

A Geophysical and Machine Learning Paradigm for High-Fidelity Digital Soil Mapping: A Case Study in Vietnam

PHAN Thien Huong^{1,*}, NGUYEN Hoang Long¹, NGUYEN Thanh Tung²

¹ Hanoi University of Mining and Geology, Hanoi, Vietnam

² Vietnam Petroleum Institute

* Corresponding email: phanthienhuong@humg.edu.vn

Abstract: *Traditional Digital Soil Mapping (DSM) has long been constrained by its reliance on sparse, point-based soil sampling, a methodology that is both costly and prone to significant interpolation uncertainty, often failing to capture fine-scale soil heterogeneity. This paper presents a modern, integrated approach that leverages the power of high-resolution geophysical surveys to overcome these limitations. We demonstrate a synergistic framework where dense data from Electromagnetic Induction (EMI), Electrical Resistivity (ER), and Gamma-ray Spectrometry are used as spatially continuous covariates. These geophysical proxies are calibrated against targeted physical soil samples using a Random Forest machine learning algorithm to predict key soil properties.*

The success of this methodological framework is demonstrated through the successful construction of a digital map of soil properties in the Nghia Dan area of Nghe An province, Vietnam, a region characterized by complex Feralit soils within a hilly, mountainous landscape. The study successfully generated high-resolution predictive maps of soil texture (sand, silt, and clay content) at three distinct depths (37.5 cm, 75 cm, and 150 cm). The results revealed complex spatial heterogeneity and significant vertical differentiation in the soil profile, patterns that are highly consistent with the known pedological processes of Feralit soil formation in an erosional topography.

The findings confirm that the integration of geophysical methods with machine learning provides a rapid, non-destructive, and economically viable paradigm for producing quantitatively robust and spatially explicit digital soil maps. These high-fidelity maps offer a superior foundation for applications in precision agriculture, and sustainable land management, and the results have been made accessible through a WebGIS platform to support practical decision-making.

Keywords: *Digital soil mapping (DSM), geophysics, machine learning, spatial heterogeneity, proximal soil sensing, precision agriculture*

1. Introduction

A Digital Soil Mapping (DSM) represents a fundamental shift from traditional cartography to a dynamic spatial information system. It is best understood not as a static map, but as a geodatabase of soil properties, where attributes are digitally encoded and managed. This structure allows for sophisticated analysis and integration, making it an indispensable tool across various scientific and economic sectors.

DSM is a tool for resource and environmental management. Digital soil maps are central to the sustainable management of land and environmental resources. For land use planning, they provide detailed spatial information on soil characteristics such as texture (clay, silt, sand content), organic matter, pH, drainage capacity, and depth. Planners can use this data to identify the most suitable areas for agriculture, forestry, urban development, or conservation, thereby minimizing risks like soil erosion, desertification, and pollution. In environmental management, DSM helps pinpoint soils vulnerable to contamination, such as those with a high potential for leaching agricultural chemicals into groundwater. It is also used to monitor soil degradation, assess the impacts of climate change on soil resources, and formulate remediation strategies.

DSM is the foundation for precision agriculture. In modern farming, digital soil maps are the cornerstone of precision agriculture, optimizing production while reducing environmental impact. They enable "zone management," where a field is divided into smaller management zones based on soil variability. Instead of uniform application, inputs like fertilizers, water, and pesticides can be tailored to the specific needs of each zone. This variable-rate application not only reduces costs but also enhances crop yields and protects the environment. Furthermore, DSM provides the critical data needed to select crop varieties that are best suited to the specific soil conditions of a particular area, ensuring optimal growth and productivity.

For the scientific community, digital soil maps are a crucial input for a wide range of predictive models. In hydrology, soil properties like permeability and water-holding capacity, derived from DSM, are essential parameters for modeling surface runoff, flood events, and groundwater recharge. In climate change research, these maps are used to estimate soil organic carbon stocks and model the flux of carbon between the soil and the atmosphere, contributing to our understanding of the global carbon cycle. Agronomic models also rely heavily on DSM, integrating soil data with weather information and management practices to forecast crop yields and assess food security.

The digital nature of DSM facilitates seamless integration with other spatial data layers within a Geographic Information System (GIS). This interoperability allows for powerful, multi-layered analyses. For instance, soil maps can be combined with remote sensing imagery (e.g., from Landsat or Sentinel satellites) to monitor crop health and soil moisture over large regions.

The transition from traditional to digital soil mapping is driven by its significant advantages over traditional paper maps. DSMs are typically created using spatial prediction models (e.g., Kriging) that incorporate field observations and related environmental covariates (like topography and climate), resulting in continuous maps with higher accuracy and spatial resolution. They can be updated efficiently and cost-effectively as new data becomes available. Most importantly, DSM allows for complex quantitative queries and spatial analyses (e.g., clay content of 20-30%), which are impractical with static maps. Finally, data can be visualized in 2D, 3D, or even 4D (to show changes over time), offering deeper insights for users.

We can see, digital soil mapping is not only an information representation tool but also a powerful analytical and predictive system. It has revolutionized the way we understand, manage and use soil resources efficiently and sustainably.

Historically, soil mapping has been predicated on a direct measurement paradigm, involving the physical extraction of soil cores at discrete geographical coordinates. The subsequent laboratory analysis yields highly accurate, quantitative data for those specific points. However, the spatial extrapolation of this point data across a landscape via interpolation algorithms (e.g., Kriging, Inverse Distance Weighting) is subject to considerable uncertainty. This approach is constrained by two primary limitations: first, the prohibitive costs and logistical demands of high-density sampling, and second, the potential for interpolation artifacts to misrepresent or entirely fail to capture fine-scale spatial heterogeneity, which is a crucial determinant of ecosystem function and agricultural productivity.

The construction of digital soil mapping (DSM) is witnessing a revolutionary shift (Searle et al, 2021), from the classical method that relies solely on soil sample analysis to the integration of modern geophysical methods. This is not just an improvement, but a fundamental shift in quality and efficiency. The superiority of geophysics in this field can be analyzed through the following core aspects: Overcoming the limitation of point data, non-destructive method, high economic efficiency, obtaining instant value, using to monitor timely decision making, huge data base helps high technology, digital transformation, and another important thing is that the physical parameters obtained from geophysical methods have a multivariate relationship with soil parameters, so when the data is large enough, combined with AI/ML, the combination of geophysical methods will provide many soil parameters at the same time such as moisture, porosity, soil structure,...

The traditional method is based on collecting soil samples at discrete locations, then analyzing them in the laboratory. Although single-point results are very accurate, they have inherent limitations: 1) Cost and time: Sampling, transporting, and analyzing in the laboratory are expensive and time-consuming; 2) Uncertain interpolation: Maps are created by interpolating data from these points. This process is essentially an “educated guess” about what happens in the space between the sampling points. It can completely miss important but small-scale anomalies or transition zones.

Geophysics has solved this problem by providing data that is either nearly continuous or extremely dense. Instead of a few dozen points per hectare, methods such as Electromagnetic Induction (EMI) (Koganti et al, 2024) or Ground Penetrating Radar (GPR) (Benedetto et al., 2013) can collect thousands of measurement points. This transforms the map from a collection of smoothed discrete points into a high-resolution “photograph” of land variation in space.

