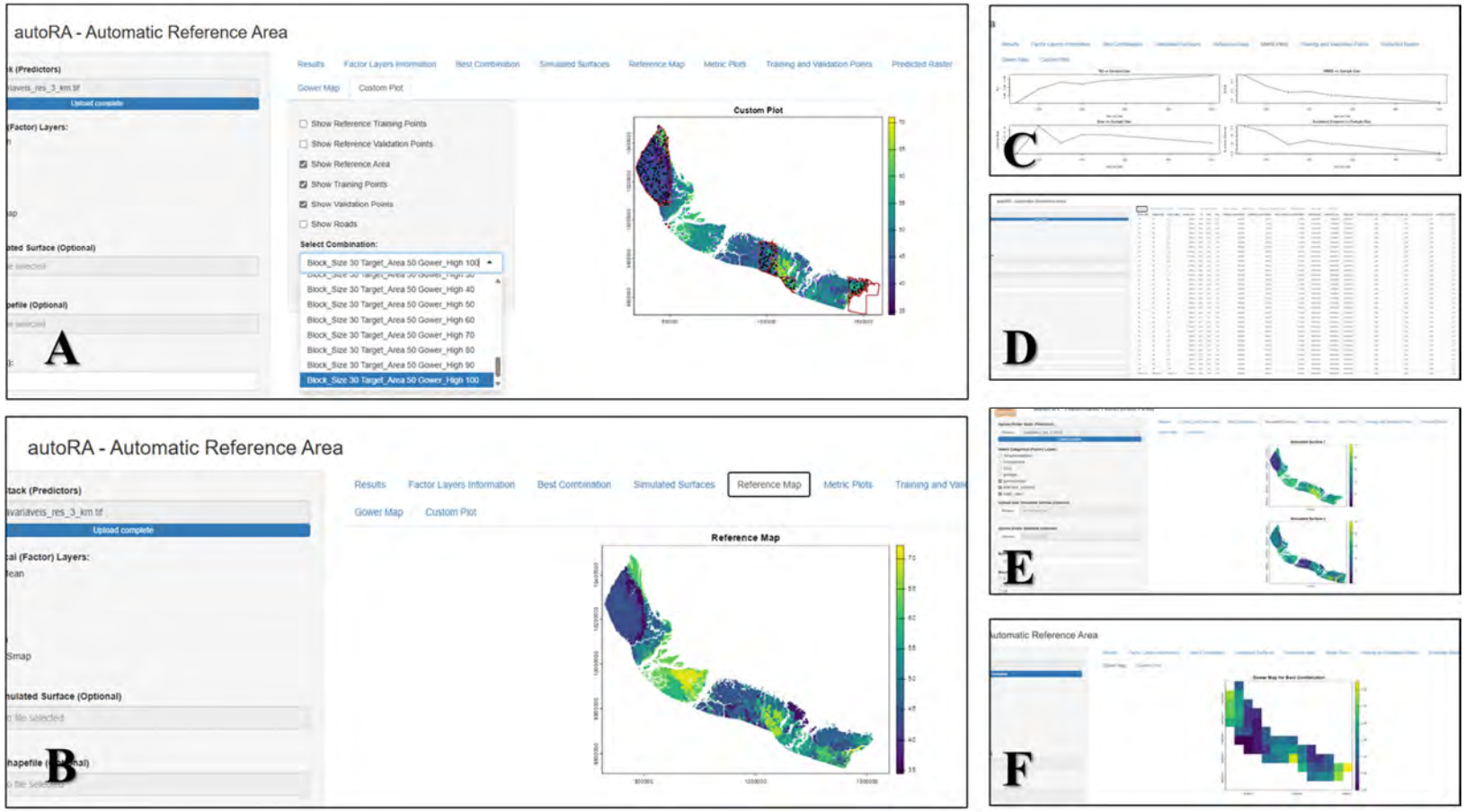


Figure 384. Screenshots of the autoRA application. A., custom layout to overlap the outputs, as the predicted STS, the RA by all combinations tested; B, output of the Reference Map, considered the benchmark map as average of the 100 STS; C. metrics tested RMSE, MAE, BIAS and Euclidean Distance using the parameters tested' D. data sheet with the metrics using the external validation dataset; E. Example of 2 out 100 STS generated; F, Gower's Index Dissimilarity map.

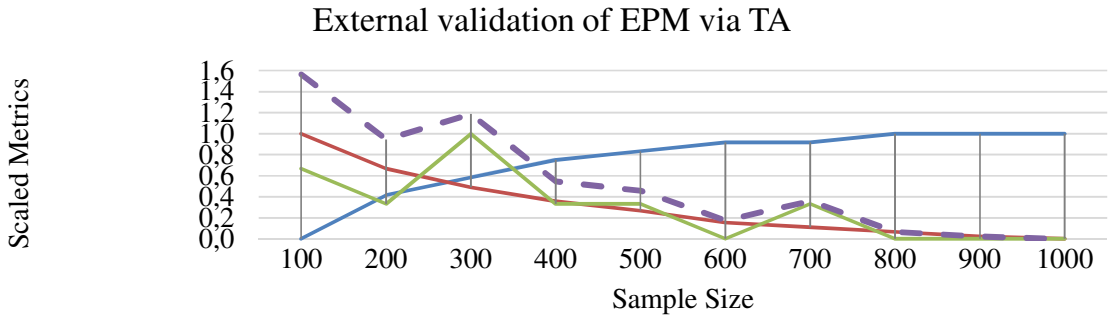
4.5 RESULTS AND DISCUSSION

The TA approach was employed based on the predefined ED threshold of 20% to determine the optimal sample size required for representing the optimized sampling size when considering the exhaustive predictive model (EPM). Various sample sizes were evaluated using RF models while maintaining a consistent external validation set of 300 samples, as defined in Equation (1) of the theoretical framework. The ED metric, derived from normalized RMSE, Bias, and R^2 values, was calculated for each sample size to assess accuracy and reliability.

The analysis identified sample sizes of n*=800 and n*=900 as the small configurations that kept the ED below the 20% cutoff compared to the benchmark n*=1000 (Figure 45). At n=800, the ED was calculated to be ED=0.07, well within the acceptable range. This sample size achieved exceptional performance metrics, including a perfect coefficient of determination (R^2=1.00), alongside minimal RMSE and Bias values, all at a total cost of $1,100,000. Increasing the sample size to n=900 reduced the ED to 0.02, comfortably within the 20% threshold, while maintaining R^2=1.00 with low RMSE and Bias. However, this improvement in ED came at an increased total cost of $1,200,000.

A sample size of n=800 was recommended for its superior cost-efficiency and excellent model performance. This configuration struck an optimal balance between keeping the ED below the 20% threshold and minimizing financial expenditure when considering the TA approach.

Given the establishment of $n^* = 800$ as the optimal sample size through the TA approach, the next phase of the autoRA methodology proceeded by leveraging this benchmark to refine further and validate the sampling strategy. According to the theoretical framework, the autoRA algorithm evaluated each remaining sample size listed by modeling 100 distinct RF models. For each model, predictions were made using the same external validation dataset of 300 samples, ensuring consistency and comparability across evaluations.



	100	200	300	400	500	600	700	800	900	1000
R2 scaled	0,00	0,42	0,58	0,75	0,83	0,92	0,92	1,00	1,00	1,00
RMSE scaled	1,00	0,67	0,49	0,36	0,27	0,16	0,11	0,07	0,02	0,00
Bias scaled	0,67	0,33	1,00	0,33	0,33	0,00	0,33	0,00	0,00	0,00
ED	1,56	0,95	1,19	0,55	0,46	0,18	0,36	0,07	0,02	0,00

Figure 395. Graph of metrics evaluated for adjusting 10 prediction models from each training set with different sizes to define the smaller set of points representing the exhaustive model and reference.

The STS was derived by averaging the results of 100 repetitions of algebraic iterations, capturing the interactions among all covariates. Figure 46 presented the final average STS map, which served as the target surface. Additionally, three example STS outputs from individual iterations were shown, illustrating the variability in spatial patterns caused by differences in the sequence of covariate interactions and applied mathematical operators during each iteration.

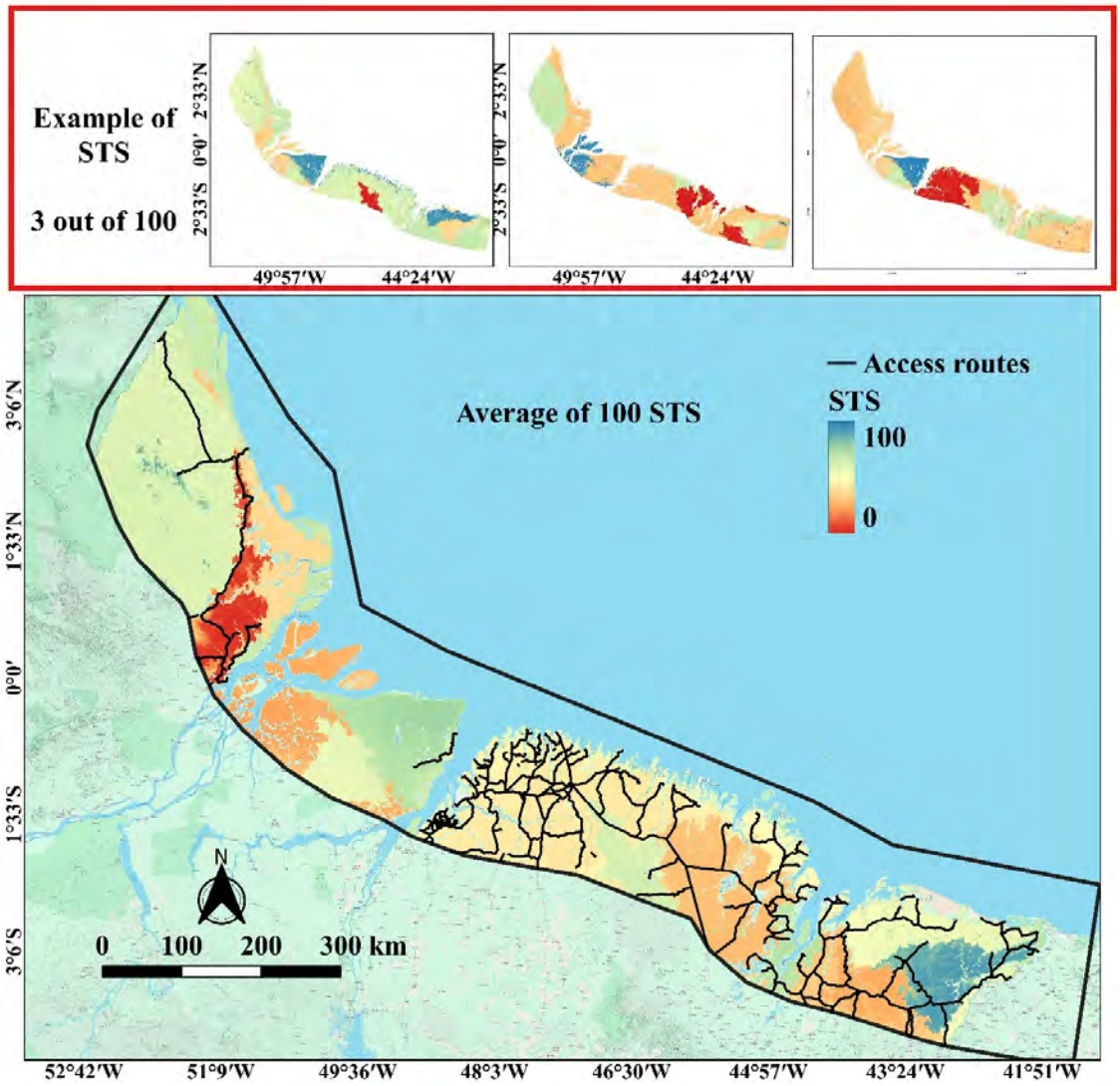


Figure 46. Simulated surface from the algebraic average of 100 simulated surfaces. A demonstrative scheme on the top line and map is considered an example without error to be predicted.

The EPM for the STS was developed using 800 training sample points and 300 external validation points (Figure 46), with covariates serving as predictors. This model demonstrated robust performance, achieving an R^2 of 0.97, an RMSE of 0.34, and a Bias of 0.02. Furthermore, it maintained an impressive ED of 0.01 relative to the benchmark, underscoring its accuracy and reliability.

The spatial distribution of predicted values, ranging from 44 to 53, effectively highlighted the regional variability captured by the model. As illustrated in Figure 47, the EPM STS map accurately represented the spatial patterns seen in the target, with black dots indicating training data points, red dots denoting external validation data, and access routes providing essential spatial context.

The cost evaluation for the mapping process using the EPM, as depicted in Figure 47, revealed significant savings by utilizing 800 samples instead of the initially considered 1,000. Specifically, the total cost was reduced from $1,300,000 (calculated as 1,000 training samples plus 300 validation samples at $1,000 each) to $1,100,000, resulting in savings of $200,000.

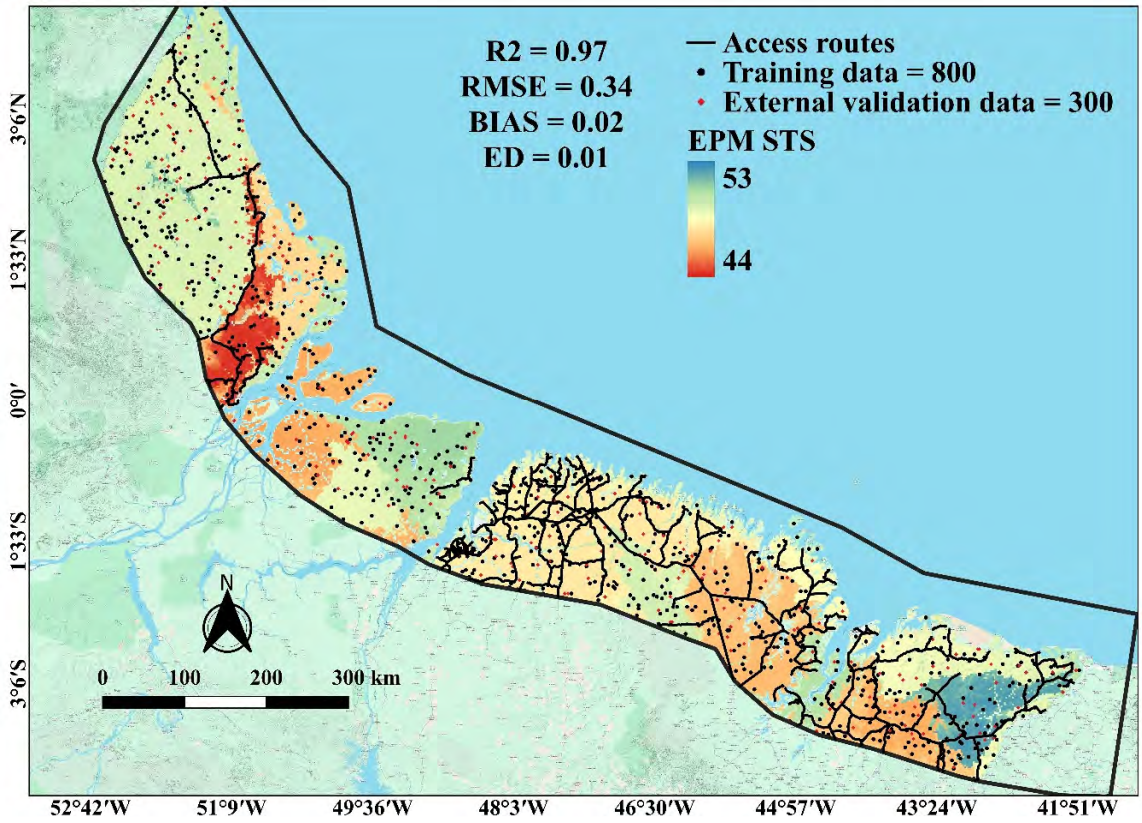


Figure 47. A simulated surface map from the data distributed throughout the study area considered 800 points; external validation was also distributed throughout the study area with 300 points.

