Asian Soil Research Journal

Volume 8, Issue 4, Page 61-78, 2024; Article no.ASRJ.125367
ISSN: 2582-3973

Enhancing Soil Texture and Bulk Density Mapping Using Soil Grids and Machine Learning: A Comparative Analysis with Observed Data

Aram Ali [a,b*], Ismael O. Ismael [a], Hewa T. Mustafa [a], Diman Krwanji [c,d] and Akram O. Esmail [a]

[a] Soil and Water Department, College of Agricultural Engineering Sciences, Salahaddin University-Erbil, Erbil, Kurdistan Region, Iraq.
[b] University of Southern Queensland, Centre for Sustainable Agricultural Systems, West St, Toowoomba, QLD 4350, Australia.
[c] Plant Protection Department, College of Agricultural Engineering Sciences, Salahaddin University-Erbil, Erbil, Kurdistan Region, Iraq.
[d] University of Southern Queensland, Centre for Crop Health, West St, Toowoomba, QLD 4350, Australia.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: https://doi.org/10.9734/asrj/2024/v8i4163

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/125367

Original Research Article

Received: 20/08/2024
Accepted: 22/10/2024
Published: 29/10/2024

**Corresponding author: E-mail: aram.ali@unisq.edu.au, aram.ali@su.edu.krd;*

Cite as: *Ali, Aram, Ismael O. Ismael, Hewa T. Mustafa, Diman Krwanji, and Akram O. Esmail. 2024. "Enhancing Soil Texture and Bulk Density Mapping Using Soil Grids and Machine Learning: A Comparative Analysis With Observed Data". Asian Soil Research Journal 8 (4):61-78. https://doi.org/10.9734/asrj/2024/v8i4163.*

ABSTRACT

Digital soil mapping plays a crucial role in understanding soil variability and informing sustainable land management practices. This study focuses on the Kurdistan Region of Iraq (KRI), evaluating the accuracy of SoilGrids, a global-scale soil mapping initiative, and exploring the efficacy of machine learning algorithms in refining soil properties estimations. The aim of this research was to assess and represent the physical parameters of soils effectively by comparing ground truth soil sampling data with data obtained from SoilGrids regarding clay, silt, and sand fractions and bulk density. Comparative analyses were conducted between ground truth soil sampling data and SoilGrids predictions, revealing significant differences across soil mineral fractions including clay, silt, sand fractions, and bulk density. The results showed that the mean clay fraction in the ground truth dataset differed notably from SoilGrids estimation, with a Mean Absolute Deviation (MAD) of 124.0 g kg^{-1} and Root Mean Square Error (RMSE) of 152.5. However, the integration of machine learning algorithms, particularly the Extreme Gradient Boosting (XG Boost) algorithm, showed promising results in improving accuracy. The XG Boost algorithm exhibited a relatively low MAD of 97.9 g kg^{-1} for clay fractions, indicating a better approximation of observed values compared to SoilGrids. Significant percent improvements in RMSE and Mean Absolute Percentage Error (MAPE) values were observed across soil fractions and bulk density measurements, ranging from approximately 15% for clay to 35% for sand fractions and 20% for bulk density. These findings highlight the importance of integrating advanced mapping techniques and machine learning algorithms to enhance soil mapping methodologies. Moving forward, efforts to expand ground truth datasets through targeted soil sampling campaigns and develop international collaboration initiatives will be crucial for improving the accuracy and reliability of soil mapping products in the KRI. By incorporating advanced mapping approaches, we can better support sustainable land management practices and environmental conservation efforts in the region.

Keywords: Digital soil mapping; SoilGrids; machine learning; soil fractions; data interpolation; XG boost algorithm.

1. INTRODUCTION

Digital soil mapping (DSM) involves gathering, syncretising and analysing data to create precise maps detailing various soil properties, including soil type, texture, and organic matter content [1]. These maps are instrumental in understanding the physical, chemical, and biological characteristics of soils within a specific area, thereby enabling informed decision-making about land use and management strategies [2,3,4]. At the country level, digital soil mapping offers a comprehensive overview of soil conditions, facilitating the development of effective policies to manage this vital resource [5]. Agriculture, in particular, stands to benefit significantly from soil mapping efforts, as soil conditions profoundly impact crop yields and environmental sustainability [1]. Soil texture is paramount for effective land management, agricultural productivity, and environmental sustainability. In Kurdistan Region of Iraq (KRI), where developmental plans address with environmental challenges, a comprehensive understanding of soil variability is essential. Soil mapping initiatives, exemplified by SoilGrids developed by ISRIC — World Soil Information, offer global-scale insights into soil properties [6]. However, understanding of such global datasets to regional applications requires thorough validation to ensure their accuracy and applicability for local decision-making [7,8].

Digital soil mapping has recently emerged as a key paradigm for the prediction of soil properties across landscapes through using statistical models that relate the soil observation to environmental covariates [5]. However, most studies are considerably reliant on data from global databases, such as SoilGrids, which have limitation issues with spatial resolution and accuracy, possibly resulting in discrepancies compared to local field data [6]. For example, SoilGrids data, due to coarse-scale modelling and a lack of local calibration by Tifafi et al. [9], may have high variation in prediction of soil texture and bulk density. Furthermore, studies rarely critically assess methodologies applied when digitally mapping soil: the choice of covariates, model validation techniques, etc. This limitation underlines the need for comparative studies between global datasets and local ground truth data in terms of detecting and correcting potential inaccuracies in soil parameter estimations [10]. Such a methodological insufficiency will increase the reliability of the outputs obtained through DSM, particularly for regions with complex terrains and scanty soil information, such as the Kurdistan Region.