To address these critical limitations, this paper introduces and validates an integrated framework that leverages high-resolution geophysical data as spatially continuous covariates, calibrated against targeted soil samples using machine learning. The primary objective is not to conduct an exhaustive soil survey of the Nghe An region, but rather to demonstrate the efficacy, efficiency, and superiority of this modern

workflow. The Nghia Dan area, with its complex Feralit soils, was selected as a challenging and ideal test site to validate our methodology

A combination of geophysical methods used to determine physical parameters to soil parameters based on statistical algorithms (Triantafilis et al, 2001) as well as AI/ML (Huong Phan et al, 2024c) has been published in specialized journals.

2. Material and methods

2.1. Material data collection

Validation of the integrated geophysical-ML framework was conducted at a site selected for its pronounced environmental heterogeneity in Nghia Dan, Nghe An province. This location serves as a rigorous testbed for methodology performance due to its diverse topography, which spans three ecological zones (mountainous, plain, and coastal) and is characterized by two predominant soil types: ferralitic soils in upland areas and fluvial soils in the lowlands. The broader study region in Nghe An is characterized by two predominant soil types: alluvial fluvial soils (~198,139 ha) supporting diverse cropping systems, and extensive ferralitic soils (~383,121 ha) in the highlands primarily used for forestry. While three sites were surveyed to capture this regional variability (Fig. 1), this article will present the results from the Nghia Dan area specifically to exemplify the application and efficacy of the geophysical methods for digital soil property transformation..

Duoc Lieu Nghia Dan Cooperative (ND): Located in the mountainous region, this site has a more varied terrain with altitudes ranging from 49 to 55 meters above sea level. Medicinal plants such as *Celastrus Hindsii* are being grown here.

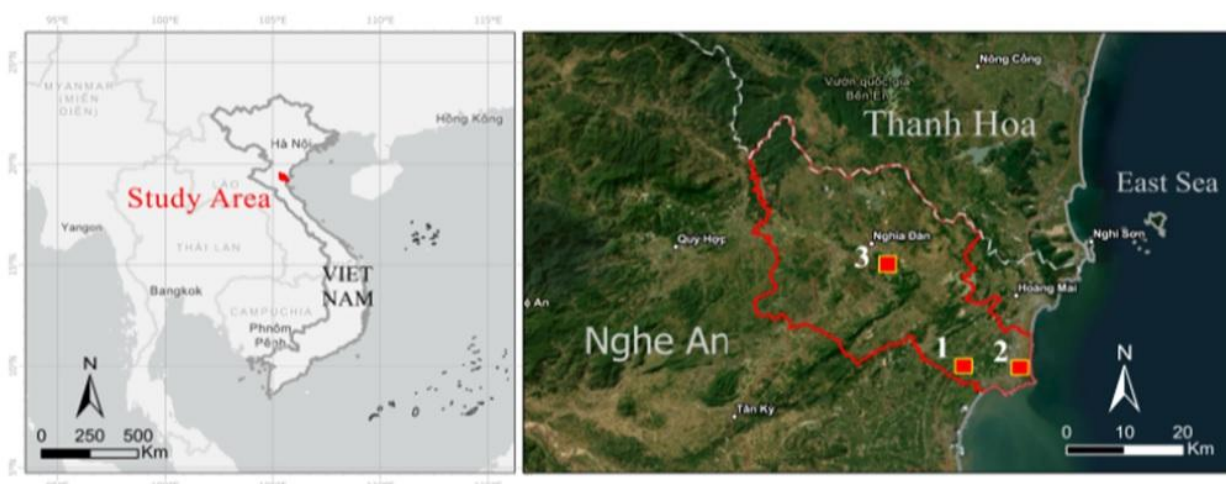


Fig. 1. Location of study areas in Nghe An province. (Red square 3: Nghia Dan Cooperative).

Geophysical data was acquired through three primary methods to serve as spatially continuous covariates for the soil prediction model.

Electromagnetic Induction (EMI): EMI measurements were conducted using an EM38-MK2 instrument. The EM38-MK2 is a fixed-frequency (14.6 kHz) electromagnetic tool with transmitter-receiver coil separations of 0.5 m and 1.0 m. By acquiring data in both the vertical and horizontal dipole modes at these two separations, it provides simultaneous measurements of apparent ground conductivity (ECa) and magnetic susceptibility (In-Phase) with effective depths of investigation corresponding to the target depths of 150 cm, 75 cm, and 37.5 cm. For EMI measurements, with grid 10 m × 10 m data was collected at 1-meter intervals along these lines. At each measurement point, both in-phase and quadrature components were recorded in both the vertical and horizontal directions, totaling 2000 data points

Electrical Resistivity (ER): Two-dimensional (2D) Electrical Resistivity surveys were conducted using a GESKA instrument with a Wenner-Schlumberger array configuration. The survey grid consisted of parallel lines spaced 10 m apart, with an in-line station spacing of 10 m. At each measurement station, data was acquired at 11 different electrode spacings to investigate various depths. From the collected apparent resistivity data, an inversion was performed to solve for the true subsurface resistivity distribution. This process generated a 2D model from which resistivity values at the specific target depths of 150 cm, 75 cm, and 37.5 cm were extracted for use in the prediction model.

Gamma-ray Spectrometry: A Gamma Surveyor Vario was used to measure the natural gamma spectrum and determine the concentrations of Potassium (K), Uranium (U), and Thorium (Th). As this is a passive method primarily sensitive to the near-surface, the results were considered representative of the shallowest soil layer and were thus used as covariates only for the 37.5 cm depth model. Gamma-ray spectrometry measurements were conducted at the grid intersections with a 10-meter spacing in one direction and a 5-meter spacing in the other. Four values (K, U, Th, and total gamma) were obtained at each point, yielding 242 data points.

To calibrate and validate the geophysical data, physical soil samples were collected from 120 designated points within the Nghia Dan site. At each location, the sampling plan targeted three distinct depths: 37.5 cm, 75 cm, and 150 cm. The number of samples retrieved from each point was contingent on the local depth to bedrock; consequently, some locations yielded fewer than the target of three samples due to shallow bedrock. This sampling protocol resulted in a total of 300 soil samples, which were collected for subsequent laboratory analysis.

2.2. Methods

A substantial body of literature has established the efficacy of geophysical methods for characterizing soil properties (Huong Phan et al., 2022). Techniques such as Electromagnetic Induction (EMI) and Electrical Resistivity (ER) measure the soil's apparent electrical conductivity (ECa) and resistivity, respectively. These parameters are intrinsically governed by key edaphic attributes including texture (specifically clay content and mineralogy), moisture content, porosity, and solute concentration (salinity). Concurrently, Gamma Spectra, a passive method, quantifies the activity concentration of naturally occurring radionuclides—namely Potassium (K), Uranium (U), and Thorium (Th)—which exhibit strong correlations with soil mineralogy, parent material, and texture. Furthermore, gamma spectra can be employed as a proxy to indirectly estimate soil moisture and salinity.

It is clear that there is a relationship between the physical parameters obtained from EMI, ER and gamma spectroscopy methods and soil parameters. In this study, a comprehensive geophysical survey was first conducted to generate spatially continuous maps of relevant proxy variables (electrical conductivity, resistivity, radioactive element content). These maps serve as a high-resolution auxiliary dataset, or dependent variable, capturing the underlying spatial structure of soil property variations. The next step is calibration, which determines the relationship between the physical parameters and soil parameters obtained from soil samples collected in the field and analyzed in the laboratory. After establishing the relationship between the physical parameters and soil parameters at the sampling points, a soil property prediction model is built from the relationship function and the physical data obtained over the entire study area. This model, combined with the entire high-density geophysical dataset, produces a final high-resolution predictive map of the desired soil properties.