The spatial distribution of Gower's Dissimilarity Index was exhibited in Figure 48 (ranging from 0.46 to 0.61). It reflected the environmental heterogeneity across the study area and was directly influenced by the covariates used: geology, geomorphology, pedology, DEM, precipitation, temperature, land use, and land cover. Regions with higher dissimilarity (lighter shades, closer to 0.61) coincided with areas of overlapping or distinct transitions in geology (e.g., Amazon and Gurupi formations) and geomorphology (e.g., fluvial and coastal systems). These regions, particularly in the northeast and central parts of the study area, also exhibited more significant variability in soil classes, ranging from U1 to O3.

In contrast, areas with lower dissimilarity (darker shades, closer to 0.46) were found predominantly in the central and western regions, corresponding to more uniform land use (e.g.,

conservation units) and homogeneous elevation patterns (low-lying coastal plains in the DEM). These areas also showed less variation in climate variables, such as precipitation (closer to 3,400 mm) and temperature (around 25–26°C).

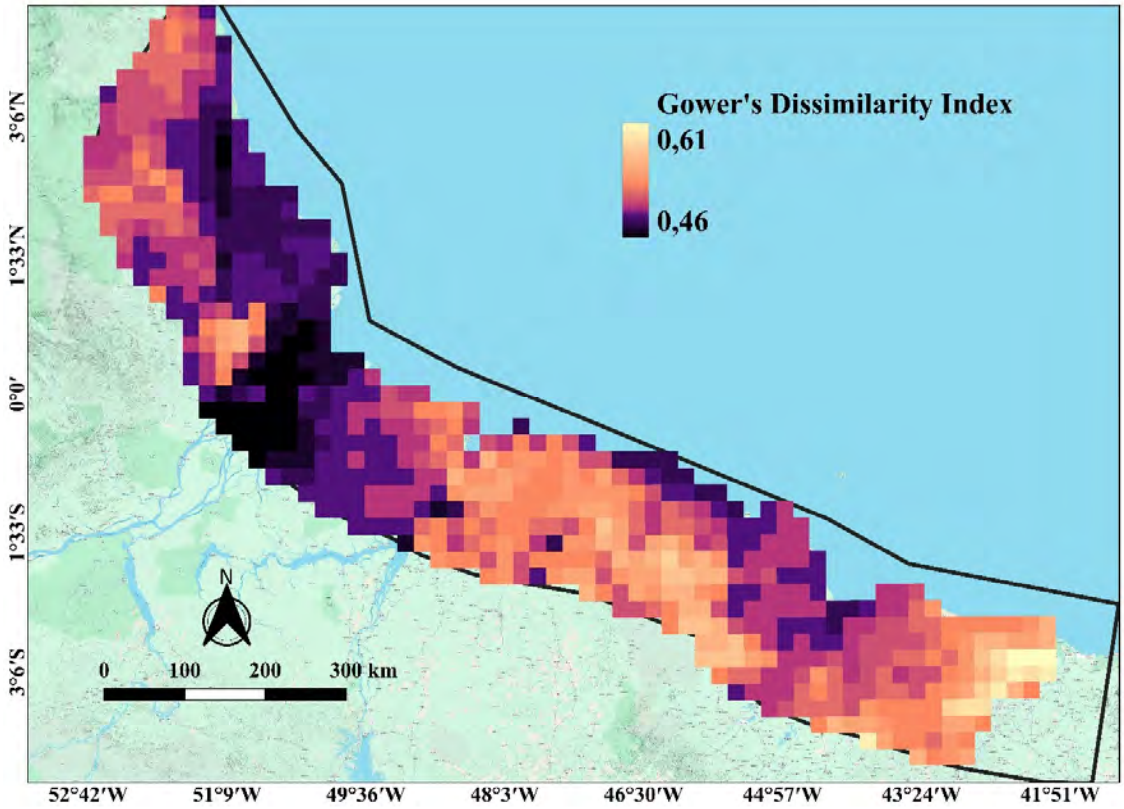


Figure 48. Gower's Dissimilarity Index for the Equatorial Margin of Brazil. The resolution of this map is associated with the block size parameter, which, in this case, was set to 10, meaning a 10 x 10 km^2 pixel resolution.

By comparing the remaining sample sizes for the RA framework part, which aimed to optimize the smallest sample size number associated with the smallest RA target area, Figure 48 showed the best ED values for each tested sample size for each proportionality of Gower's Dissimilarity Index high and low aggregation at target area ratios of 10%, 20%, 30%, 40%, and 50%, with a benchmark set at sample size = 800. The 20% improvement cutoff for ED was applied, comparing smaller sample sizes to the benchmark, validated by 300 external points scattered across the total area. Sample sizes for which ED exceeded the 20% cutoff compared to the benchmark were considered insufficient.

The graph presented in Figure 49 evaluated the relationship between sample size and Euclidean Distance (ED) for different target area sizes (autoRA 10%, 20%, 30%, 40%, and 50%) under varying Gower's dissimilarity index proportions. The x-axis showed the proportional aggregation of pixels based on high and low dissimilarity indices. For example, for a target area ratio of 10%, the test started with 10% of the area's pixels representing high dissimilarity indices and 90% representing low dissimilarity indices, progressing in steps until 100% of the pixels

represented only high dissimilarity indices. This process was repeated for each target area size (autoRA 10%, 20%, 30%, etc.).

For the benchmark (sample size = 800), ED values remained consistently low, demonstrating their ability to capture heterogeneity effectively across all proportions of high and low dissimilarity indices. When the sample size was reduced to below 800, ED increased, particularly at lower ratios of high dissimilarity pixels, indicating that smaller sample sizes struggled to capture the environmental complexity of these areas. Sample sizes below 600 frequently exceeded the 20% ED improvement cutoff, making them unsuitable for robust modeling.

For the target area of 10%, sample sizes of 700 and 800 met the 20% improvement cutoff for ED across most proportions. In contrast, smaller sizes failed, especially when the proportion of high dissimilarity pixels was low (10–40%). For target areas 20% and 30%, sample sizes of 600, 700, and 800 consistently stayed within the 20% cutoff, particularly as the proportion of high dissimilarity pixels increased. The performance improved for target areas by 40% and 50%, with sample sizes of 600 meeting the cutoff for most proportions of high dissimilarity pixels, reflecting a better representation of heterogeneity at larger target area sizes.

In conclusion, the sample size 800 remained the most reliable across all target areas and Gower proportions. However, sample sizes of 600 and 700 were sufficient for target areas 20%, 30%, 40%, and 50%, mainly when higher proportions of high dissimilarity pixels were used. Sample sizes below 600 failed to meet the 20% ED improvement threshold and were unsuitable for accurately capturing heterogeneity—larger sample sizes beyond 800 offered minimal additional improvement, indicating diminishing returns. For cost-efficient modeling, sample size = 600 was recommended for higher target areas and proportions of high dissimilarity pixels, whereas sample size = 800 was ideal for ensuring the highest accuracy.

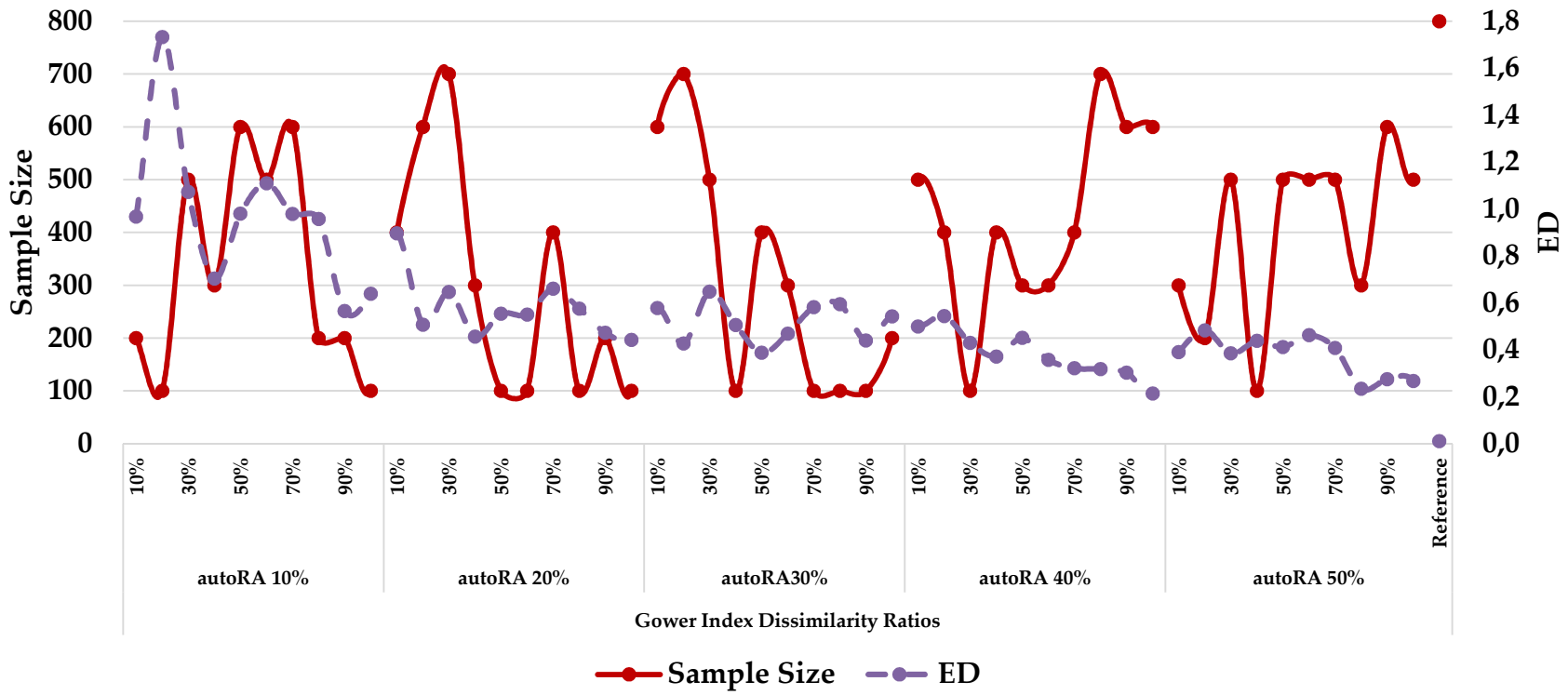


Figure 409. Graph the Euclidean Distance (ED) calculated for the models adjusted from the different sets of remaining points sampled within each reference area with the proportionalities of the Gower dissimilarity index clustering.

The RAM with the autoRA RA 40% configuration was designed to validate that its STS map was as accurate as the benchmark EPM map while reducing the sampled area by 60% and the training sample size from 800 to 600 points. The RAM and EPM maps were initially validated using the same 300 external validation points scattered across the total area to ensure a valid comparison. This approach allowed a direct evaluation of performance metrics between the two approaches.

The RAM achieved strong results, with R^2=0.87, RMSE=0.68, BIAS=0.08, and ED=0.21 (Figure 50). These metrics were within the 20% ED improvement cutoff relative to the EPM benchmark (R^2=0.97, RMSE=0.34, BIAS=0.02, ED=0.01), demonstrating that the RAM STS map was highly comparable to the EPM while significantly reducing field effort. By focusing on the 40% most heterogeneous areas (as determined by Gower's Dissimilarity Index), the RAM effectively captured key environmental gradients without requiring exhaustive coverage as the EPM demanded under the TA approach.

After confirming these results, a new proportional external validation dataset was generated by randomly reducing the number of samples in the original validation dataset to 300 points. This proportional dataset corresponded to the 40% RA used by the RAM and included 257 validation points (maintaining the original 70:30 split between training and validation). This adjustment ensured consistency in validation proportionality between the training and external validation datasets, aligning with the reduced spatial scope of the RA configuration.

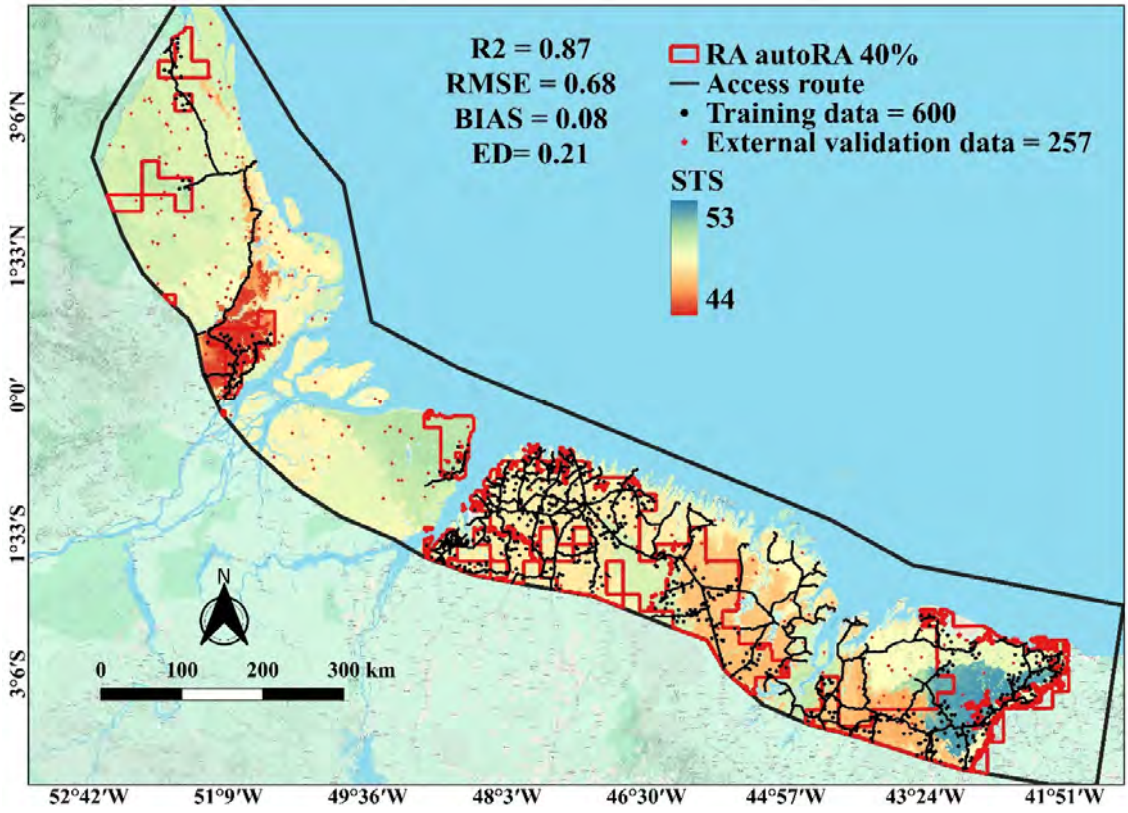


Figure 50. A simulated surface map was predicted using a model with a target_area of 40%, 600 training points, and a distribution of the points in the buffer 10 km from the roads.