A significant methodological challenge in this study is establishing a robust transfer function to relate the geophysical measurements to the target soil properties. This challenge arises from the complex, multivariate, and non-unique nature of these relationships. Specifically, a single geophysical parameter such as resistivity is not a function of one soil property, but rather a composite response to multiple factors, which often have competing influences. For example, high soil water content (moisture) tends to decrease resistivity, whereas low dissolved ion concentration (salinity) tends to increase it. Therefore, without a model capable of deconvolving and quantifying these multivariate effects, knowing that a soil has both high moisture and low salinity makes it impossible to definitively predict whether the bulk resistivity will increase or decrease. The same principle applies to conductivity.

This physical complexity is demonstrated statistically in the Pearson correlation matrix for the 37.5 cm depth layer (Fig. 2). The heatmap reveals that the direct correlations between individual geophysical proxies (e.g., R37, Cond_H05) and soil texture components (sand, silt) are weak, with coefficients generally falling below $|r| = 0.3$. The matrix confirms that no single geophysical parameter serves as a strong standalone predictor. Consequently, a multivariate modeling approach, such as the Random Forest algorithm used in this study, is essential to capture these complex interdependencies and accurately predict soil characteristics from the combined geophysical dataset.

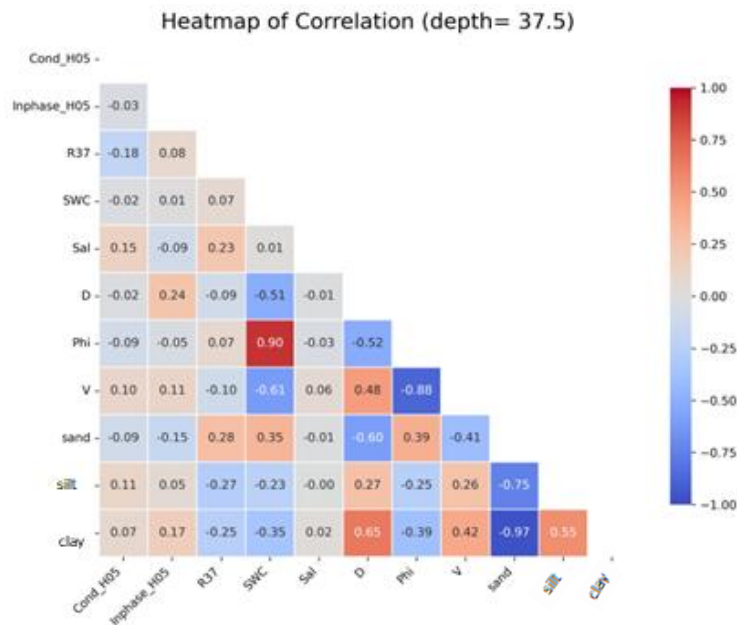


Fig. 2. Heatmap of correlations between parameters from EMI, ER with soil properties at Nghia Dan of depth 37.5 cm.

To resolve this multivariate problem, The Random Forest algorithm was selected for its ability to effectively handle the complex, non-linear relationships among multiple input variables, and because it achieved the lowest prediction error compared to other algorithms. The specifics of this ML approach, which leverages the synergistic effects of multiple physical parameters, are detailed further in Huong Phan et al. (2024c). For the purpose of random forest regression algorithm, the dataset is divided into. a training set (90%) and a testing set (10%). The number of fold K = 10.

Because in this study three Machine Learning (Random Forest) models were built separately: one for predicting sand content, one for predicting silt, and one for predicting clay. Each model was optimized to minimize its own prediction error (e.g., reduce the RMSE for predicting sand) without “knowing” the existence of the other two models. Therefore, there is no guarantee that at any point on the map, the sum of the three independent predicted values (predicted_sand + predicted_limon + predicted_clay) will be exactly 100. This sum could be 98%, 105%, or any other arbitrary number. This creates an unphysical result. A soil sample cannot have a composition that adds up to anything other than 100%. It reduces the reliability and scientific validity of the result. To fix this error, we performed a post-processing normalization step.

Post-processing normalization is a crucial data correction step performed after initial predictions have been made by a machine learning model. In the context of this study, it is necessary to address the issue of compositional data, where the sum of several components (sand, silt, and clay) must equal a whole, or 100%. The normalization workflow corrects this issue through the following steps:

1. Obtain raw predictions: The process begins with the initial, uncorrected predicted values for sand, silt, and clay content for each data point, as generated by the Random Forest algorithm.
2. Calculate the unconstrained sum: For each individual data point, the raw predicted values for sand, silt, and clay are added together. This sum, referred to as texture in the provided data, often deviates from 100 (e.g., 104.57, 98.67).
3. Rescale Each Component: Each raw prediction is then divided by its corresponding unconstrained sum and multiplied by 100 to calculate the final normalized percentage. The formula is:

$$\text{Normalized Value (\%)} = \frac{\text{Raw Predicted Value}}{\text{Unconstrain Sum}} * 100 \quad (1)$$

This calculation is performed independently for the sand, silt, and clay components of each data point.

4. Verify the Final Output: The result of this process is a new, corrected set of values for sand, silt, and clay. When summed together, these new values equal exactly 100.00, ensuring the final dataset is physically consistent and adheres to the closure constraint required for compositional data.

This paper presents the resulting predictive maps for soil texture (sand, silt, and clay content) at three distinct depths (37.5, 75, and 150 cm) for the Nghia Dan site (Figs 3, 4, and 5).

3. Result and discussion

Based on the provided map (Fig. 3), the sand distribution in the Nghia Dan area exhibits significant spatial heterogeneity and remarkable vertical differentiation across the three analyzed depths. The overall pattern suggests a complex interplay between the underlying parent material and soil development processes in the upper layers.

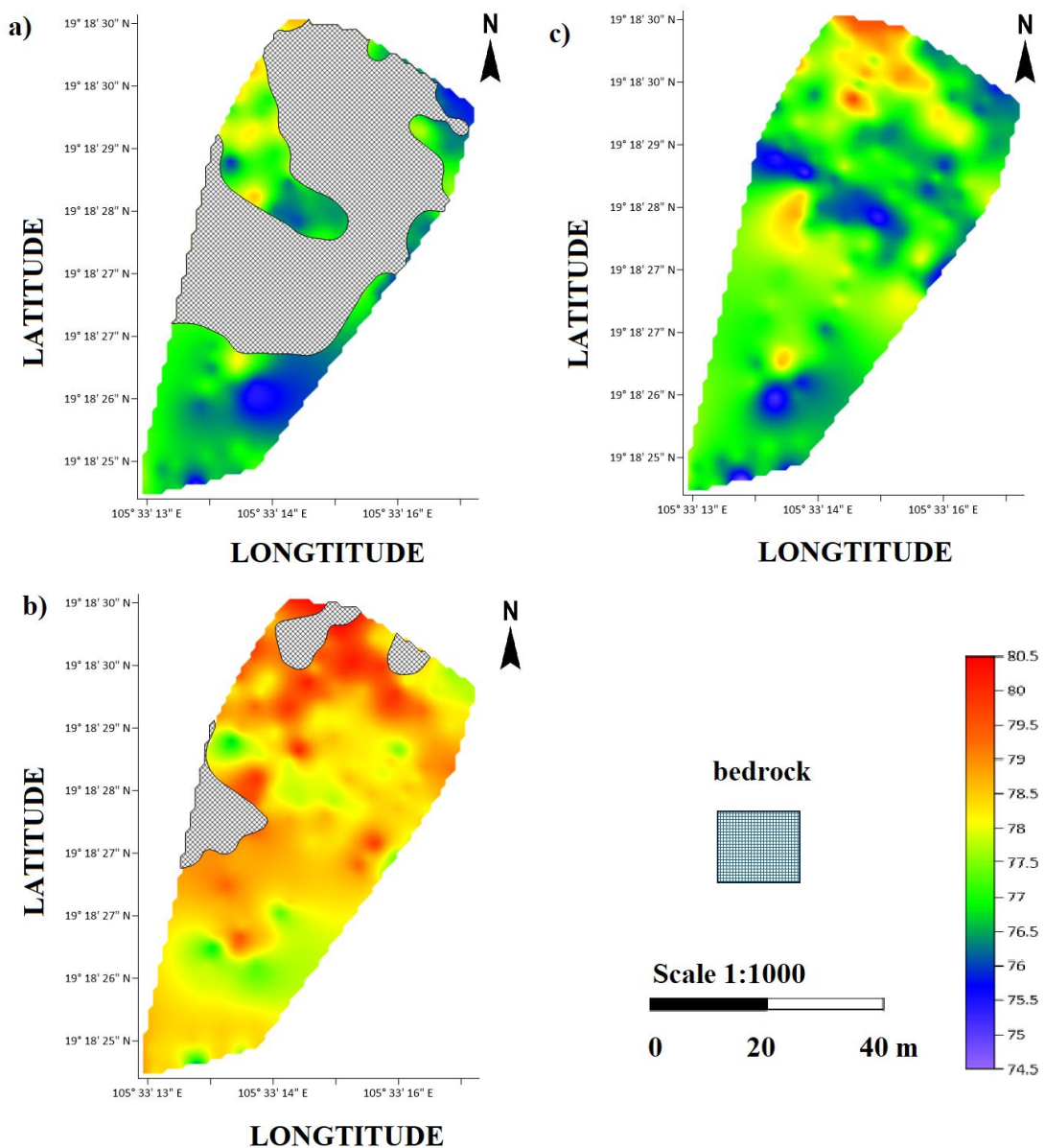


Fig. 3. Predicted sand content at Duoc lieu Cooperative Nghia Dan at a depth a) of 150cm; b) of 75 cm and c) 37.5cm (scale weight percents)

At 150 cm Depth (map a): At this depth, the landscape is heavily dominated by a large, continuous bedrock outcrop that covers the northern and central portions of the survey area. The sand distribution is

therefore spatially constrained to the southern half of the map. In these soil-covered areas, the sand content is relatively low and fairly uniform, with values primarily ranging from approximately 74.5% to 77.5% (weight percent), as indicated by the blue and green colors. The lowest concentrations are notably clustered in the southeastern corner.

At 75 cm Depth (map b): A dramatic transformation in the soil profile is evident at the 75 cm depth. The influence of bedrock is significantly diminished, reduced to smaller, fragmented outcrops in the far north. This map reveals a pronounced spatial zonation across the entire area: a large, prominent zone of high sand content (approximately 79% to 80.5% by weight, shown in orange and red) dominates the west-central region. In stark contrast, the eastern portion of the area is characterized by significantly lower sand content, with values ranging from 75% to 77.5% (weight percent) (green tones). This clear east-west division suggests that different soil characteristics or processes are dominant in these two distinct zones at this depth.

At 37.5 cm Depth (map c): The surface layer at 37.5 cm shows a spatial pattern that is remarkably consistent with the 75 cm depth. No bedrock is present at the surface. The strong zonation persists, with the primary features being. A large area of high sand percentage (predominantly 78% to 80% by weight, yellow and orange colors) continues to occupy the west-central region. The eastern half remains a distinct area of lower sand content (green tones, ~76% to 77.5% by weight). The overall structure is maintained from the 75 cm layer, indicating that the processes generating the upper soil profile are consistent.

The most significant change occurs between the 150 cm and 75 cm depths. This suggests a transition from a deeper soil layer heavily influenced by the parent bedrock to a distinct upper soil horizon with well-defined spatial patterns. The strong similarity between the 75 cm and 37.5 cm maps indicates that the soil-forming, erosional, or depositional processes that created the clear east-west division are dominant and consistent throughout the upper ~75 cm of the soil profile.

Tab. 1. Evaluation test

Detailed metrics

Explained Variance Score 	0.9160 (± 0.061)
Mean Absolute Error (MAE) 	2.791 (± 0.93)
Mean Absolute Percentage Error 	3.79% (± 1.2%)
Mean Squared Error (MSE) 	17.07 (± 12)
Root Mean Squared Error (RMSE) 	4.064 (± 1.5)
Root Mean Squared Logarithmic Error (RMSLE) 	0.05438 (± 0.019)
Pearson coefficient 	0.9575 (± 0.032)
R2 Score 	0.9135 (± 0.061)

Tab. 2. Post-Processing Normalization Workflow for Predicted Soil Texture Data.

Longitude all	Latitude all	sand dudoan	silt dudoan	clay dudoan	texture_te sand (normalization)	silt (normalization)	clay (normalization)	sum	sand error	silt (error)	clay (error)	
105.55458000	19.30786000	81.24	5.71	17.62	104.57	77.69	5.46	16.85	100.00	3.55	0.25	0.77
105.55455000	19.30782000	77.08	5.26	16.33	98.67	78.11	5.33	16.55	100.00	-1.04	-0.07	-0.22
105.55453000	19.30779000	77.60	5.51	17.06	100.16	77.47	5.50	17.03	100.00	0.12	0.01	0.03
105.55451000	19.30776000	77.89	6.01	17.84	101.74	76.56	5.91	17.53	100.00	1.33	0.10	0.31
105.55449500	19.30773000	76.20	6.16	17.76	100.12	76.11	6.15	17.74	100.00	0.09	0.01	0.02
105.55448000	19.30770000	77.20	6.15	17.87	101.23	76.27	6.07	17.66	100.00	0.93	0.07	0.22
105.55444500	19.30766000	76.18	6.03	17.54	99.74	76.37	6.04	17.58	100.00	-0.20	-0.02	-0.05
105.55441000	19.30762000	76.75	5.93	18.54	101.22	75.83	5.85	18.32	100.00	0.93	0.07	0.22
105.55438470	19.30759120	76.41	5.96	18.63	101.00	75.65	5.91	18.44	100.00	0.76	0.06	0.18
105.55435940	19.30756239	75.72	6.01	18.08	99.81	75.86	6.02	18.11	100.00	-0.15	-0.01	-0.03
105.55434230	19.30753688	76.39	5.98	17.55	99.92	76.45	5.98	17.57	100.00	-0.06	0.00	-0.01
105.55432510	19.30751136	78.26	5.58	16.50	100.34	77.99	5.56	16.45	100.00	0.27	0.02	0.06
105.55465740	19.30813212	79.09	6.32	18.10	103.52	76.40	6.11	17.49	100.00	2.69	0.22	0.62
105.55462870	19.30809606	77.07	6.22	17.73	101.02	76.29	6.16	17.55	100.00	0.78	0.06	0.18
105.55460000	19.30806000	78.83	6.34	17.79	102.96	76.57	6.15	17.28	100.00	2.27	0.18	0.51
105.55456000	19.30800500	77.89	6.19	17.90	101.98	76.37	6.07	17.56	100.00	1.52	0.12	0.35
105.55452000	19.30795000	77.49	5.82	17.79	101.11	76.64	5.76	17.60	100.00	0.85	0.06	0.19
105.55449790	19.30791195	78.38	5.75	17.42	101.54	77.19	5.66	17.15	100.00	1.19	0.09	0.26
105.55447590	19.30787391	77.82	5.53	16.72	100.07	77.77	5.53	16.71	100.00	0.05	0.00	0.01
105.55445290	19.30784195	76.56	5.98	17.76	100.30	76.33	5.96	17.71	100.00	0.23	0.02	0.05
105.55443000	19.30781000	78.70	6.20	17.65	102.54	76.74	6.05	17.21	100.00	1.95	0.15	0.44
105.55441430	19.30777863	76.74	6.06	17.78	100.58	76.30	6.02	17.68	100.00	0.44	0.03	0.10
105.55439850	19.30774726	75.81	5.28	16.74	97.83	77.49	5.40	17.11	100.00	-1.68	-0.12	-0.37
105.55437430	19.30770863	75.98	5.62	17.21	98.81	76.90	5.69	17.42	100.00	-0.91	-0.07	-0.21
105.55435000	19.30767000	76.32	5.30	16.48	98.10	77.80	5.40	16.80	100.00	-1.48	-0.10	-0.32
105.55432500	19.30763000	82.35	5.50	17.37	105.22	78.27	5.22	16.51	100.00	4.08	0.27	0.86