5. CONCLUSIONS

Applying the autoRA (Automated Reference Area) approach within the Equatorial Margin of Brazil proved to be a highly effective strategy for optimizing soil sampling methodologies that balanced model accuracy with cost efficiency. Central to this success was incorporating sensitivity tests facilitated by 100 algebraic iterations of the Simulated Theoretical Surface (STS). These iterations were meticulously designed to reflect feasible, dimensionless covariate scenarios associated with SCORPAN (Soil, Climate, Organisms, Relief, Parent Material, Location, and Time) soil-derived properties. By performing 100 STS simulations and averaging the results, autoRA ensured a robust representation of environmental variability, critical for accurate predictive modeling.

Through the Total Area (TA) methodology, a sample size of n=800 was identified as optimal, achieving a high coefficient of determination (R^2=0,97), low Root Mean Square Error (RMSE=0,34), minimal Bias (0,02), and an impressive Euclidean Distance (ED=0,01) relative to the benchmark model. This configuration delivered exceptional model performance and resulted in substantial cost savings, reducing expenditure by $200,00 compared to a larger sample size of 1,00. The spatial distribution analysis confirmed that the model effectively captured regional environmental heterogeneity, which is essential for accurate predictive insights in complex terrains.

Furthermore, the autoRA methodology's ability to utilize the same map covariates in both pre-fieldwork planning and post-fieldwork modeling stages underscored its utility as a comprehensive tool for project planning. By aligning the sampling strategy with the covariates used in digital soil mapping (DSM), autoRA allowed users to accurately dimension critical project parameters, including budget allocation, survey labor days, and acceptable levels of sampling accuracy. This integration ensured the sampling strategy was economically feasible and scientifically rigorous, enabling informed decision-making tailored to specific project requirements.

The validation of the Reference Area Model (RAM) with a reduced sample size of 600 points further demonstrated autoRA's flexibility and efficiency. Achieving R^2=0.87, RMSE=0.68, Bias=0.08, and ED=0.21, the RAM maintained substantial model accuracy while significantly reducing field effort and associated costs. This validation highlighted autoRA's capacity to focus sampling efforts on the most heterogeneous regions, as determined by Gower's Dissimilarity Index, thereby optimizing resource utilization without compromising the integrity of the predictive model.

Notably, the metrics presented for both the Exhaustive Predicted Model (EPM) and the Reference Area Model (RAM) were derived from Random Forest (RF) external validations of the STS-predicted models. Each RF model was rigorously tested against a consistent external validation set of 300 samples, ensuring that performance assessments were reliable and comparable across different sampling strategies. This systematic validation approach underscored autoRA's robustness in handling varying target area sizes and proportions of Gower's Dissimilarity Index, affirming its applicability across diverse environmental contexts.

The autoRA approach offered a robust, statistically validated, cost-effective framework for optimizing soil sampling strategies in environmentally diverse and resource-constrained regions like the Equatorial Margin of Brazil. Its ability to integrate sensitivity testing through simulated theoretical surface iterations, maximize entropy by capturing spatial variability, and align pre-planning with post-fieldwork covariate usage ensured that sampling projects were economically viable and scientifically sound. Consequently, autoRA stood out as an invaluable tool for enhancing the efficiency and accuracy of environmental studies, facilitating sustainable and informed decision-making in soil science and related disciplines.

Moreover, the study confirmed the hypothesis that by systematically adjusting RA size, proportions of high/low dissimilarity values, and repeated sampling for Random Forest modeling, autoRA identified the minimal portion of the TA that yielded predictive performance comparable to full-coverage sampling.

6. GENERAL CONCLUSIONS

Collectively, the three studies provide substantial theoretical and experimental evidence that the autoRA (Automated Reference Area) framework is a unique and reliable tool for optimizing soil sampling strategies in soil mapping. Unlike any existing methodology, autoRA fully automates the delineation of reference areas by systematically integrating Gower's Dissimilarity Index, environmental heterogeneity, and cost–accuracy trade-offs—no other approach offers this level of automation and comprehensiveness. Across different geographic settings, autoRA has repeatedly demonstrated its ability to identify optimal configurations of target area coverage and sample size, ensuring maximum representativeness of soil-forming factors while substantially reducing fieldwork costs.

The results show that autoRA not only refines RA delineation but also optimizes the number of samples required for accurate DSM. This dual optimization—reference area size and sampling density—was validated in multiple case studies (Rio de Janeiro, Florida, and the Equatorial Margin of Brazil), where autoRA's configurations closely matched benchmark accuracies at a fraction of the cost, often exceeding 60% in savings. Sensitivity tests using Simulated Theoretical Surface (STS) iterations confirmed that autoRA consistently captures critical environmental gradients, achieving strong predictive performance even with reduced samples. Such performance gains underscore the method's theoretical soundness and practical reliability.

By reducing subjective expert input and ensuring reproducibility, autoRA holds particular value for resource-limited or logistically challenging regions, where field sampling must be scientifically rigorous and financially feasible. Moreover, the ability to generate pre-fieldwork accuracy estimates—using the same covariates for planning and modeling—makes autoRA a powerful decision-support tool. It allows project managers to accurately anticipate labor days, budget allocations, and acceptable error thresholds before field campaigns begin. Consequently, autoRA stands as a trustworthy solution for future soil sampling initiatives, offering a robust, cost-effective, and scientifically validated framework that advances both the efficiency and the accuracy of digital soil mapping.

7. FINAL CONSIDERATIONS ON AUTORA IMPLEMENTATION AND APPLICABILITY

The autoRA methodology introduces a systematic, data-driven approach to optimizing reference areas (RAs) for Digital Soil Mapping (DSM). By leveraging Gower's Dissimilarity Index, sensitivity analysis, and machine learning (Random Forest), autoRA refines soil sampling locations, enhancing predictive efficiency. This section addresses key questions regarding computational demand, required expertise, cost-reduction potential, recommended conditions, applicability, and limitations of autoRA.

1) **What is the computational demand required to use autoRA?** The computational complexity of autoRA depends on the parameters Study Area Size (A): A larger area increases the number of computational operations; Raster Grid Resolution (R): A finer resolution significantly increases processing time and memory usage. Number of Environmental Covariates (C): More covariates result in higher computational loads for distance calculations. Sensitivity Analysis Complexity (B, T): The number of block sizes (B) and target areas (T) tested increase the number of iterations. Number of Simulated Surfaces (S): The STS (Simulated Theoretical Surface) method requires multiple simulations to generate robust training data. Sampling Density (N): More sample points also increase model training time.

If the total number of raster cells in a study area is given by:

$$\boldsymbol{N_{cells} = \frac{A}{R^2}}$$

The total number of pairwise Gower dissimilarity calculations required is approximately:

$$\boldsymbol{O\left(N_{cells}^2 \cdot C\right)}$$

For model training, assuming a sample size N, the Random Forest complexity is:

$$\boldsymbol{O(N \cdot C \cdot T)}$$

Where T is the number of trees in the forest.

For a 100 km² area at different resolutions:

Table 4. Estimation of computational demand of autoRA applied at a 100 km^2 area.

Resolution	Number of Cells (N_{cells})	Gower Calculations ($N_{cells}^2 \cdot C$)	Estimated Processing Time*
100m (1 ha)	1,000,000	10^{12}	1-2 hours
30m (0.09 ha)	11,111,111	10^{14}	10-20 hours
10m (0.01 ha)	100,000,000	10^{16}	Days to Weeks
1m (1m²)	10,000,000,000	10^{20}	Infeasible without HPC

(*) Estimated time depends on RAM availability, CPU parallelization, and algorithm optimizations.

For large-scale analyses (e.g., PronaSolos level mapping), it is recommended to downsample input rasters to 100m or 1000m resolutions or use cloud computing (HPC clusters) to process higher resolutions for test simulations before proceeding with the final reference area outlined shape.

2) **What expertise is required for autoRA compared to traditional methods?** autoRA was designed to reduce the need for deep technical expertise in soil mapping by automating key processes and providing a user-friendly interface via Shiny applications. A user unfamiliar with R programming can easily interact with autoRA by pointing and clicking to configure key parameters such as block size and target area proportions. Meanwhile, the Random Forest model tuning and Simulated Theoretical Surface (STS) generation run in the background with default settings, making the tool accessible to professionals with little coding experience.

However, users should still have some GIS and soil science expertise to understand the input covariates and their relevance to soil formation. Interpreting the autoRA outputs—including Gower dissimilarity maps and sensitivity analyses—requires this necessary expertise, similar to both the autoRA approach and traditional methods, since statistics is a core methodology in soil science. The primary output generated by autoRA is the delineation of a reference area for planning field sampling projects, meaning that the user must be familiar with soil-forming factor interactions and check the plausibility of the delineated output.

autoRA intends not to dispense with the deep pedological knowledge in manually delineating a reference area. Instead, autoRA is designed for specialists who intend to use the reference area hypothesis to plan soil sampling. It offers cost optimization by testing the possibility of identifying heterogeneous regions concerning the maps grouped by the specialist and used in the autoRA process. The output generated by autoRA can then be criticized and compared to what would supposedly be the ideal reference area, assuming the user understands that the manual delineation method is the "correct" one and, consequently, the standard, with autoRA presenting a similar format in its generated reference area.

In this way, while autoRA minimizes subjectivity and enhances reproducibility through data-driven sampling optimization, traditional mapping demands higher field expertise and manual interpretation, which varies by mapper. autoRA thus lowers the technical barrier for digital soil mapping, allowing a broader range of professionals to leverage automated sample optimization using the reference area hypothesis.

3) **Is it possible to reconcile strategies to reduce the number of samples based on field campaigns, thus reducing the cost of the following steps as expensive as (the laboratory)?** Yes, autoRA allows for strategic sample reduction while maintaining prediction accuracy, helping to cut down expensive laboratory analyses. Some strategies that can be pulled together are prioritizing high Gower Dissimilarity areas and focusing the sampling in regions that maximize model performance while avoiding redundant samples.

Legacy data could complement new campaign planning by replacing the suggested sample positioning with the autoRA output. Also, a multi-stage sampling approach can be adopted, starting with a first-stage sampling using low-cost methods (e.g., spectroscopy) to refine reference areas before sending samples for laboratory analysis.

4) **From the results in the 3 chapters, under what conditions (environments) do you recommend using the algorithm, and what are the minimum data requirements, considering the demands of the survey scale proposed in PronaSolos?** autoRA is applicable in any situation that requires soil sampling planning. Unlike traditional large-scale mapping, where systematic grids are often used, autoRA can be applied to extensive and small-scale projects, regardless of the spatial scale.

autoRA can calculate the Gower Dissimilarity Index for a single environmental layer if the user considers it a dominant factor in soil variability. If the multiple environmental layers are combined, the non-sensitivity of Gower's metric is leveraged to the number of layers (i.e., adding more layers does not overly inflate distance values).

A minimum computational power adequate for raster processing (e.g., multi-core CPU, 16GB RAM minimum) is desirable but not mandatory.

5) **Is autoRA more efficient for predicting soil properties or classes?** autoRA does not inherently prefer continuous or categorical predictions. The Gower's Dissimilarity Index allows for both categorical and continuous data. The Simulated Theoretical Surface (STS) handles categorical and continuous raster layers correctly. Random Forest, the core modeling algorithm in autoRA, can be used for regression (continuous) and classification (categorical) tasks.

The autoRA's approach to data preparation handles categorical and continuous covariates, which are standardized and used together in model training. The STS modeling ensures proper treatment of categorical and continuous inputs, reducing bias toward one data type. The final output depends on the predicted response variable (whether it is a soil class or a continuous property like pH).

autoRA was dedicated and developed for pedometrics and digital soil mapping, where categorical (e.g., soil class, geology) and continuous (e.g., pH, organic matter) variables are commonly used.

The core methodology of autoRA allows any combination of data types to be used in Gower's Dissimilarity Index calculus, STS modeling, and Random Forest predictions.

6) **Limitations and Applicability of autoRA –** The output of the autoRA will be a reference area boundary with accuracy based on Gower's Dissimilarity Index and the STS, which are byproducts of the raster input resolution. In this way, the autoRA output quality depends on the quality of environmental data entered. Coarse predictor resolution leads to weak model performance.

The input of very fine spatial resolution maps (i.e. 3 meters pixel resolution) and an extensive area will be computationally intensive to generate the outputs. The tradeoff can be handled in the resampling strategies, such as resampling the input for a coarse (but not too much coarse) or using the block size argument with different values of pixel aggregation to calculate Gower's Dissimilarity Index. In this way, the sensitivity test will provide several outputs that can be interpreted by the user and can be decided if the tradeoff is acceptable given the desired accuracy and the resources to implement the strategy on the field and at the computer.

The applicability of the autoRA is related to all the fields of optimized sampling strategies, which provide guided sampling campaigns in a statistically sound way.

The data-driven approach implemented at the autoRA and the reliance on pixel formats maps allow the autoRA to be implemented in any situation where a raster map is available, available for large and small-scale soil mapping, assisting in reference area selection across multiple scales.

The autoRA was intentionally developed to improve the tradeoff of cost-effective DSM workflow by reducing laboratory costs by focusing resources on high-variance areas to be sampled.