Tab. 1 is the evaluations to the model prediction of the sand percent in the Duoclieu Cooperative - Nghia dan area at a depth of 37.5 cm. Based on the provided metrics, the model for predicting sand content at the 37.5 cm depth in the Nghia Dan area demonstrates excellent performance. The high degree of accuracy is immediately evident from the R2 Score of 0.9135 and a nearly identical Explained Variance Score, which indicate the model successfully explains approximately 91% of the variability in the sand content data. This strong fit is further corroborated by a Pearson coefficient of 0.9575, signifying a near-perfect positive linear correlation between predicted and actual values. In terms of predictive error, the model's precision is notable, with a Mean Absolute Error (MAE) of 2.791, meaning that predictions are, on average, off by only ± 2.8 percentage points. The Root Mean Squared Error (RMSE) of 4.064, which penalizes larger errors more heavily, is still well within an acceptable range for this application. On a relative scale, the Mean Absolute Percentage Error (MAPE) is an impressively low 3.79%, confirming the model's practical utility. Finally, the small standard deviations shown for all metrics suggest that the model is stable and its performance is consistent, making it a reliable tool for this soil mapping task.

Tab. 2 illustrates the data normalization workflow, a critical post-processing step for handling the compositional nature of soil texture predictions. This process is broken down into two stages:

- **Raw Predictions (sand, silt, clay):** These columns represent the initial, uncorrected outputs from the independent machine learning models. Because each component was predicted separately, these values do not adhere to the closure constraint—a fundamental rule for compositional data requiring the sum of all parts to equal 100%. This deviation is explicitly shown in the texture column, which lists sums such as 104.57 and 98.67.
- **Normalized Values (sand (norm), silt (norm), clay (norm)):** These are the corrected, physically consistent outputs. The raw values have been rescaled to enforce the closure constraint, ensuring that their sum for every data point is exactly 100.00.

The key takeaway from this table is the small magnitude of the adjustments required during normalization, as shown in the error columns (the difference between raw and normalized values). The adjustments for sand and clay are generally minor, typically within a few percentage points. The adjustments for silt are consistently very small, often less than one percentage point.

This demonstrates that the initial predictions from the independent ML models were already very close to a physically correct total. The small size of these corrections underscores the high accuracy and reliability of applying machine learning for predicting individual soil texture components.

The spatial distribution of silt (Fig. 4) in the Nghia Dan area reveals a highly complex pattern with dramatic changes between the deeper and shallower soil layers. The overall silt content is generally low, and its distribution is significantly influenced by depth and the presence of bedrock. The prediction model for silt percentage at a depth of 37.5 cm demonstrated strong performance. The key evaluation metrics were highly favorable, with an R-squared (R^2) of 0.85 and an Explained Variance Score (EVS) of 0.85, indicating the model successfully explains 85% of the data's variability. Furthermore, the error metrics were low, with a Mean Absolute Error (MAE) of 0.48 and a Root Mean Squared Error (RMSE) of 0.79.

At 150 cm depth (map a), the soil map is defined by a large bedrock outcrop in the central and northern regions. In the surrounding soil, particularly across the entire southern portion of the area, there is a large, continuous zone of the highest silt content. Values in this zone are approximately 6.2% to over 6.5% by weight (indicated by the orange and red colors). This indicates that at this depth, the fine-grained silt particles are concentrated in a single, dominant southern region. At 75 cm depth (map b), a significant transformation occurs at this depth. The large, high-silt zone seen at 150 cm disappears completely. The bedrock is reduced to smaller, isolated patches. The overall silt content is lower, and the spatial pattern becomes much more fragmented and heterogeneous. The map is predominantly a mosaic of moderate (~5.8% to 6.1% weight percent, yellow tones) and lower (~5.1% to 5.7% weight percent, green tones) silt content, with no clear large-scale zonation. The surface layer at 37.5 cm (map c) largely mirrors the pattern observed at 75 cm, with no surface bedrock. The distribution remains complex and patchy, without a single dominant trend. The majority of the area exhibits moderate to low silt content, primarily ranging from 5.1% to 6.1% by weight (green and yellow colors). Small, isolated pockets of slightly higher silt content (orange tones) can be seen in the southwest and northeast corners.

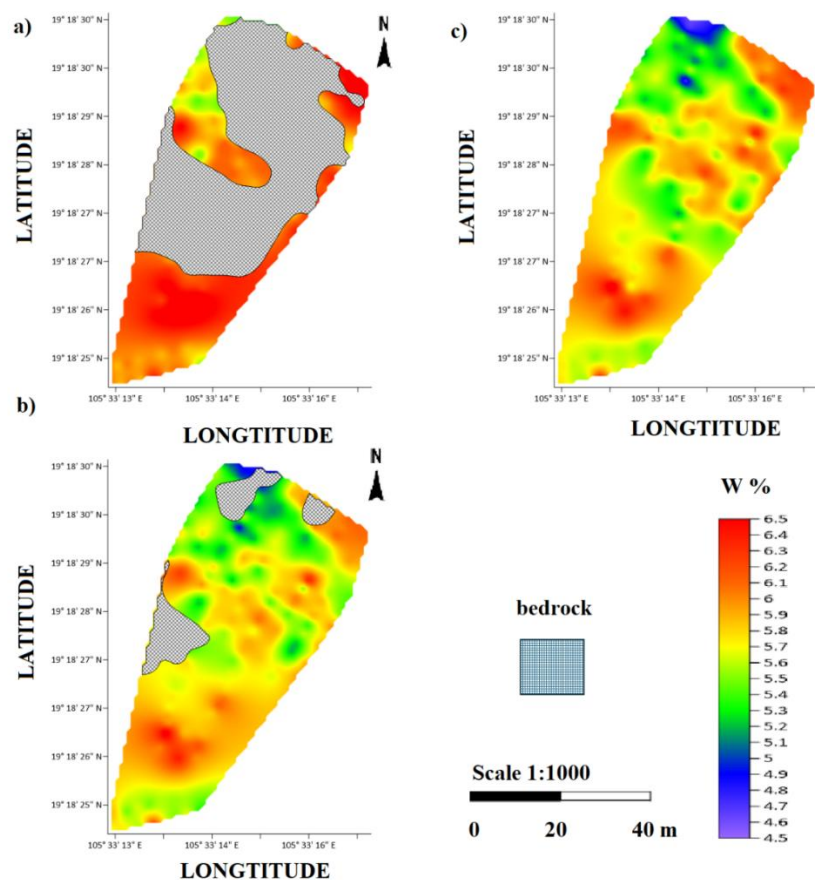


Fig. 4. Predicted silt content at Duoc lieu Cooperative Nghia Dan at a depth a) of 150cm; b) of 75 cm and c) 37.5cm

The most striking feature is the dramatic vertical shift in silt distribution. The profile changes from having a single, large area of high silt concentration at 150 cm to a complex, heterogeneous, and generally lower-silt-content pattern in the upper soil profile (75 cm and 37.5 cm). This sharp change suggests a distinct soil horizon boundary between 150 cm and 75 cm. The loss of the concentrated silt zone in the shallower layers could be indicative of pedological processes such as weathering and translocation, where fine particles are moved and redeposited within the soil profile, leading to the more complex mosaic observed closer to the surface.

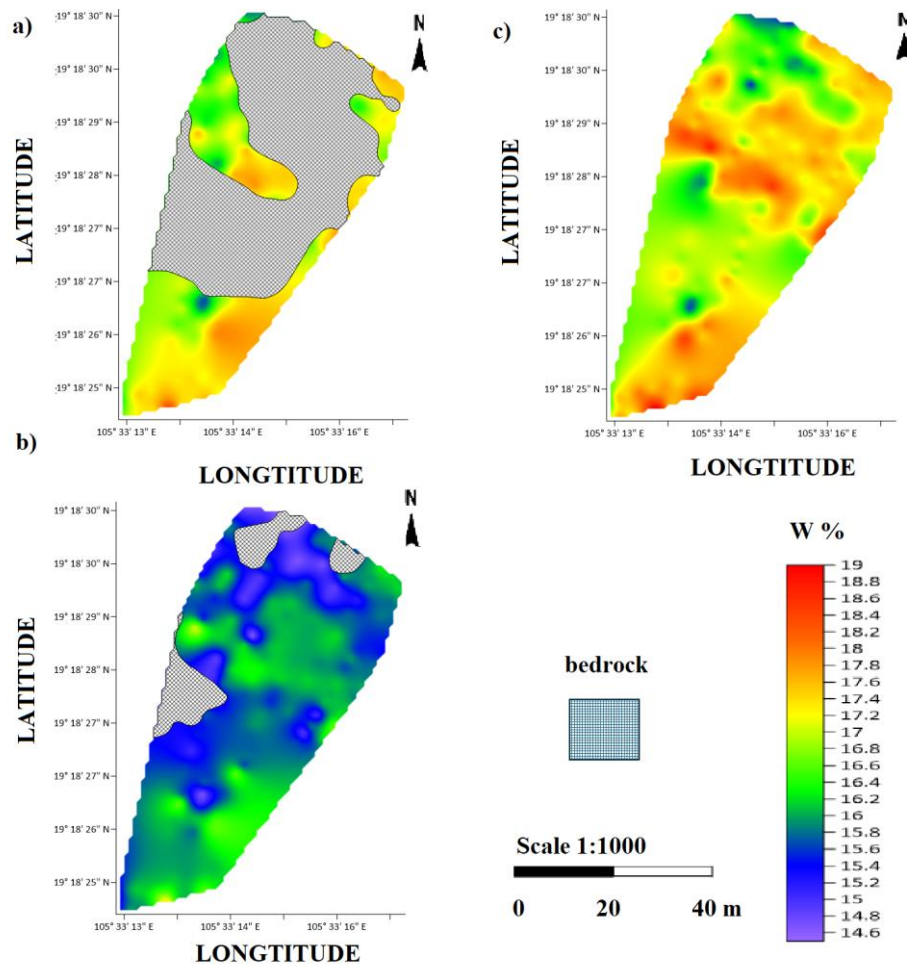


Fig. 5. Predicted clay content at Duoc lieu Cooperative Nghia Dan at a depth a) of 150cm; b) of 75 cm and c) 37.5cm

The distribution of clay in the Nghia Dan area (Fig. 5), is characterized by extreme vertical heterogeneity, with dramatic and distinct shifts in content between each of the three analyzed depths. The prediction model for clay percentage at a depth of 37.5 cm demonstrated robust performance. The key evaluation metrics were favorable, with an R-squared (R^2) of 0.78 and an Explained Variance Score (EVS) of 0.78, indicating the model successfully explains 78% of the data's variability. Furthermore, the prediction errors were low, with a Mean Absolute Error (MAE) of 1.05 and a Root Mean Squared Error (RMSE) of 1.61. This suggests a complex and well-developed soil profile with distinct horizons. At the 150 cm depth (map a), surrounding a large central bedrock feature, the soil exhibits a moderate to relatively high clay content. The values predominantly range from 17.2% to 18.0% by weight (indicated by the yellow and light green colors). The highest concentrations (orange tones, ~18.2% by weight) are located in the southern part of the map, indicating a general trend of increasing clay content towards the south at this depth. A striking and abrupt change occurs at the 75 cm depth (map b). The clay content decreases sharply across almost the entire survey area. The map is overwhelmingly dominated by the lowest values on the scale, ranging from approximately 14.6% to 16.2% (weight percent), as shown by the dark blue and green colors. This indicates the presence of a distinct soil horizon characterized by significant clay depletion compared to the layer beneath it. The surface layer at 37.5 cm (map c) presents another completely different and complex pattern. The clay content increases again relative to the 75 cm layer. The distribution is heterogeneous, featuring a wide range of values from moderate (~16.8% weight percent, green) to high (~18.0% to 18.8% weight percent, yellow and orange). There is a general trend of higher clay content in the western and southern parts of the area, while the northeastern region shows comparatively lower values.