7) **Future Steps for the autoRA** – The development of autoRA has laid the foundation for optimizing soil sampling strategies by automating the delineation of reference areas based on spatial heterogeneity. Moving forward, three key advancements are envisioned to enhance the applicability and efficiency of autoRA in digital soil mapping and monitoring.

The first enhancement involves incorporating legacy data from the study region as additional information to refine sampling strategies. By leveraging existing soil observations, historical surveys, and remote sensing archives, autoRA can assess whether certain regions require fewer new samples while maintaining predictive accuracy. This approach has the

potential to significantly reduce sampling costs and field effort by integrating prior knowledge into the selection process. A methodological framework will be developed to quantify how legacy data contribute to reference area selection and to establish a decision criterion for sample reduction without compromising model performance.

The second advancement focuses on introducing a state-space analysis approach for the planning of long-term monitoring and the prediction of soil attributes over time. Soil properties are dynamic, influenced by environmental changes, land use, and management practices. Incorporating state-space models will allow autoRA to move beyond static reference area delineation and support spatiotemporal predictions. This approach will enable adaptive monitoring strategies, where sampling density and locations are optimized over time based on observed changes in soil attributes. By integrating temporal dynamics, autoRA can contribute to more robust soil management frameworks, facilitating decision-making in agricultural and environmental monitoring contexts.

The third improvement involves introducing alternative dissimilarity measures and sampling designs to enhance the flexibility of autoRA. While Gower's dissimilarity has been the primary metric for capturing spatial heterogeneity, other distance measures such as Mahalanobis distance, geostatistical variograms, or machine learning-based dissimilarity metrics could be explored. Additionally, integrating different sampling design strategies, such as stratified, adaptive, or model-based sampling, may improve the efficiency of reference area selection, particularly in heterogeneous landscapes. Evaluating the impact of these alternative methods on prediction accuracy and sampling efficiency will be an essential step in the continued development of autoRA.

By implementing these advancements, autoRA will evolve into a more comprehensive tool capable of integrating past knowledge, predicting future soil conditions, and leveraging diverse statistical approaches, ultimately enhancing its utility for soil monitoring, precision agriculture, and sustainable land management.

8. BIBLIOGRAPHIC REFERENCES

ABDULRAHEEM, M. I.; ZHANG, W.; LI, S.; MOSHAYEDI, A. J.; FAROOQUE, A. A.; HU, J. Advancement of Remote Sensing for Soil Measurements and Applications: A Comprehensive Review. **Sustainability**, vol. 15, no. 21, p. 15444, Jan. 2023. DOI 10.3390/su152115444. Available at: https://www.mdpi.com/2071-1050/15/21/15444. Accessed on: 10 Jan. 2025.

ABREU, S. F. Brazilian Oil Fields and Oil-Shale Reserves. **Bulletin of the American Association of Petroleum Geologists**, vol. 33, no. 9, p. 10, 1949. DOI 10.1306/3D933DF4-16B1-11D7-8645000102C1865D. Available at: http://search.datapages.com/data/doi/10.1306/3D933DF4-16B1-11D7-8645000102C1865D. Accessed on: 23 Jan. 2025.

ADAMS, R. A.; FOURNIER, J. J. F. (Eds.). 4 - The Sobolev Imbedding Theorem. **Pure and Applied Mathematics**. Sobolev Spaces. [*S. l.*]: Elsevier, 2003. vol. 140, p. 79–134. DOI 10.1016/S0079-8169(03)80006-5. Available at: https://www.sciencedirect.com/science/article/pii/S0079816903800065. Accessed on: 29 Jan. 2025.

ADRIANO, L.; ZALÁN, P.; HIDALGO-GATO, M.; CUNHA, A.; YALAMANCHILI, R.; SILVA, D. Integrated interpretation of the Western Portion of the Pará – Maranhão Basin – Brazilian Equatorial Margin. In: 14TH INTERNATIONAL CONGRESS OF THE BRAZILIAN GEOPHYSICAL SOCIETY & EXPOGEF, RIO DE JANEIRO, BRAZIL, 3-6 AUGUST 2015, 6 Aug. 2015. **14th International Congress of the Brazilian Geophysical Society & EXPOGEF, Rio de Janeiro, Brazil, 3-6 August 2015** [...]. Rio de Janeiro: Brazilian Geophysical Society, 6 Aug. 2015. p. 43–48. DOI 10.1190/sbgf2015-008. Available at: https://library.seg.org/doi/10.1190/sbgf2015-008. Accessed on: 23 Jan. 2025.

ARRUDA, G. P.; DEMATTÊ, J. A. M.; CHAGAS, C. S.; FIORIO, P. R.; SOUZA, A. B.; FONGARO, C. T. Digital soil mapping using reference area and artificial neural networks. **Scientia Agricola**, vol. 73, p. 266–273, 2016. DOI 10.1590/0103-9016-2015-0131. Available at: https://www.scielo.br/j/sa/a/HWJLXB9TBj9Y43X8GN3T4Qf/?lang=en. Accessed on: 22 Sep. 2024.

BARROS, G. M.; SANTOS, J. C. B.; SOUZA JÚNIOR, V. S.; DELARMELINDA, E. A.; ARAÚJO FILHO, J. C.; CÂMARA, E. R. G. Association between parent materials and soil attributes along different geological environments in western Pará, Brazil. **Acta Amazonica**, vol. 48, p. 261–270, Sep. 2018. DOI https://doi.org/10.1590/1809-4392201703322. Available at: https://www.scielo.br/j/aa/a/Nqs5KGps8hWDSzPFgrgVfKD/?lang=en. Accessed on: 23 Jan. 2025.

BARUCK, J.; NESTROY, O.; SARTORI, G.; BAIZE, D.; TRAIDL, R.; VRŠČAJ, B.; BRÄM, E.; GRUBER, F. E.; HEINRICH, K.; GEITNER, C. Soil classification and mapping in the Alps: The current state and future challenges. **Geoderma**, Soil mapping, classification, and modelling: history and future directions. vol. 264, p. 312–331, 15 Feb. 2016. DOI 10.1016/j.geoderma.2015.08.005. Available at: https://www.sciencedirect.com/science/article/pii/S0016706115300343. Accessed on: 23 Dec. 2024.

BIJOS, N. R.; DA SILVA, D. P.; MUNHOZ, C. B. R. Soil texture and fertility determine the beta diversity of plant species in veredas in Central Brazil. **Plant and Soil**, vol. 492, no. 1, p. 241–259, 1 Nov. 2023. DOI 10.1007/s11104-023-06168-3. Available at: https://doi.org/10.1007/s11104-023-06168-3. Accessed on: 23 Jan. 2025.

BISWAS, A.; ZHANG, Y. Sampling Designs for Validating Digital Soil Maps: A Review. **Pedosphere**, vol. 28, no. 1, p. 1–15, Feb. 2018. DOI 10.1016/S1002-0160(18)60001-3. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1002016018600013. Accessed on: 30 Sep. 2024.

BOETTINGER, J. L. (Ed.). **Digital soil mapping: bridging research, environmental application, and operation**. Dordrecht [Netherlands] ; London: Springer, 2010(Progress in soil science, 2).

BOVOLO, C. I.; PEREIRA, R.; PARKIN, G.; KILSBY, C.; WAGNER, T. Fine-scale regional climate patterns in the Guianas, tropical South America, based on observations and reanalysis data. **International Journal of Climatology**, vol. 32, no. 11, p. 1665–1689, 2012. DOI 10.1002/joc.2387. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/joc.2387. Accessed on: 23 Jan. 2025.

BRUNGARD, C. W.; BOETTINGER, J. L. Conditioned Latin Hypercube Sampling: Optimal Sample Size for Digital Soil Mapping of Arid Rangelands in Utah, USA. In: BOETTINGER, J. L.; HOWELL, D. W.; MOORE, A. C.; HARTEMINK, A. E.; KIENAST-BROWN, S. (eds.). **Digital Soil Mapping: Bridging Research, Environmental Application, and Operation**. Dordrecht: Springer Netherlands, 2010. p. 67–75. DOI 10.1007/978-90-481-8863-5_6. Available at: https://doi.org/10.1007/978-90-481-8863-5_6. Accessed on: 29 Dec. 2024.

BRUS, D. J. Statistical sampling approaches for soil monitoring. **European Journal of Soil Science**, vol. 65, no. 6, p. 779–791, 2014. DOI 10.1111/ejss.12176. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/ejss.12176. Accessed on: 21 Sep. 2024.

BRUS, D. J.; KEMPEN, B.; HEUVELINK, G. B. M. Sampling for validation of digital soil maps. **European Journal of Soil Science**, vol. 62, no. 3, p. 394–407, 2011. DOI 10.1111/j.1365-2389.2011.01364.x. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2389.2011.01364.x. Accessed on: 30 Sep. 2024.

BRUS, D. J.; VAŠÁT, R.; HEUVELINK, G. B. M.; KNOTTERS, M.; DE VRIES, F.; WALVOORT, D. J. J. Towards a Soil Information System with quantified accuracy. A prototype for mapping continuous soil properties. **Wageningen, Statutory Research Tasks Unit for Nature and the Environment**, vol. WOt-werkdocument 197, p. 160, 2009. Available at: https://edepot.wur.nl/148932. Accessed on: 21 Sep. 2024.

CANAVESI, V.; SEGONI, S.; ROSI, A.; TING, X.; NERY, T.; CATANI, F.; CASAGLI, N. Different Approaches to Use Morphometric Attributes in Landslide Susceptibility Mapping Based on Meso-Scale Spatial Units: A Case Study in Rio de Janeiro (Brazil). **Remote Sensing**, vol. 12, no. 11, p. 1826, 5 Jun. 2020. DOI 10.3390/rs12111826. Available at: https://www.mdpi.com/2072-4292/12/11/1826. Accessed on: 23 Dec. 2024.

CARNEIRO, P. B. M.; XIMENES NETO, A. R.; JUCÁ-QUEIROZ, B.; TEIXEIRA, C. E. P.; FEITOSA, C. V.; BARROSO, C. X.; MATTHEWS-CASCON, H. DE MORAIS, J. O.;

FREITAS, J. E. P.; SANTANDER-NETO, J.; DE ARAÚJO, J. T.; MONTEIRO, L. H. U.; PINHEIRO, L. S.; BRAGA, M. D. A.; CORDEIRO, R. T. S.; ROSSI, S.; BEJARANO, S.; SALANI, S.; GARCIA, Tatiane M.; SOARES, M. O. Interconnected marine habitats form a single continental-scale reef system in South America. **Scientific Reports**, vol. 12, no. 1, p. 17359, 17 Oct. 2022. DOI 10.1038/s41598-022-21341-x. Available at: https://www.nature.com/articles/s41598-022-21341-x. Accessed on: 23 Jan. 2025.

CARTER, M. R.; GREGORICH, E. G. (Eds.). **Soil Sampling and Methods of Analysis**. 0 ed. [*S. l.*]: CRC Press, 2007. DOI 10.1201/9781420005271. Available at: https://www.taylorfrancis.com/books/9781420005271. Accessed on: 19 Oct. 2024.

CARVALHO, W.; PEREIRA, N. R.; FERNANDES, E. I.; CALDERANO, B.; PINHEIRO, H. S. K.; CHAGAS, C. S.; BHERING, S. B.; PEREIRA, V. R.; LAWALL, S. Sample design effects on soil unit prediction with machine: randomness, uncertainty, and majority map. **Revista Brasileira de Ciência do Solo**, vol. 44, p. e0190120, 6 Aug. 2020. DOI 10.36783/18069657rbcs20190120. Available at: https://www.rbcsjournal.org/article/sample-design-effects-on-soil-unit-prediction-with-machine-randomness-uncertainty-and-majority-map/. Accessed on: 2 Jan. 2025.

CASA, R.; CASTALDI, F.; PASCUCCI, S.; BASSO, B.; PIGNATTI, S. Geophysical and Hyperspectral Data Fusion Techniques for In-Field Estimation of Soil Properties. **Vadose Zone Journal**, vol. 12, no. 4, p. vzj2012.0201, 2013. DOI 10.2136/vzj2012.0201. Available at: https://onlinelibrary.wiley.com/doi/abs/10.2136/vzj2012.0201. Accessed on: 17 Aug. 2024.

CEDDIA, M. B.; VILLELA, A. L. O.; PINHEIRO, É. F. M.; WENDROTH, O. Spatial variability of soil carbon stock in the Urucu river basin, Central Amazon-Brazil. **Science of The Total Environment**, vol. 526, p. 58–69, 1 Sep. 2015. DOI 10.1016/j.scitotenv.2015.03.121. Available at: https://www.sciencedirect.com/science/article/pii/S0048969715004052. Accessed on: 29 Nov. 2024.

CESCO, S.; SAMBO, P.; BORIN, M.; B., B.; ORZES, G.; MAZZETTO, F. Smart agriculture and digital twins: Applications and challenges in a vision of sustainability. **European Journal of Agronomy**, vol. 146, p. 126809, May 2023. DOI 10.1016/j.eja.2023.126809. Available at: https://linkinghub.elsevier.com/retrieve/pii/S1161030123000771. Accessed on: 24 Feb. 2025.

CHEN, L.; WANG, W.; WANG, C.; YAN, X.; ZHANG, Yuhan; SHEN, Zhenyao. From field soil sampling to watershed model: Upscaling by integrating information entropy and interpolation method. **Journal of Environmental Management**, vol. 360, p. 121119, 1 Jun. 2024. DOI 10.1016/j.jenvman.2024.121119. Available at: https://www.sciencedirect.com/science/article/pii/S0301479724011058. Accessed on: 23 Dec. 2024.

CHENG, X.; LUO, Y.; XU, X.; SHERRY, R.; ZHANG, Q. Soil organic matter dynamics in a North America tallgrass prairie after 9 yr of experimental warming. **Biogeosciences**, vol. 8, no. 6, p. 1487–1498, 9 Jun. 2011. DOI 10.5194/bg-8-1487-2011. Available at: https://bg.copernicus.org/articles/8/1487/2011/. Accessed on: 10 Sep. 2024.

CLINGENSMITH, Christopher M.; GRUNWALD, Sabine. Predicting Soil Properties and Interpreting Vis-NIR Models from across Continental United States. **Sensors**, vol. 22, no. 9, p.

3187, Jan. 2022. DOI 10.3390/s22093187. Available at: https://www.mdpi.com/1424-8220/22/9/3187. Accessed on: 2 Aug. 2024.

CLOTHIER, B. E.; POLLOK, J. A.; SCOTTER, D. R. Mottling in Soil Profiles Containing a Coarse-textured Horizon. **Soil Science Society of America Journal**, vol. 42, no. 5, p. 761–763, 1978. DOI 10.2136/sssaj1978.03615995004200050022x. Available at: https://onlinelibrary.wiley.com/doi/abs/10.2136/sssaj1978.03615995004200050022x. Accessed on: 19 Jan. 2025.

COSTA, L. R. F.; MAIA, R. P.; BARRETO, L. L.; SALES, V. C. C. Geomorphology of the Northern Brazilian Northeast: A Classification Proposal. **Revista Brasileira de Geomorfologia**, vol. 21, no. 1, 1 Jan. 2020. DOI 10.20502/rbg.v21i1.1447. Available at: https://rbgeomorfologia.org.br/rbg/article/view/1447. Accessed on: 23 Jan. 2025.

COSTA, E. M.; CEDDIA, M. B.; SANTOS, F. N.; SILVA, L. O.; REZENDE, I. P. T.; FERNANDES, D. A. C. Training pedologist for soil mapping: Contextualizing methods and its accuracy using the project pedagogy approach. **Revista Brasileira de Ciência do Solo**, vol. 45, p. 33, 10 Mar. 2021. DOI 10.36783/18069657rbcs20200130. Available at: https://www.scielo.br/j/rbcs/a/XZMn6LVn6HdJKF8wpSZQ5Yk/?lang=en. Accessed on: 25 Nov. 2024.

COSTA, E. M.; RODRIGUES, H. M.; FERREIRA, A. C. S.; CEDDIA, M. B.; FERNANDES, D. A. C. Using Legacy Soil Data to Plan New Data Collection: Study Case of Rio de Janeiro State: Brazil. In: DE CARVALHO JUNIOR, W.; PINHEIRO, H. S. K.; VALLADARES, G. S. (eds.). **Pedometrics in Brazil**. Progress in Soil Science. Cham: Springer Nature Switzerland, 2024. p. 101–113. DOI 10.1007/978-3-031-64579-2_8. Available at: https://link.springer.com/10.1007/978-3-031-64579-2_8. Accessed on: 22 Feb. 2025.

DA SILVA FREITAS, L. C.; CAVALCANTI, L. C. S.; NETO, J. J. F. Geoenvironmental Diagnosis of the Protected Areas of the Spix's Macaw, Bahia. **Revista Brasileira de Geografia Fisica**, vol. 17, no. 5, p. 3416–3449, 2024. https://doi.org/10.26848/rbgf.v17.5.p3416-3449.

DANTAS, L. G.; DOS SANTOS, C. A. C.; SANTOS, C. A. G.; MARTINS, E. S. P. R.; ALVES, L. M. Future Changes in Temperature and Precipitation over Northeastern Brazil by CMIP6 Model. **Water**, vol. 14, no. 24, p. 4118, Jan. 2022. DOI 10.3390/w14244118. Available at: https://www.mdpi.com/2073-4441/14/24/4118. Accessed on: 23 Jan. 2025.

DE ALKMIM, F. F. Geological Background: A Tectonic Panorama of Brazil. *In*: VIEIRA, Bianca Carvalho; SALGADO, André Augusto Rodrigues; SANTOS, Leonardo José Cordeiro (eds.). **Landscapes and Landforms of Brazil**. Dordrecht: Springer Netherlands, 2015. p. 9–17. DOI 10.1007/978-94-017-8023-0_2. Available at: https://doi.org/10.1007/978-94-017-8023-0_2. Accessed on: 23 Jan. 2025.

DE CARVALHO JUNIOR, W.; PINHEIRO, H. S. K.; CEDDIA, M. B.; VALLADARES, G. S. **Pedometrics in Brazil**. Cham, Switzerland: Springer, 2024.

DE MORISSON VALERIANO, M.; DE FÁTIMA ROSSETTI, D. Topodata: Brazilian full coverage refinement of SRTM data. **Applied Geography**, vol. 32, no. 2, p. 300–309, 1 Mar. 2012. DOI 10.1016/j.apgeog.2011.05.004. Available at: https://www.sciencedirect.com/science/article/pii/S0143622811000786. Accessed on: 12 Jan. 2025.

DEFRAEYE, T.; SHRIVASTAVA, C.; BERRY, T.; VERBOVEN, Pi.; ONWUDE, D.; SCHUDEL, S.; BÜHLMANN, A.; CRONJE, P.; ROSSI, R. M. Digital twins are coming: Will we need them in supply chains of fresh horticultural produce? **Trends in Food Science & Technology**, vol. 109, p. 245–258, Mar. 2021. DOI 10.1016/j.tifs.2021.01.025. Available at: https://linkinghub.elsevier.com/retrieve/pii/S092422442100025X. Accessed on: 24 Feb. 2025.

DEVINE, S. M.; STEENWERTH, K. L.; O'GEEN, A. T. A regional soil classification framework to improve soil health diagnosis and management. **Soil Science Society of America Journal**, vol. 85, no. 2, p. 361–378, 2021. DOI 10.1002/saj2.20200. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/saj2.20200. Accessed on: 10 Jan. 2025.

DORNIK, A.; CHEŢAN, M. A.; DRĂGUŢ, L.; DICU, D. D.; ILIUŢĂ, A. Optimal scaling of predictors for digital mapping of soil properties. **Geoderma**, vol. 405, p. 115453, 1 Jan. 2022. DOI 10.1016/j.geoderma.2021.115453. Available at: https://www.sciencedirect.com/science/article/pii/S0016706121005334. Accessed on: 23 Dec. 2024.

ELLILI, Y.; WALTER, C.; MICHOT, D.; PICHELIN, P.; LEMERCIER, B. Mapping soil organic carbon stock change by soil monitoring and digital soil mapping at the landscape scale. **Geoderma**, vol. 351, p. 1-8, 1 Oct. 2019. DOI 10.1016/j.geoderma.2019.03.005. Available at: https://www.sciencedirect.com/science/article/pii/S0016706117320633. Accessed on: 23 Dec. 2024.

FALCONER, K. J.; MARSH, D. T. On the Lipschitz equivalence of Cantor sets. **Mathematika**, vol. 39, no. 2, p. 223–233, 1992. DOI 10.1112/S0025579300014959. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1112/S0025579300014959. Accessed on: 29 Jan. 2025.

FARR, T. G.; ROSEN, P. A.; CARO, E.; CRIPPEN, R.; DUREN, R.; HENSLEY, S.; KOBRICK, M.; PALLER, M.; RODRIGUEZ, E.; ROTH, L.; SEAL, D.; SHAFFER, S.; SHIMADA, J.; UMLAND, J.; WERNER, M.; OSKIN, M.; BURBANK, D.; ALSDORF, D. The Shuttle Radar Topography Mission. **Reviews of Geophysics**, vol. 45, no. 2, 2007. DOI 10.1029/2005RG000183. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1029/2005RG000183. Accessed on: 5 Dec. 2024.

FAVROT, J. C. Cartographie, caractérisation et interpretation des données pédologique en vue du drainage agricole. 1986. **Sol et eau: résumés des actes du séminaire** [...]. Paris, France: ORSTOM, 1986. p. 551–577.

FERNANDES, K.; JÚNIOR, J. M.; RIBON, A. A.; DE ALMEIDA, G. M.; MOITINHO, M. R.; DELARICA, D. L. D.; BAHIA, A. S. R. S.; OLIVEIRA, D. M. S. Characterization and detailed mapping of C by spectral sensor for soils of the Western Plateau of São Paulo. **Scientific Reports**, vol. 14, no. 1, p. 17311, 27 Jul. 2024. DOI 10.1038/s41598-024-66369-3. Available at: https://www.nature.com/articles/s41598-024-66369-3. Accessed on: 23 Dec. 2024.

FERREIRA, A. C. S.; CEDDIA, M. B.; COSTA, E. M.; PINHEIRO, É. F. M.; NASCIMENTO, M. M.; VASQUES, G. M. Use of Airborne Radar Images and Machine Learning Algorithms to Map Soil Clay, Silt, and Sand Contents in Remote Areas under the Amazon Rainforest. **Remote Sensing**, vol. 14, no. 22, p. 5711, Jan. 2022. DOI 10.3390/rs14225711. Available at: https://www.mdpi.com/2072-4292/14/22/5711. Accessed on: 21 Nov. 2024.

FERREIRA, A. C. S.; PINHEIRO, E. E. M.; COSTA, E. M.; CEDDIA, M. B. Predicting soil carbon stock in remote areas of the Central Amazon region using machine learning techniques. **Geoderma Regional**, vol. 32, p. e00614, 1 Mar. 2023. DOI 10.1016/j.geodrs.2023.e00614. Available at: https://www.sciencedirect.com/science/article/pii/S235200942300010X. Accessed on: 30 Sep. 2024.

FERRY, S.; WEINBERGER, S. Quantitative algebraic topology and Lipschitz homotopy. **Proceedings of the National Academy of Sciences**, vol. 110, no. 48, p. 19246–19250, 26 Nov. 2013. DOI 10.1073/pnas.1208041110. Available at: https://pnas.org/doi/full/10.1073/pnas.1208041110. Accessed on: 29 Jan. 2025.

FICK, S. E.; HIJMANS, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. **International Journal of Climatology**, vol. 37, no. 12, p. 4302–4315, 2017. DOI 10.1002/joc.5086. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/joc.5086. Accessed on: 23 Nov. 2024.

FILIPPINI-ALBA, J. M.; FLORES, C. A.; BERNARDI, A. C. Pedology in Precision Agriculture from a Brazilian context. **Revista de Ciências Agrícolas**, vol. 40, no. 3, p. e3216, 31 Oct. 2023. DOI 10.22267/rcia.20234003.216. Available at: https://revistas.udenar.edu.co/index.php/rfacia/article/view/7753. Accessed on: 23 Dec. 2024.

FONTANA, A.; CHAGAS, C. S.; DONAGEMMA, G. K.; MENEZES, A. R.; CALDERANO, B. Soils Developed on Geomorphic Surfaces in the Mountain Region of the State of Rio de Janeiro. **Rev. Bras. Ciênc. Solo**, vol. 41, p., 5 Dec. 2017. DOI 10.1590/18069657rbcs20160574. Available at: https://www.rbcsjournal.org/pt-br/article/soils-developed-on-geomorphic-surfaces-in-the-mountain-region-of-the-state-of-rio-de-janeiro/. Accessed on: 26 Dec. 2024.

FRANCINI-FILHO, R. B.; ASP, N. E.; SIEGLE, E.; HOCEVAR, J.; LOWYCK, K.; D'AVILA, N.; VASCONCELOS, A. A.; BAITELO, R.; REZENDE, Carlos E.; OMACHI, Claudia Y.; THOMPSON, Cristiane C.; THOMPSON, Fabiano L. Perspectives on the Great Amazon Reef: Extension, Biodiversity, and Threats. **Frontiers in Marine Science**, vol. 5, 23 Apr. 2018. DOI 10.3389/fmars.2018.00142. Available at: https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2018.00142/full. Accessed on: 23 Jan. 2025.

FREIRE, L. M.; LIMA, J. S.; VERÍSSIMO, C. U. V.; SILVA, E. V. Carste em Rochas Não Carbonáticas: contribuição ao estudo geomorfológico em cavernas de arenito da Amazônia Paraense (Karst in Non-Carbonate Rocks: contribution in geomorphological study in sandstones caves of the Paraense Amazon). **Revista Brasileira de Geografia Física**, vol. 10, no. 6, p. 1829–1845, 10 Jul. 2017. DOI 10.26848/rbgf.v10.6.p1829-1845. Available at: https://periodicos.ufpe.br/revistas/index.php/rbgfe/article/view/234054. Accessed on: 23 Jan. 2025.

FURTADO, A. M. M.; PONTE, F. C. Mapeamento de Unidades de Relevo do Estado do Pará. **Revista Geoamazônia**, vol. 2, no. 1, p. 56–67, 31 Dec. 2013. DOI 10.17551/2358-1778/geoamazonia.n1v2p56-67. Available at: http://www.bibliotekevirtual.org/index.php/2013-02-07-03-02-35/2013-02-07-03-03-11/2014-07-19-06-15-59/720-geoamazonia/n01v02/6615-mapeamento-de-unidades-de-relevo-do-estado-do-para.html. Accessed on: 24 Jan. 2025.

GAULD, D. B. Topological Properties of Manifolds. **The American Mathematical Monthly**, vol. 81, no. 6, p. 633–636, 1 Jun. 1974. DOI 10.1080/00029890.1974.11993635. Available at: https://doi.org/10.1080/00029890.1974.11993635. Accessed on: 29 Jan. 2025.

GELSLEICHTER, Y. A.; COSTA, E. M.; ANJOS, L. H. C.; MARCONDES, R. A. T.. Enhancing Soil Mapping with Hyperspectral Subsurface Images generated from soil lab Vis-SWIR spectra tested in southern Brazil. **Geoderma Regional**, vol. 33, p. e00641, Jun. 2023. DOI 10.1016/j.geodrs.2023.e00641. Available at: https://linkinghub.elsevier.com/retrieve/pii/S2352009423000378. Accessed on: 25 Nov. 2024.

GOMES, L. C.; BEUCHER, A. M.; MØLLER, A. B.; IVERSEN, B. V.; BØRGESEN, C. D.; ADETSU, D. V.; SECHU, G. L.; HECKRATH, G. J.; KOCH, J.; ADHIKARI, K.; KNADEL, M.; LAMANDÉ, M.; GREVE, M. B.; JENSEN, N. H.; GUTIERREZ, S.; BALSTRØM, T.; KOGANTI, T.; ROELL, Y.; PENG, Y.; GREVE, M. H.. Soil assessment in Denmark: Towards soil functional mapping and beyond. **Frontiers in Soil Science**, vol. 3, 31 Jan. 2023. DOI 10.3389/fsoil.2023.1090145. Available at: https://www.frontiersin.org/journals/soil-science/articles/10.3389/fsoil.2023.1090145/full. Accessed on: 10 Jan. 2025.

GONÇALVES, D. A. M.; PEREIRA, W. V. S.; JOHANNESSON, K. H.; PÉREZ, D. V.; GUILHERME, L. R. G.; FERNANDES, A. R.. Geochemical Background for Potentially Toxic Elements in Forested Soils of the State of Pará, Brazilian Amazon. **Minerals**, vol. 12, no. 6, p. 674, Jun. 2022. DOI 10.3390/min12060674. Available at: https://www.mdpi.com/2075-163X/12/6/674. Accessed on: 23 Jan. 2025.

GONÇALVES, R. V. S.; CARDOSO, J. C. F.; OLIVEIRA, P. E.; RAYMUNDO, D.; DE OLIVEIRA, D. C. The role of topography, climate, soil and the surrounding matrix in the distribution of Veredas wetlands in central Brazil. **Wetlands Ecology and Management**, vol. 30, no. 6, p. 1261–1279, 1 Dec. 2022. DOI 10.1007/s11273-022-09895-z. Available at: https://doi.org/10.1007/s11273-022-09895-z. Accessed on: 26 Dec. 2024.