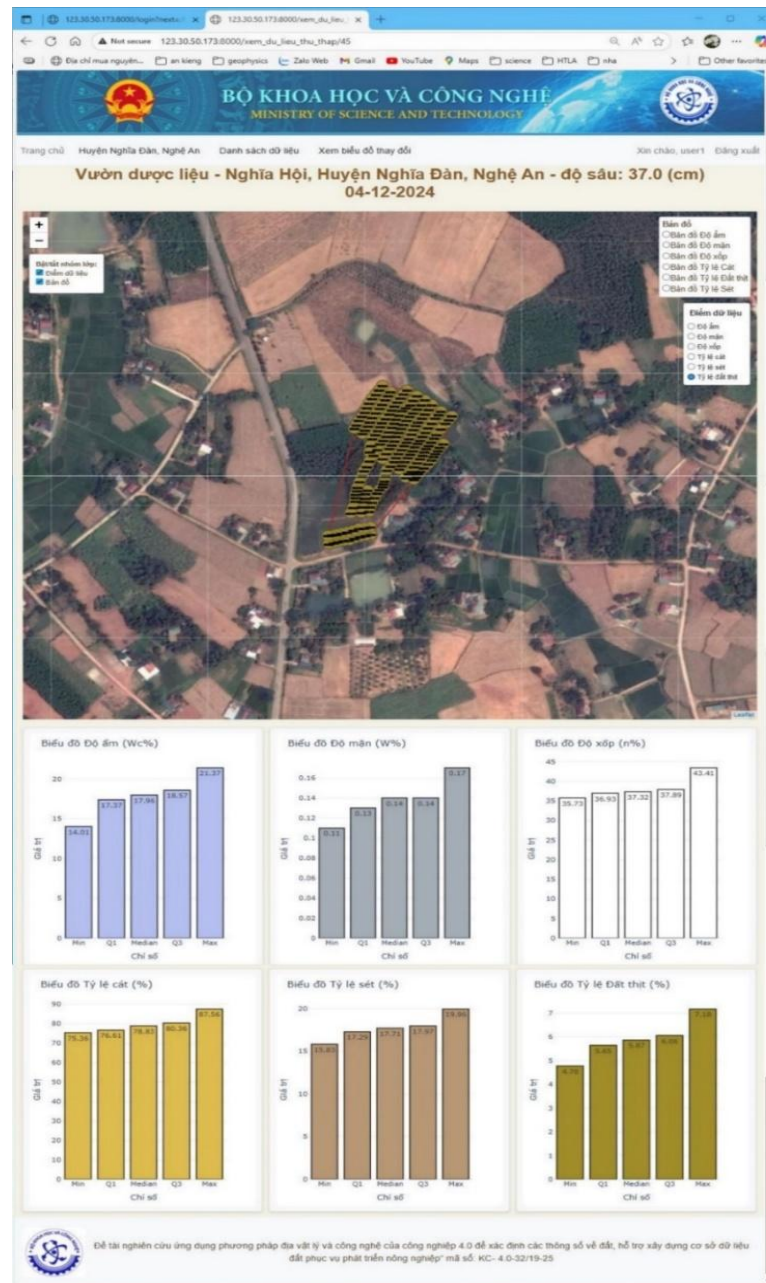


Fig. 6. The WebGIS platform visualizing the high-density discrete data points from the geophysical survey at the 37.5 cm depth.

The clay distribution reveals a classic example of a differentiated soil profile. There is a layer of moderate clay accumulation at 150 cm, followed by a sharply defined, clay-depleted (eluviated) horizon at 75 cm, and then a surface layer with renewed, complex clay distribution. This non-linear, layered structure is a strong indicator of advanced pedological processes, such as clay translocation (lessivage), where clay particles are leached from an upper horizon and accumulate in a lower one. This complex vertical layering is a key characteristic of the soil profile in this region.

In fact, the detailed results for sand, silt, and clay distribution are not only consistent with, but also provide compelling evidence for, the characteristics of a Feralit soil developing in a hilly, mountainous region. The analyses tell a coherent geological story that aligns precisely with this description.

A hilly landscape is characterized by uneven erosion on slopes and deposition in valleys, which creates a complex mosaic of soil properties rather than a uniform cover. The analyzed maps reflect this reality perfectly. They reveal a patchwork of distinct zones, such as the sand-rich western area versus the sand-poor eastern area, and the localization of silt and clay into specific pockets. This spatial heterogeneity is a direct and expected consequence of the varied topography.

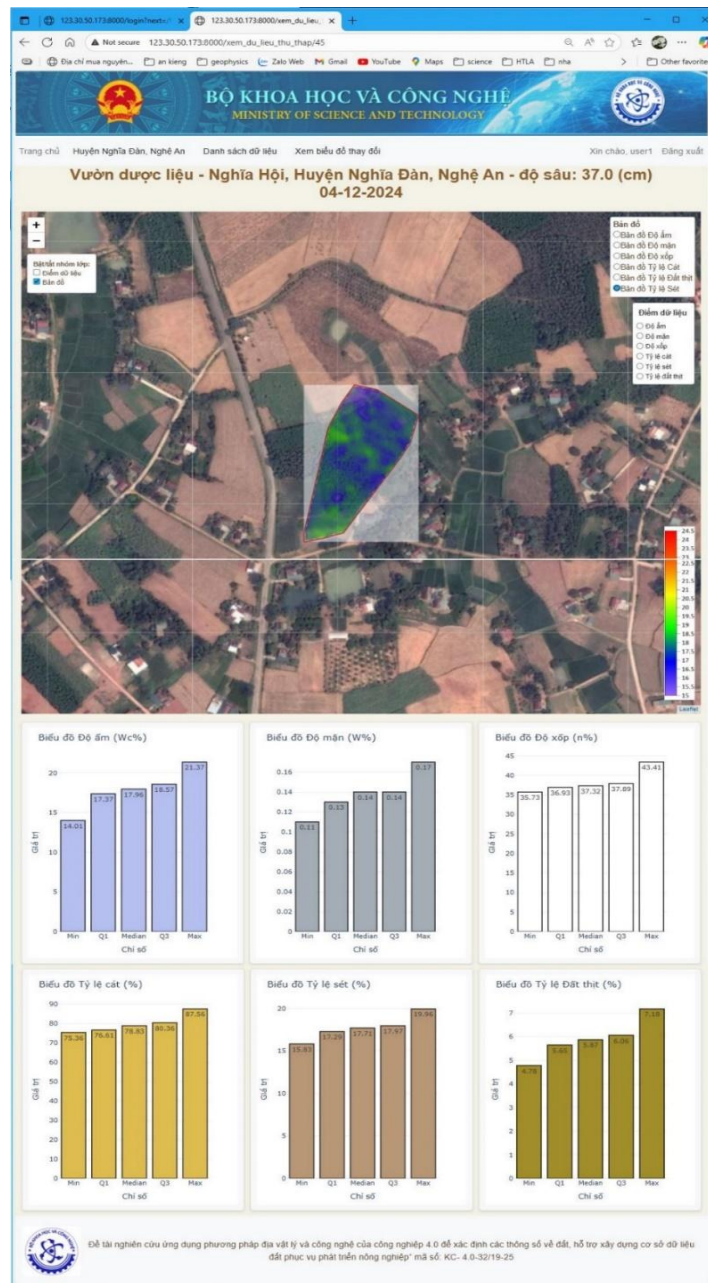


Fig. 7. The WebGIS platform presenting the final predictive map as a continuous surface, illustrating the spatial distribution of a selected soil property at the 37.5 cm depth.

Furthermore, the complex vertical layering observed in the soil profile is a classic diagnostic feature of Feralit soil formation. This process involves intense weathering and the translocation of clay particles, where clay is leached from an upper horizon and accumulates in a lower one. The clay profile analysis provides a textbook example, showing a layer of moderate clay at 150 cm, followed by an abrupt drop to a low-clay layer at 75 cm, and another complex pattern at the surface. Such a dramatic, non-linear vertical structure is a strong indicator of advanced soil development processes.

Finally, the presence of basement rock near the surface, as shown in the deeper maps, is consistent with an erosional landscape where soil cover can be thin. In conclusion, the geophysical data has successfully captured the unique imprint of both the topographical controls and the geological processes of ferralitization that have shaped the soils of the Nghia Dan area.

To translate the study's complex findings into actionable information for end-users, a WebGIS platform was developed to provide farmers and land managers with direct and rapid access to the results (Fig. 6 and Fig. 7). This platform employs WebGIS, a technology that integrates Geographic Information Systems (GIS) with web-based accessibility, to effectively distribute and visualize spatial data. The fundamental role of the platform is to transform complex numerical data into intuitive graphical forms,

such as maps and charts, enabling users to easily observe, understand, and extract valuable insights for decision-making.