GONÇALVES, T. G.; PONS, N. A. D.; MELLONI, E. G. P.; MANCINI, M.; CURI, N. Digital soil mapping: Predicting soil classes distribution in large areas based on existing soil maps from similar small areas. **Ciência e Agrotecnologia**, vol. 45, p. e007921, 2021. DOI 10.1590/1413-7054202145007921. Available at: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-70542021000100226&tlng=en. Accessed on: 23 Dec. 2024.

GOOLEY, L.; HUANG, J.; PAGÉ, D.; TRIANTAFILIS, J. Digital soil mapping of available water content using proximal and remotely sensed data. **Soil Use and Management**, vol. 30, no. 1, p. 139–151, 2014. DOI 10.1111/sum.12094. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/sum.12094. Accessed on: 18 Aug. 2024.

GOWER, J. C. A General Coefficient of Similarity and Some of Its Properties. **Biometrics**, vol. 27, no. 4, p. 857–871, 1971. DOI 10.2307/2528823. Available at: https://www.jstor.org/stable/2528823. Accessed on: 21 Nov. 2024.

GRIEVES, M. Intelligent digital twins and the development and management of complex systems. **Digital Twin**, vol. 2, p. 8, 25 May 2022. DOI 10.12688/digitaltwin.17574.1. Available at: https://digitaltwin1.org/articles/2-8/v1. Accessed on: 22 Feb. 2025.

GRIEVES, M.; HUA, E. Y. (Eds.). **Digital Twins, Simulation, and the Metaverse: Driving Efficiency and Effectiveness in the Physical World through Simulation in the Virtual Worlds**. Cham: Springer Nature Switzerland, 2024(Simulation Foundations, Methods and Applications). DOI 10.1007/978-3-031-69107-2. Available at: https://link.springer.com/10.1007/978-3-031-69107-2. Accessed on: 24 Feb. 2025.

GRIEVES, M.; VICKERS, J. Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. In: KAHLEN, F. J.; FLUMERFELT, S.; ALVES, Anabela (eds.). **Transdisciplinary Perspectives on Complex Systems**. Cham: Springer International Publishing, 2017. p. 85–113. DOI 10.1007/978-3-319-38756-7_4. Available at: http://link.springer.com/10.1007/978-3-319-38756-7_4. Accessed on: 22 Feb. 2025.

GRUIJTER, J. J. (Ed.). **Sampling for natural resource monitoring**. Berlin ; New York: Springer, 2006.

GRUNWALD, S. Current State of Digital Soil Mapping and What Is Next. In: BOETTINGER, Janis L.; HOWELL, D. W.; MOORE, A. C.; HARTEMINK, A. E.; KIENAST-BROWN, S. (eds.). **Digital Soil Mapping: Bridging Research, Environmental Application, and Operation**. Dordrecht: Springer Netherlands, 2010. p. 3–12. DOI 10.1007/978-90-481-8863-5_1. Available at: https://doi.org/10.1007/978-90-481-8863-5_1. Accessed on: 21 Nov. 2024.

GRUNWALD, S.; THOMPSON, J. A.; BOETTINGER, J. L. Digital Soil Mapping and Modeling at Continental Scales: Finding Solutions for Global Issues. **Soil Science Society of America Journal**, vol. 75, no. 4, p. 1201–1213, 2011. DOI 10.2136/sssaj2011.0025. Available at: https://onlinelibrary.wiley.com/doi/abs/10.2136/sssaj2011.0025. Accessed on: 20 Aug. 2024.

GRUNWALD, Sabine. Grand Challenges in Pedometrics-AI Research. **Frontiers in Soil Science**, vol. 1, 20 Aug. 2021. DOI 10.3389/fsoil.2021.714323. Available at: https://www.frontiersin.org/journals/soil-science/articles/10.3389/fsoil.2021.714323/full. Accessed on: 30 Aug. 2024.

GRUNWALD, S.; MURAD, M. O. F.; FARRINGTON, S.; WALLACE, W.; ROONEY, D.. Multi-Sensor Soil Probe and Machine Learning Modeling for Predicting Soil Properties. **Sensors**, vol. 24, no. 21, p. 6855, Jan. 2024. DOI 10.3390/s24216855. Available at: https://www.mdpi.com/1424-8220/24/21/6855. Accessed on: 29 Oct. 2024.

GRUNWALD, S.; VASQUES, Gustavo M.; RIVERO, R. G. Fusion of Soil and Remote Sensing Data to Model Soil Properties. **Advances in Agronomy**. [*S. l.*]: Elsevier, 2015. vol. 131, p. 1–109. DOI 10.1016/bs.agron.2014.12.004. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0065211314000169. Accessed on: 20 Aug. 2024.

GUIMARAES FILHO, A. G.; BORBA, P. Methodology for Land Mapping of Amapa State-A Special Case of Amazon Radiography Project. In: IGARSS 2020 - 2020 IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, 26 Sep. 2020. **IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium** [...]. Waikoloa, HI, USA: IEEE, 26 Sep. 2020. p. 1540–1543. DOI 10.1109/IGARSS39084.2020.9324673. Available at: https://ieeexplore.ieee.org/document/9324673/. Accessed on: 24 Jan. 2025.

HARTEMINK, A. E.; MCBRATNEY, A.; MENDONÇA-SANTOS, M. L. (Eds.). **Digital Soil Mapping with Limited Data**. Dordrecht: Springer Netherlands, 2008(SpringerLink Bücher). https://doi.org/10.1007/978-1-4020-8592-5.

HEIL, K.; SCHMIDHALTER, U. The Application of EM38: Determination of Soil Parameters, Selection of Soil Sampling Points and Use in Agriculture and Archaeology. **Sensors**, vol. 17, no. 11, p. 2540, Nov. 2017. DOI 10.3390/s17112540. Available at: https://www.mdpi.com/1424-8220/17/11/2540. Accessed on: 18 Aug. 2024.

HEILBRON, M.; SILVA, L. G. E.; ALMEIDA, J. C. H.; TUPINAMBÁ, M.; PEIXOTO, C.; VALERIANO, C. M.; LOBATO, M.; RODRIGUES, S. W. O.; RAGATKY, C. D.; SILVA, M. A.; MONTEIRO, T.; FREITAS, N. C.; MIGUENS, D.; GIRÃO, R. Proterozoic to Ordovician geology and tectonic evolution of Rio de Janeiro State, SE-Brazil: insights on the central Ribeira Orogen from the new 1:400,000 scale geologic map. **Brazilian Journal of Geology**, vol. 50, p. e20190099, 27 Apr. 2020. DOI https://doi.org/10.1590/2317-4889202020190099. Available at: https://www.scielo.br/j/bjgeo/a/5d4VCzVqGVRKttNYLNhn8Pf/. Accessed on: 22 Feb. 2025.

HENGL, T.; JESUS, J. M.; MACMILLAN, R. A.; BATJES, N. H.; HEUVELINK, G. B. M.; RIBEIRO, E.; SAMUEL-ROSA, A.; KEMPEN, B.; LEENAARS, J. G. B.; WALSH, M. G.; GONZALEZ, M. R.. SoilGrids1km — Global Soil Information Based on Automated Mapping. **PLOS ONE**, vol. 9, no. 8, p. e105992, 29 Aug. 2014. DOI 10.1371/journal.pone.0105992. Available at: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0105992. Accessed on: 21 Nov. 2024.

HENRYS, P. A.; MONDAIN-MONVAL, T. O.; JARVIS, S. G. Adaptive sampling in ecology: Key challenges and future opportunities. **Methods in Ecology and Evolution**, vol. 15, no. 9, p. 1483–1496, 2024. DOI 10.1111/2041-210X.14393. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14393. Accessed on: 10 Jan. 2025.

HORST-HEINEN, T. Z.; DALMOLIN, R. S. D.; TEN CATEN, A.; MOURA-BUENO, J. M.; GRUNWALD, S.; PEDRON, F. A.; RODRIGUES, M. F. ROSIN, N. A.; DA SILVA-SANGOI, D. V. Soil depth prediction by digital soil mapping and its impact in pine forestry productivity in South Brazil. **Forest Ecology and Management**, vol. 488, p. 118983, 15 May 2021. DOI 10.1016/j.foreco.2021.118983. Available at: https://www.sciencedirect.com/science/article/pii/S0378112721000724. Accessed on: 30 Aug. 2024.

HUANG, J.; MCBRATNEY, A. B.; MINASNY, B.; TRIANTAFILIS, J. Monitoring and modelling soil water dynamics using electromagnetic conductivity imaging and the ensemble Kalman filter. **Geoderma**, vol. 285, p. 76–93, Jan. 2017. DOI 10.1016/j.geoderma.2016.09.027. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0016706116304931. Accessed on: 3 Sep. 2024.

IBGE. **Base Cartográfica Contínua do Estado do Rio de Janeiro, escala 1:25.000**. Rio de Janeiro, Brazil: IBGE, Coordenação de Cartografia, 2018a.

IBGE. **Mapeamento de recurso naturais do Brasil escala 1:250.000**. Rio de Janeiro, Brazil: Coordenação de Recursos Naturais e Estudos Ambientais, 2018b.

J. C. FAVROT. Pour une approche raisoneé du drainage agricole en France. La méthode des secteurs de référence. **C.R. Académie d'Agriculture de France**, , p. 716–723, 1981. .

JEAN-MARC ROBBEZ-MASSON. **Reconnaissance et délimitation de motifs d'organisation spatiale. Application à la cartographie des pédopaysages.** 1994. 192 f. Thesis – École Nationale Supérieure Agronomique de Montpellier, Montepellier, 1994.

JENNY, H. **Factors of soil formation: a system of quantitative pedology**. Foreword by Ronald Amundson. Originally published: New York : McGraw-Hill, 1941. With new foreword. Includes bibliographical references and index. New York, USA: Dover Publications, 1994(Dover Books on Earth Sciences).

JI, W.; ADAMCHUK, V. I.; CHEN, S.; MAT SU, A. S.; ISMAIL, A.; GAN, Q.; SHI, Z.; BISWAS, Asim. Simultaneous measurement of multiple soil properties through proximal sensor data fusion: A case study. **Geoderma**, vol. 341, p. 111–128, 1 May 2019. DOI 10.1016/j.geoderma.2019.01.006. Available at: https://www.sciencedirect.com/science/article/pii/S0016706117311217. Accessed on: 18 Aug. 2024.

JUNIOR, P. R. P. R.; NASCIMENTO, Flávio Rodrigues Do. Environment, geology-geomorphology and water availability in the Guandu river basin/Rio de Janeiro. **William Morris Davis - Revista de Geomorfologia**, vol. 3, no. 1, p. 1–20, 26 Jul. 2022. DOI 10.48025/ISSN2675-6900.v3n1.2022.147. Available at: //williammorrisdavis.uvanet.br/index.php/revistageomorfologia/article/view/147. Accessed on: 12 Dec. 2024.

KESKIN, H.; GRUNWALD, S.; HARRIS, W. G. Digital mapping of soil carbon fractions with machine learning. **Geoderma**, vol. 339, p. 40–58, 1 Apr. 2019. DOI 10.1016/j.geoderma.2018.12.037. Available at: https://www.sciencedirect.com/science/article/pii/S001670611732030X. Accessed on: 30 Aug. 2024.

KHALEDIAN, Y.; MILLER, B. A. Selecting appropriate machine learning methods for digital soil mapping. **Applied Mathematical Modelling**, vol. 81, p. 401–418, May 2020. DOI 10.1016/j.apm.2019.12.016. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0307904X19307565. Accessed on: 23 Dec. 2024.

KHOMUTININ, Y.; FESENKO, S.; LEVCHUK, S.; ZHEBROVSKA, K.; KASHPAROV, V. Optimising sampling strategies for emergency response: Soil sampling. **Journal of Environmental Radioactivity**, vol. 222, p. 106344, 1 Oct. 2020. DOI 10.1016/j.jenvrad.2020.106344. Available at: https://www.sciencedirect.com/science/article/pii/S0265931X19310288. Accessed on: 21 Nov. 2024.

KIM, S.; HEO, S. An agricultural digital twin for mandarins demonstrates the potential for individualized agriculture. **Nature Communications**, vol. 15, no. 1, p. 1561, 20 Feb. 2024. DOI 10.1038/s41467-024-45725-x. Available at: https://www.nature.com/articles/s41467-024-45725-x. Accessed on: 18 Aug. 2024.

LAGACHERIE, P.; LEGROS, J. P.; BURROUGH, P. A. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. **Geoderma**, vol.

65, no. 3–4, p. 283–301, 1995. DOI 10.1016/0016-7061(94)00040-H. Available at: https://linkinghub.elsevier.com/retrieve/pii/001670619400040H. Accessed on: 22 Sep. 2024.

LAGACHERIE, P.; MCBRATNEY, A. B.; VOLTZ, M. (Eds.). **Digital soil mapping: an introductory perspective**. 1st ed. Amsterdam ; Boston: Elsevier, 2007(Developments in soil science, v. 31).

LAGACHERIE, P.; ROBBEZ-MASSON, J. M.; NGUYEN-THE, N.; BARTHÈS, J. P. Mapping of reference area representativity using a mathematical soilscape distance. **Geoderma**, vol. 101, no. 3–4, p. 105–118, 2001. DOI 10.1016/S0016-7061(00)00101-4. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0016706100001014. Accessed on: 22 Sep. 2024.

LAGACHERIE, P; VOLTZ, M. Predicting soil properties over a region using sample information from a mapped reference area and digital elevation data: a conditional probability approach. **Geoderma**, vol. 97, no. 3–4, p. 187–208, 2000. DOI 10.1016/S0016-7061(00)00038-0. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0016706100000380. Accessed on: 22 Sep. 2024.

LAGACHERIE, P.; LEDREUX, C.; LEGROS, J. P. Modélisation de la connaissance d'un pédologue cartographe. **Mappemonde**, vol. 32, no. 4, p. 12–13, 1993. DOI 10.3406/mappe.1993.1094. Available at: https://www.persee.fr/doc/mappe_0764-3470_1993_num_32_4_1094. Accessed on: 22 Sep. 2024.

LAPIERRE, G. D. J.; IRIZARRY, N. D. M.; ANDREU, M. G. Florida Soil Series and Natural Community Associations: FOR384 FR455, 5 2022. **EDIS**, vol. 2022, no. 3, 2 Jun. 2022. DOI 10.32473/edis-fr455-2022. Available at: https://journals.flvc.org/edis/article/view/129590. Accessed on: 12 Dec. 2024.

LIAW, Andy; WIENER, Matthew. Classification and Regression by randomForest. **R News**, vol. 2, no. 3, p. 18–22, 2002. Available at: https://CRAN.R-project.org/doc/Rnews/.

LIMA, L. A. S.; NEUMANN, M. R. B.; REATTO, A.; ROIG, H. L. **Mapeamento de solos do tradicional ao digital**, n. M386. Planaltina, DF, Brasil: Embrapa Cerrados, 2013.

LOPES, L. C. M.; MARIANO-NETO, E.; AMORIM, A. M. Can soil types explain species distributions? Evaluating the woody understory component of a tropical forest in Brazil. **Brazilian Journal of Botany**, vol. 39, no. 1, p. 251–259, 1 Mar. 2016. DOI 10.1007/s40415-015-0235-x. Available at: https://doi.org/10.1007/s40415-015-0235-x. Accessed on: 23 Jan. 2025.

M. BORNAND; FAVROT, J. C. Cartographie des sols et gestion de l'eau, depuis l'échelle régionale jusqu'a l'échelon parcellaire: l'exemple en France du Languedoc-Roussillon. **Bulletin du Reseau Erosion**, vol. 18, 1998. .

MA, T.; BRUS, D. J.; ZHU, A.; ZHANG, L.; SCHOLTEN, T.. Comparison of conditioned Latin hypercube and feature space coverage sampling for predicting soil classes using simulation from soil maps. **Geoderma**, vol. 370, p. 114366, 1 Jul. 2020. DOI 10.1016/j.geoderma.2020.114366. Available at: https://www.sciencedirect.com/science/article/pii/S001670612030135X. Accessed on: 2 Jan. 2025.

MAIA, R.; BEZERRA, F. The Geomorphology of the Northeast: Classical and Current Perspectives. In: MAIA, R.; BEZERRA, F. (eds.). **Structural Geomorphology in Northeastern Brazil**. Cham: Springer International Publishing, 2020. p. 31–40. DOI 10.1007/978-3-030-13311-5_3. Available at: https://doi.org/10.1007/978-3-030-13311-5_3. Accessed on: 23 Jan. 2025.

MALLAVAN, B. P.; MINASNY, B.; MCBRATNEY, A. B. Homosoil, a Methodology for Quantitative Extrapolation of Soil Information Across the Globe. In: BOETTINGER, J. L.; HOWELL, David W.; MOORE, A. C.; HARTEMINK, A. E.; KIENAST-BROWN, S. (eds.). **Digital Soil Mapping**. Dordrecht: Springer Netherlands, 2010. p. 137–150. DOI 10.1007/978-90-481-8863-5_12. Available at: http://link.springer.com/10.1007/978-90-481-8863-5_12. Accessed on: 4 Sep. 2024.

MALONE, B. P.; MINANSY, B.; BRUNGARD, C. Some methods to improve the utility of conditioned Latin hypercube sampling. **PeerJ**, vol. 7, p. e6451, 25 Feb. 2019. DOI 10.7717/peerj.6451. Available at: https://peerj.com/articles/6451. Accessed on: 29 Dec. 2024.

MALONE, B. P.; MINASNY, B.; MCBRATNEY, A. B. **Using R for Digital Soil Mapping**. Cham: Springer International Publishing, 2017(Progress in Soil Science). DOI 10.1007/978-3-319-44327-0. Available at: http://link.springer.com/10.1007/978-3-319-44327-0. Accessed on: 12 Feb. 2025.

MALONE, B. P.; STYC, Q.; MINASNY, B.; MCBRATNEY, A. B. Digital soil mapping of soil carbon at the farm scale: A spatial downscaling approach in consideration of measured and uncertain data. **Geoderma**, vol. 290, p. 91–99, 15 Mar. 2017. DOI 10.1016/j.geoderma.2016.12.008. Available at: https://www.sciencedirect.com/science/article/pii/S0016706116309922. Accessed on: 23 Dec. 2024.

MARQUES, J. A.; COSTA, P. G.; MARANGONI, L. F. B.; PEREIRA, C. M.; ABRANTES, D. P.; CALDERON, E. N.; CASTRO, C. B.; BIANCHINI, A. Environmental health in southwestern Atlantic coral reefs: Geochemical, water quality and ecological indicators. **Science of The Total Environment**, vol. 651, p. 261–270, 15 Feb. 2019. DOI 10.1016/j.scitotenv.2018.09.154. Available at: https://www.sciencedirect.com/science/article/pii/S0048969718335988. Accessed on: 23 Jan. 2025.

MCBRATNEY, A. B; MENDONÇA SANTOS, M. L; MINASNY, B. On digital soil mapping. **Geoderma**, vol. 117, no. 1, p. 3–52, 2003. DOI 10.1016/S0016-7061(03)00223-4. Available at: https://www.sciencedirect.com/science/article/pii/S0016706103002234. Accessed on: 22 Sep. 2024.

MCBRATNEY, A. B.; WEBSTER, R. Spatial dependence and classification of the soil along a transect in northeast Scotland. **Geoderma**, vol. 26, p. 63–82, 1 Jul. 1981. DOI 10.1016/0016-7061(81)90076-8. Available at: https://www.sciencedirect.com/science/article/pii/0016706181900768. Accessed on: 25 Nov. 2024.

MCBRATNEY, A. B.; MINASNY, B.; STOCKMANN, U. (Eds.). **Pedometrics**. Cham: Springer International Publishing, 2018(Progress in Soil Science). DOI 10.1007/978-3-319-

63439-5. Available at: http://link.springer.com/10.1007/978-3-319-63439-5. Accessed on: 27 Aug. 2024.

MCBRATNEY, A. B.; ODEH, I. O. A.; BISHOP, T. F. A.; DUNBAR, M. S.; S., T. M. An overview of pedometric techniques for use in soil survey. **Geoderma**, vol. 97, no. 3, p. 293–327, 1 Sep. 2000. DOI 10.1016/S0016-7061(00)00043-4. Available at: https://www.sciencedirect.com/science/article/pii/S0016706100000434. Accessed on: 25 Nov. 2024.

MEYER, H.; PEBESMA, E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. **Methods in Ecology and Evolution**, vol. 12, no. 9, p. 1620–1633, 2021. DOI 10.1111/2041-210X.13650. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13650. Accessed on: 21 Nov. 2024.

MINASNY, B.; MCBRATNEY, A. B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. **Computers & Geosciences**, vol. 32, no. 9, p. 1378–1388, Nov. 2006. DOI 10.1016/j.cageo.2005.12.009. Available at: https://linkinghub.elsevier.com/retrieve/pii/S009830040500292X. Accessed on: 30 Sep. 2024.

MOMOLI, R. S.; COOPER, M. Water erosion on cultivated soil and soil under riparian forest. **Pesquisa Agropecuaria Brasileira**, vol. 51, no. 9, p. 1295–1305, 2016. https://doi.org/10.1590/S0100-204X2016000900029.

MOREIRA, Fattima M. S.; SIQUEIRA, J. O.; BRUSSAARD, L. **Soil biodiversity in Amazonian and other Brazilian ecosystems**. Wallingford: CABI publ, 2006.

MOURA, D. M. B.; OLIVEIRA, I. J.; NASCIMENTO, D. T. F.; SOUSA, F. A. Refinamento do mapa de solos da alta bacia hidrográfica do Ribeirão Santa Marta, estado de Goiás, Brasil. **Caderno de Geografia**, vol. 30, no. 62, p. 865, 3 Aug. 2020. DOI 10.5752/P.2318-2962.2020v30n62p865. Available at: http://periodicos.pucminas.br/index.php/geografia/article/view/22936. Accessed on: 23 Dec. 2024.

MOURA, R. L.; AMADO-FILHO, G. M.; MORAES, F. C.; BRASILEIRO, P. S.; SALOMON, P. S.; MAHIQUES, M. M.; BASTOS, A. C.; ALMEIDA, M. G.; SILVA, J. M.; ARAUJO, B. F.; BRITO, F. P.; RANGEL, T. P.; OLIVEIRA, B. C. V.; BAHIA, R. G.; PARANHOS, R. P.; DIAS, R. J. S.; SIEGLE, E.; FIGUEIREDO, A. G.; PEREIRA, R. C.; THOMPSON, F. L. An extensive reef system at the Amazon River mouth. **Science Advances**, vol. 2, no. 4, p. e1501252, 22 Apr. 2016. DOI 10.1126/sciadv.1501252. Available at: https://www.science.org/doi/10.1126/sciadv.1501252. Accessed on: 23 Jan. 2025.

NAUMAN, T. W.; KIENAST-BROWN, S.; ROECKER, S. M.; BRUNGARD, C.; WHITE, D.; PHILIPPE, J.; THOMPSON, J. A. Soil landscapes of the United States (SOLUS): Developing predictive soil property maps of the conterminous United States using hybrid training sets. **Soil Science Society of America Journal**, vol. 88, no. 6, p. 2046–2065, 2024. DOI 10.1002/saj2.20769. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/saj2.20769. Accessed on: 23 Dec. 2024.

NAWAR, S.; CORSTANJE, R.; HALCRO, G.; MULLA, D.; MOUAZEN, A. M. Delineation of Soil Management Zones for Variable-Rate Fertilization. **Advances in Agronomy**. [*S. l.*]:

Elsevier, 2017. vol. 143, p. 175–245. DOI 10.1016/bs.agron.2017.01.003. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0065211317300032. Accessed on: 18 Aug. 2024.

NEIVA, H.; DA SILVA, M.; CARDOSO, C. Analysis of Climate Behavior and Land Use in the City of Rio de Janeiro, RJ, Brazil. **Climate**, vol. 5, no. 3, p. 52, 14 Jul. 2017. DOI 10.3390/cli5030052. Available at: https://www.mdpi.com/2225-1154/5/3/52. Accessed on: 26 Dec. 2024.

NENKAM, A. M.; WADOUX, A. M. J. C.; MINASNY, B.; MCBRATNEY, A. B.; TRAORE, P. C. S.; FALCONNIER, G. N.; WHITBREAD, A. M. Using homosoils for quantitative extrapolation of soil mapping models. **European Journal of Soil Science**, vol. 73, no. 5, p. e13285, Sep. 2022. DOI 10.1111/ejss.13285. Available at: https://bsssjournals.onlinelibrary.wiley.com/doi/10.1111/ejss.13285. Accessed on: 4 Sep. 2024.

NENKAM, A. M.; WADOUX, A. M. J. C.; MINASNY, B.; MCBRATNEY, A. B.; TRAORE, P. C. S.; WHITBREAD, A. M. Using homosoils to enrich sparse soil data infrastructure: An example from Mali. **CATENA**, vol. 223, p. 106862, Apr. 2023. DOI 10.1016/j.catena.2022.106862. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0341816222008487. Accessed on: 4 Sep. 2024.

NEYESTANI, M. SARMADIAN, F.; JAFARI, A.; KESHAVARZI, A.; SHARIFIFAR, A. Digital mapping of soil classes using spatial extrapolation with imbalanced data. **Geoderma Regional**, vol. 26, p. e00422, Sep. 2021. DOI 10.1016/j.geodrs.2021.e00422. Available at: https://linkinghub.elsevier.com/retrieve/pii/S2352009421000675. Accessed on: 4 Jan. 2025.

ODGERS, N. P.; MCBRATNEY, A. B.; CARRÉ, F. Soil Profile Classes. In: MCBRATNEY, A. B.; MINASNY, B.; STOCKMANN, U. (eds.). **Pedometrics**. Cham: Springer International Publishing, 2018. p. 265–288. DOI 10.1007/978-3-319-63439-5_9. Available at: https://doi.org/10.1007/978-3-319-63439-5_9. Accessed on: 10 Jan. 2025.

OWENS, P. R.; RUTLEDGE, E. M. MORPHOLOGY. In: HILLEL, D. (ed.). **Encyclopedia of Soils in the Environment**. Oxford: Elsevier, 2005. p. 511–520. DOI 10.1016/B0-12-348530-4/00002-3. Available at: https://www.sciencedirect.com/science/article/pii/B0123485304000023. Accessed on: 19 Jan. 2025.

PADARIAN, J.; MINASNY, B.; MCBRATNEY, A. B. Using deep learning for digital soil mapping. **SOIL**, vol. 5, no. 1, p. 79–89, 2019. DOI 10.5194/soil-5-79-2019. Available at: https://soil.copernicus.org/articles/5/79/2019/. Accessed on: 1 Oct. 2024.

PELADARINOS, N.; PIROMALIS, D.; CHEIMARAS, V.; TSEREPAS, E.; MUNTEANU, R. A.; PAPAGEORGAS, P. Enhancing Smart Agriculture by Implementing Digital Twins: A Comprehensive Review. **Sensors**, vol. 23, no. 16, p. 7128, 11 Aug. 2023. DOI 10.3390/s23167128. Available at: https://www.mdpi.com/1424-8220/23/16/7128. Accessed on: 24 Feb. 2025.

PEREIRA, M. G.; ANJOS, L. H. C. Formas extraíveis de ferro em solos do estado do Rio de Janeiro. **Revista Brasileira de Ciência do Solo**, vol. 23, no. 2, p. 371–382, Jun. 1999. DOI 10.1590/S0100-06831999000200020. Available at:

http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-06831999000200020&lng=pt&tlng=pt. Accessed on: 25 Nov. 2024.

PEREIRA, M. G.; ANJOS, L. H. C.; NETO, E. C. S.; JUNIOR, C. R. P. **Solos do Rio de Janeiro - Gênese, classificação e limitações ao uso agrícola**. 1st ed. [*S. l.*]: Atena Editora, 2023. DOI 10.22533/at.ed.273232510. Available at: https://atenaeditora.com.br/catalogo/ebook/solos-do-rio-de-janeiro-genese-classificacao-e-limitacoes-ao-uso-agricola. Accessed on: 22 Feb. 2025.

PEREIRA, P. H. C.; LIMA, G. V.; PONTES, A. V. F.; CÔRTES, L. G. F.; GOMES, E.; SAMPAIO, Cláudio L. S.; PINTO, T. K.; MIRANDA, R. J.; CARDOSO, A. T. C.; ARAUJO, J. C.; SEOANE, José Carlos Sícoli. Unprecedented Coral Mortality on Southwestern Atlantic Coral Reefs Following Major Thermal Stress. **Frontiers in Marine Science**, vol. 9, 20 May 2022. DOI 10.3389/fmars.2022.725778. Available at: https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2022.725778/full. Accessed on: 23 Jan. 2025.

PINHEIRO JUNIOR, C. R.; PEREIRA, M. G.; AZEVEDO, A. C.; VAN HUYSSTEEN, C.; ANJOS, L. H. C.; FONTANA, A.; SILVA NETO, E. C.; VIEIRA, J. N.; SANTOS, T. G. Genesis and classification of carbonate soils in the State of Rio de Janeiro, Brazil. **Journal of South American Earth Sciences**, vol. 108, p. 103183, 1 Jun. 2021. DOI 10.1016/j.jsames.2021.103183. Available at: https://www.sciencedirect.com/science/article/pii/S0895981121000304. Accessed on: 26 Dec. 2024.