The platform's interface is centered around interactive digital maps, which are the most effective method for visualizing this type of spatial data. It offers two primary data visualization modes:

- **Discrete Data Point Visualization:** This mode displays the underlying data framework as an overlay of discrete points on a satellite base map. Each point represents a specific location within the survey grid where soil properties have been predicted. This view is crucial for illustrating the high spatial density of the data that informs the final predictive models.
- **Continuous Predictive Surface Visualization:** This mode presents the final output of the machine learning model as a continuous, interpolated surface map. It utilizes a color ramp to intuitively represent the spatial distribution and magnitude of a selected soil property. This powerful visualization allows users to instantly identify spatial patterns and heterogeneity within the field, which is essential for delineating management zones in precision agriculture.

These two visualization modes are complemented by a series of bar charts that provide quantitative statistical summaries (Minimum, Q1, Median, Q3, Maximum) for each soil property across the entire surveyed area. By combining these interactive spatial displays with quantitative summaries, the WebGIS platform serves as a powerful decision-support tool, bridging the gap between advanced scientific modeling and practical on-farm application.

Finally, the complete methodological workflow for this study is illustrated in Fig. 8. This schematic diagram provides a step-by-step overview of the integrated framework, beginning with the parallel acquisition of high-density geophysical data (physical proxies) and sparse physical soil samples (ground truth). It then details the key stages of data processing, Random Forest model training, and the application of the model to generate predictive maps. The workflow concludes with the crucial post-processing normalization step to ensure physical consistency, and the final delivery of the high-resolution digital soil maps via the WebGIS platform.

4. Conclusion

This study successfully demonstrated that an integrated framework of geophysical methods and machine learning can produce high-fidelity digital soil maps in a complex Feralit landscape in Nghia Dan, Vietnam. The workflow effectively overcame the limitations of traditional sparse sampling to generate predictive maps of soil texture at three distinct depths. Specifically, the Random Forest model predicted sand content at the 37.5 cm depth with excellent performance, achieving an R-squared (R^2) of 0.9135 and a Root Mean Squared Error (RMSE) of 4.064. The models for silt and clay at this depth also showed strong predictive accuracy ($R^2 = 0.85$ and $R^2 = 0.78$, respectively). The resulting maps revealed complex spatial heterogeneity and vertical layering that are highly consistent with the known pedological processes of Feralit soil formation in an erosional topography.

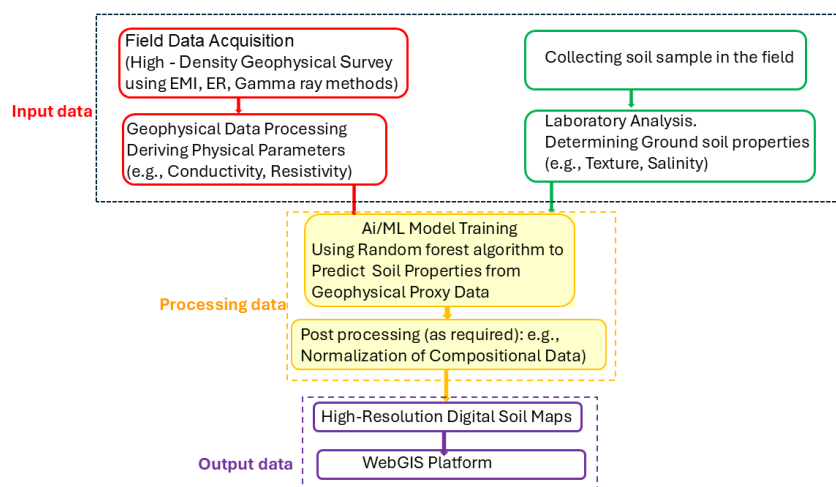


Fig. 8. schematic workflow illustrating the integrated geophysical and machine learning approach for high-resolution digital soil mapping

The success of this case study provides compelling evidence for a paradigm shift in Digital Soil Mapping (DSM). By leveraging dense geophysical measurements, this approach transforms DSM from an interpolation-reliant practice to a data-rich, model-driven science. It overcomes the severe spatial limitations of point-based sampling, converting a smoothed map based on an "educated guess" into a high-resolution "photograph" of soil variability. As a rapid and non-destructive methodology, it is also a more economically viable solution for large-scale mapping. The vast datasets generated create a synergistic relationship with AI algorithms, enabling the simultaneous prediction of multiple soil properties and representing a transformative leap in analytical efficiency.

While this study successfully demonstrated the workflow's effectiveness across three diverse topographies in Nghe An province (mountainous, plain, and coastal), a primary limitation is the scope of the predicted soil attributes, which was focused on fundamental texture (sand, silt, and clay). Future research should aim to expand this suite of predictions. To achieve this, the integration of complementary geophysical methods, such as Ground Penetrating Radar (GPR), should be considered. Incorporating GPR data could provide additional valuable covariates related to subsurface structure and moisture, potentially enabling the prediction of a wider range of soil properties and further improving model accuracy.

Acknowledgement

The authors acknowledge the support of the Ministry of Science and Technology (Vietnam) and the National Science and Technology Program 4.0/19-25 for research funding. We also thank Hanoi University of Mining and Geology for their support.

References

- Yen N.T.H, 2018. Phat trien nong nghiep Nghe An theo huong hien dai trong dieu kien hoi nhap kinh te quoc te. (Doctoral's thesis – Vietnam Academy of Social science)
- Huong Phan, Phong Nguyen, Kha Tran, 2022, A review of geophysical application in Agriculture and their potential in Vietnam, Proceedings of the 13th International Conference on Application of Information Technology in Agriculture Asia-Pacific Region- Application of Smart technologies for achieving sustainable agriculture, pp 135-140.
- Huong Phan, Huu Tran, Duy Nguyen. 2024a. Apply electromagnetic induction method in precise agriculture. In: 16th International Conference of the East and Southeast Asia Federation of Soil Science Societies, Vietnam; 2024. PS1E-3. DOI: 10.13140/RG.2.2.28154.20167.
- Huong T. Phan, Duong H. Vu, Hung D. Tran, Tung T. Nguyen. 2024b. Application of geophysical methods in agriculture and their potential in Vietnam. Journal of Mining and Earth Science. 65(2):86-98.
- Huong Phan, Tung Nguyen, Phong Nguyen. 2024c. Machine learning analysis Geophysical data to study soil's properties. In: Advances in the earth, mining and environmental sciences for safe and sustainable development. Hanoi: NXB Khoa học và Kỹ thuật; 2024. p. 223-229. DOI: 10.15625/vap.2024.0342
- Searle, R., McBratney, A., Grundy, M., Kidd, D., Malone, B., Arrouays, D., Stockman, U., Zund, P., Wilson, P., Wilford, J. and Van Gool, D., 2021. Digital soil mapping and assessment for Australia and beyond: A propitious future. Geoderma Regional, 24, p.e00359.
- Koganti, T., De Smedt, P., Farzamian, M., Knadel, M., Triantafilis, J., Christiansen, A.V. and Greve, M.H., 2024. Digital soil mapping using electromagnetic sensors. Frontiers in Soil Science, 4, p.1536797.
- Benedetto, F. and Tosti, F., 2013. GPR spectral analysis for clay content evaluation by the frequency shift method. Journal of Applied Geophysics, 97, pp.89-96.
- Triantafilis, J., Odeh, I.O.A. and McBratney, A.B., 2001. Five geostatistical models to predict soil salinity from electromagnetic induction data across irrigated cotton. Soil Science Society of America Journal, 65(3), pp.869-878.