POGGIO, L.; DE SOUSA, L. M.; BATJES, N. H.; HEUVELINK, G. B. M.; KEMPEN, B.; RIBEIRO, E.; ROSSITER, D. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. **SOIL**, vol. 7, no. 1, p. 217–240, 14 Jun. 2021. DOI 10.5194/soil-7-217-2021. Available at: https://soil.copernicus.org/articles/7/217/2021/. Accessed on: 21 Nov. 2024.

PURCELL, W.; NEUBAUER, T.. Digital Twins in Agriculture: A State-of-the-art review. **Smart Agricultural Technology**, vol. 3, p. 100094, Feb. 2023. DOI 10.1016/j.atech.2022.100094. Available at: https://linkinghub.elsevier.com/retrieve/pii/S2772375522000594. Accessed on: 24 Feb. 2025.

PYLIANIDIS, C.; OSINGA, S.; ATHANASIADIS, I. N. Introducing digital twins to agriculture. **Computers and Electronics in Agriculture**, vol. 184, p. 105942, May 2021. DOI 10.1016/j.compag.2020.105942. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0168169920331471. Accessed on: 24 Feb. 2025.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2024. Available at: https://www.R-project.org/.

RODRIGUES, F. A.; BRAMLEY, R. G. V.; GOBBETT, D. L. Proximal soil sensing for Precision Agriculture: Simultaneous use of electromagnetic induction and gamma radiometrics in contrasting soils. **Geoderma**, vol. 243–244, p. 183–195, 1 Apr. 2015. DOI 10.1016/j.geoderma.2015.01.004. Available at: https://www.sciencedirect.com/science/article/pii/S0016706115000051. Accessed on: 15 Aug. 2024.

RODRIGUES, H. M.; CEDDIA, M. B.; TASSINARI, W.; VASQUES, G. M.; BRANDÃO, Z. N.; MORAIS, J. P. S.; OLIVEIRA, R. P.; NEVES, M. L.; TAVARES, S. R. L. Remote and Proximal Sensors Data Fusion: Digital Twins in Irrigation Management Zoning. **Sensors**, vol. 24, no. 17, p. 5742, 4 Sep. 2024. DOI 10.3390/s24175742. Available at: https://www.mdpi.com/1424-8220/24/17/5742. Accessed on: 22 Sep. 2024.

ROSSETTI, D. F. Evolução sedimentar miocênica nos estados do Pará e Maranhão. **Geologia USP. Série Científica**, vol. 6, no. 2, p. 7–18, 1 Oct. 2006. DOI 10.5327/S1519-874X2006000300003. Available at: https://www.revistas.usp.br/guspsc/article/view/27420. Accessed on: 23 Jan. 2025.

ROUDIER, P. **clhs: a R package for conditioned Latin hypercube sampling.** [*S. l.*: *s. n.*], 2011.

SANTOS, H. G.; JÚNIOR, W. C.; DART, R. O.; ÁGLIO, M. L. D.; SOUSA, J S; PARES, J. G.; FONTANA, A.; MARTINS, A. L. S.; OLIVEIRA, A. P. **O novo mapa de solos do Brasil: legenda atualizada.** Rio de Janeiro, Brazil: Embrapa Solos, 2011(Documentos 130, Documentos 130). Available at: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/123772/1/DOC-130-O-novo-mapa-de-solos-do-Brasil.pdf. Accessed on: 25 Nov. 2024.

SAURETTE, D. D.; BISWAS, A.; HECK, R. J.; GILLESPIE, A. W.; BERG, A. A. Determining minimum sample size for the conditioned Latin hypercube sampling algorithm. **Pedosphere**, vol. 34, no. 3, p. 530–539, 1 Jun. 2024. DOI 10.1016/j.pedsph.2022.09.001. Available at: https://www.sciencedirect.com/science/article/pii/S1002016022000868. Accessed on: 26 Dec. 2024.

SAURETTE, D. D.; HECK, R. J.; GILLESPIE, A. W.; BERG, A. A.; BISWAS, A. Divergence metrics for determining optimal training sample size in digital soil mapping. **Geoderma**, vol. 436, p. 116553, Aug. 2023. DOI 10.1016/j.geoderma.2023.116553. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0016706123002306. Accessed on: 26 Dec. 2024.

SAURETTE, D. D.; HECK, R. J.; GILLESPIE, A. W.; BERG, A. A.; BISWAS, A. Sample Size Optimization for Digital Soil Mapping: An Empirical Example. **Land**, vol. 13, no. 3, p. 365, 14 Mar. 2024. DOI 10.3390/land13030365. Available at: https://www.mdpi.com/2073-445X/13/3/365. Accessed on: 24 Dec. 2024.

SCARPELLI, W.; HORIKAVA, E. H. Chromium, iron, gold and manganese in Amapá and northern Pará, Brazil. **Brazilian Journal of Geology**, vol. 48, p. 415–433, Sep. 2018. DOI https://doi.org/10.1590/2317-4889201820180026. Available at: https://www.scielo.br/j/bjgeo/a/4gJXnwJ8bG4db5xgwq739Cj/?lang=en. Accessed on: 23 Jan. 2025.

SCARPELLI, W.; HORIKAVA, É. H.. Gold, iron and manganese in central Amapá, Brazil. **Brazilian Journal of Geology**, vol. 47, no. 4, p. 703–721, Dec. 2017. DOI 10.1590/2317-4889201720170114. Available at: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2317-48892017000400703&lng=en&tlng=en. Accessed on: 23 Jan. 2025.

SELLARDS, E. H. Geology of Florida. **The Journal of Geology**, vol. 27, no. 4, p. 286–302, May 1919. DOI 10.1086/622662. Available at: https://www.journals.uchicago.edu/doi/10.1086/622662. Accessed on: 12 Dec. 2024.

SENA, N. C.; VELOSO, G. V.; LOPES, A. O.; FRANCELINO, M. R.; FERNANDES-FILHO, E. I.; SENRA, E. O.; SILVA FILHO, L. A.; CONDÉ, V. F.; SILVA, D. L. A.; ARAÚJO, R. W. Soil sampling strategy in areas of difficult access using the cLHS method. **Geoderma Regional**, vol. 24, p. e00354, 1 Mar. 2021. DOI 10.1016/j.geodrs.2020.e00354. Available at: https://www.sciencedirect.com/science/article/pii/S2352009420301036. Accessed on: 26 Dec. 2024.

SHADDAD, S. M.; MADRAU, S.; CASTRIGNANÒ, A.; MOUAZEN, A. M. Data fusion techniques for delineation of site-specific management zones in a field in UK. **Precision Agriculture**, vol. 17, no. 2, p. 200–217, Apr. 2016. DOI 10.1007/s11119-015-9417-6. Available at: http://link.springer.com/10.1007/s11119-015-9417-6. Accessed on: 18 Aug. 2024.

SILVA, M. D. S. D.; BARRETO-GARCIA, P. A. B.; MONROE, P. H. M.; PEREIRA, M. G.; PINTO, L. A. D. S. R.; NUNES, M. R. Physically protected carbon stocks in a Brazilian Oxisol under homogeneous forest systems. **Geoderma Regional**, vol. 40, 2025. https://doi.org/10.1016/j.geodrs.2024.e00915.

SOARES, M. O.; ROSSI, S.; GURGEL, A. R.; LUCAS, C. C.; TAVARES, T. C. L.; DINIZ, B.; FEITOSA, C. V.; RABELO, E. F.; PEREIRA, P. H. C.; KIKUCHI, R. K. P.; LEÃO, Z. M. A. N.; CRUZ, I. C. S.; CARNEIRO, P. B. M.; ALVAREZ-FILIP, L. Impacts of a changing environment on marginal coral reefs in the Tropical Southwestern Atlantic. **Ocean & Coastal Management**, vol. 210, p. 105692, 1 Sep. 2021. DOI 10.1016/j.ocecoaman.2021.105692. Available at: https://www.sciencedirect.com/science/article/pii/S0964569121001769. Accessed on: 23 Jan. 2025.

SOIL SURVEY STAFF, NATURAL RESOURCES CONSERVATION SERVICE. Web Soil Survey. 2016. Available at: http://websoilsurvey.nrcs.usda.gov. Accessed on: 23 Dec. 2024.

STUMPF, F.; SCHMIDT, K.; BEHRENS, T.; SCHÖNBRODT-STITT, S.; BUZZO, G.; DUMPERTH, C.; WADOUX, A.; XIANG, W.; SCHOLTEN, T.. Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. **Journal of Plant Nutrition and Soil Science**, vol. 179, no. 4, p. 499–509, 2016. DOI 10.1002/jpln.201500313. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/jpln.201500313. Accessed on: 29 Dec. 2024.

TEEGAVARAPU, R. S. V.; SHARMA, P. J. Influences of climate variability on regional precipitation and temperature associations. **Hydrological Sciences Journal**, vol. 66, no. 16, p. 2395–2414, 10 Dec. 2021. DOI 10.1080/02626667.2021.1994976. Available at: https://doi.org/10.1080/02626667.2021.1994976. Accessed on: 23 Jan. 2025.

TEN CATEN, A.; DALMOLIN, R. S. D.; PEDRON, F. A.; SANTOS, M. L. M. Extrapolação das relações solo-paisagem a partir de uma área de referência. **Ciência Rural**, vol. 41, p. 812–816, May 2011. DOI 10.1590/S0103-84782011000500012. Available at: https://www.scielo.br/j/cr/a/zbfLpXgYxGsWqkdhk6w7Xsc/?lang=pt. Accessed on: 30 Sep. 2024.

TERUIYA, R. K.; PARADELLA, W. R.; DOS SANTOS, A. R.; DALL'AGNOL, R.; VENEZIANI, P. Integrating airborne SAR, Landsat TM and airborne geophysics data for improving geological mapping in the Amazon region: the Cigano Granite, Carajás Province, Brazil. **International Journal of Remote Sensing**, vol. 29, no. 13, p. 3957–3974, 1 Jul. 2008. DOI 10.1080/01431160801891838. Available at: https://doi.org/10.1080/01431160801891838. Accessed on: 24 Jan. 2025.

U.S. CENSUS BUREAU. **Primary and secondary roads state-based shapefile**. Department of Commerce - TIGER: [*s. n.*], 2019.

USDA/NRCS. **1981-2010 Annual Average Maximum Temperature by State**. Texas, USA: USDA/NRCS - National Geospatial Center of Excellence, 2012a.

USDA/NRCS. **1981-2010 Annual Average Precipitation by State**. Texas, USA: USDA/NRCS - National Geospatial Center of Excellence, 2012b.

USGS MINERAL RESOURCES. **State geologic maps**. Reston, USA: [*s. n.*], 2017.

VAN WESTEN, C. J.; CASTELLANOS, E.; KURIAKOSE, S. L. Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. **Engineering Geology**, Landslide Susceptibility, Hazard and Risk Zoning for Land Use Planning. vol. 102, no. 3, p. 112–131, 1 Dec. 2008. DOI 10.1016/j.enggeo.2008.03.010. Available at: https://www.sciencedirect.com/science/article/pii/S0013795208001786. Accessed on: 9 Aug. 2024.

VASCONCELOS, B. N. F.; BRAVO, J. V. M.; CUNHA, J. E. F.; FERNANDES-FILHO, E. I. Mapping the Soil Frontiers with Legacy Soil Data: An Approach for Covering the Lack of Updated Reference Maps of Minas Gerais, Brazil. **Anuário do Instituto de Geociências**, vol. 46, 3 May 2023. DOI 10.11137/1982-3908_2023_46_49327. Available at: https://revistas.ufrj.br/index.php/aigeo/article/view/49327. Accessed on: 23 Dec. 2024.

VERDOUW, C.; TEKINERDOGAN, B.; BEULENS, A.; WOLFERT, S. Digital twins in smart farming. **Agricultural Systems**, vol. 189, p. 103046, Apr. 2021. DOI 10.1016/j.agsy.2020.103046. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0308521X20309070. Accessed on: 18 Aug. 2024.

VOLTZ, M.; LAGACHERIE, P.; LOUCHART, X. Predicting soil properties over a region using sample information from a mapped reference area. **European Journal of Soil Science**, vol. 48, no. 1, p. 19–30, 1997. DOI 10.1111/j.1365-2389.1997.tb00181.x. Available at: https://bsssjournals.onlinelibrary.wiley.com/doi/10.1111/j.1365-2389.1997.tb00181.x. Accessed on: 22 Sep. 2024.

WADOUX, A. M. J. C.; HEUVELINK, G. B. M.; LARK, R. M.; LAGACHERIE, P; BOUMA, J.; MULDER, V. L.; LIBOHOVA, Z.; YANG, L.; MCBRATNEY, A. B. Ten challenges for the future of pedometrics. **Geoderma**, vol. 401, p. 11, 2021. DOI 10.1016/j.geoderma.2021.115155. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0016706121002354. Accessed on: 21 Sep. 2024.

WATTS, F. C.; COLLINS, M. E. Formation of The Soils in Florida. **Soils of Florida**. 1st ed. Madison, WI, USA: American Society of Agronomy and Soil Science Society of America, 2008. p. 1–28. DOI 10.2136/2008.soilsofflorida.c1. Available at: http://doi.wiley.com/10.2136/2008.soilsofflorida.c1. Accessed on: 12 Dec. 2024.

YIGINI, Y.; PANAGOS, P. Reference Area Method for Mapping Soil Organic Carbon Content at Regional Scale. **Procedia Earth and Planetary Science**, Geochemistry of the Earth's surface GES-10 Paris France, 18-23 August, 2014. vol. 10, p. 330–338, 1 Jan. 2014. DOI 10.1016/j.proeps.2014.08.028. Available at: https://www.sciencedirect.com/science/article/pii/S1878522014000903. Accessed on: 21 Nov. 2024.

ZHANG, L.; ZHU, A. X.; LIU, J.; MA, T.; YANG, L.; ZHOU, C. An adaptive uncertainty-guided sampling method for geospatial prediction and its application in digital soil mapping. **International Journal of Geographical Information Science**, vol. 37, no. 2, p. 476–498, 1 Feb. 2023. DOI 10.1080/13658816.2022.2125973. Available at: https://doi.org/10.1080/13658816.2022.2125973. Accessed on: 10 Jan. 2025.

ZHANG, Y.; SAURETTE, D. D.; EASHER, T. H.; JI, W.; ADAMCHUK, V. I.; BISWAS, A. Comparison of sampling designs for calibrating digital soil maps at multiple depths. **Pedosphere**, vol. 32, no. 4, p. 588–601, 1 Aug. 2022. DOI 10.1016/S1002-0160(21)60055-3. Available at: https://www.sciencedirect.com/science/article/pii/S1002016021600553. Accessed on: 26 Dec. 2024.