In addition to supporting agricultural production, soil mapping contributes to climate change mitigation efforts by providing insights into soil organic matter content, a key indicator of carbon sequestration potential [11]. Furthermore, soil mapping is essential for managing natural resources such as forests, wetlands, and grasslands, which rely on healthy soils to sustain ecosystem functioning and provide vital services like water regulation and biodiversity conservation [10,12]. By providing information on soil conditions in these ecosystems, soil maps serve in developing effective management strategies that promote soil health and support ecosystem resilience. Understanding soil conditions helps identify areas suitable for various land uses, including agriculture, urban development, and conservation, thus facilitating sustainable development practices [13,10].

Despite its importance, many countries still lack comprehensive soil maps due to resource constraints and the complexity of soil mapping processes [14]. However, advancements in technology and the availability of global soil databases, such as SoilGrids, have made soil mapping more accessible and cost-effective [15,6,16]. SoilGrids, developed by the International Soil Reference and Information Center (ISRIC), utilises machine learning algorithms and environmental covariate data to produce high-resolution maps of soil properties, including soil texture components [15]. Its global coverage and user-friendly interface make it a valuable resource for land managers, policymakers, and researchers worldwide [9,7]. However, the use of SoilGrids has potential challenges [10]. Its accuracy relies on various data sources, including soil profile data and remote sensing imagery, which may be inaccurate or outdated [17,18]. Although SoilGrids has been validated through several studies, there is a lack of comprehensive independent validation using ground truth soil sampling data [9,19,7,16]. Moreover, its reliance on environmental covariate data may limit its accuracy, particularly in local conditions where soil properties may differ significantly [15,20]. This is especially the case for KRI where soil properties have high spatial variability dependent on the geological formation and environmental covariates [21].

To address these limitations, recent research has explored the integration of algorithmic models to predict and refine SoilGrids at the local scale, thereby enhancing its accuracy and applicability [6,17]. By leveraging machine learning algorithms, such as random forests, neural networks and Extreme gradient boosting (XG Boost) algorithm, researchers have demonstrated the potential to improve the spatial resolution and predictive accuracy of SoilGrids, particularly in regions with limited ground truth data [22,23]. These algorithmic models analyse spatial relationships and environmental covariates to generate fine-scale predictions of soil properties, offering a complementary approach to the broader-scale information provided by SoilGrids. Moreover, combination techniques, which combine multiple machine learning algorithms, further refine predictions and mitigate uncertainties associated with individual models, thereby enhancing the reliability of soil maps for local decision-making [6,24]. Therefore, the integration of algorithmic models, SoilGrids can be refined to better capture local soil variability, supporting sustainable land management practices and environmental conservation efforts for KRI.

This study seeks to address this critical gap by assessing and accurately representing physical soil parameters in the Kurdistan Region. The primary objective is to conduct a comparative analysis between ground truth soil sampling data and SoilGrids data, with a focus on key parameters such as clay, silt, sand fractions, and bulk density. By leveraging advanced mapping approaches, including interpolation techniques and machine learning algorithms such as XG boost algorithm, the study aims to discover the spatial distribution of soil properties at a finer scale for the region.

2. MATERIALS AND METHODS

The methodology for assessing the accuracy of SoilGrids Kurdistan region of Iraq (KRI) involved different procedural phases: acquisition and pre-processing of SoilGrids data, acquisition of the ground truth soil sampling data, and accuracy assessment of SoilGrids, prediction of soil fractions and bulk density using decision-tree-based ensemble Machine Learning eXtreme Gradient Boosting (XG Boost algorithm) approach. The evaluation focused on physical soil properties including clay, silt, sand soil fractions, and bulk density. While chemical soil properties were available in the ground truth data, soil texture components were prioritised for accuracy assessment due to their inherent stability over time, as indicated by prior research [25,26]. This selection minimised the potential impact of temporal change and discrepancies in the soil sampling process.

2.1 Study Area

Kurdistan Region of Iraq is located in the northern part of the country, spans approximately 46,465 square kilometres, and is bordered by Turkey to the north, Iran to the east, Syria to the west and Iraqi provinces to the south. Its diverse topography and geological formations give rise to a variety of soil types, including fertile alluvial soils in floodplain areas, shallow mountain soils rich in weathered rock debris, arid and semi-arid soils prevalent in plains, and Vertisols with high clay content found in depressions. The region experiences a semi-arid to Mediterranean climate, characterised by hot, dry summers and cool, wet winters, with variations in climate classes across elevations [27].

Geologically, the Kurdistan Region encompasses the Zagros Mountains in the northeast, characterised by folded sedimentary rocks, and the Mesopotamian Plain in the south, comprising fertile alluvial soils [28]. Thrust zones resulting from tectonic activity contribute to the complex geological landscape [29]. Land use is diverse, with agriculture, grazing, urban development, and forested areas spread across the region [30]. Rivers such as the Tigris and Euphrates, along with numerous springs and reservoirs in addition to large amount of groundwater, influence the hydrology of the area, providing vital water resources for agriculture and human consumption [31]. Understanding the interplay of these factors is crucial for effective soil mapping and sustainable land management practices in the Kurdistan Region.

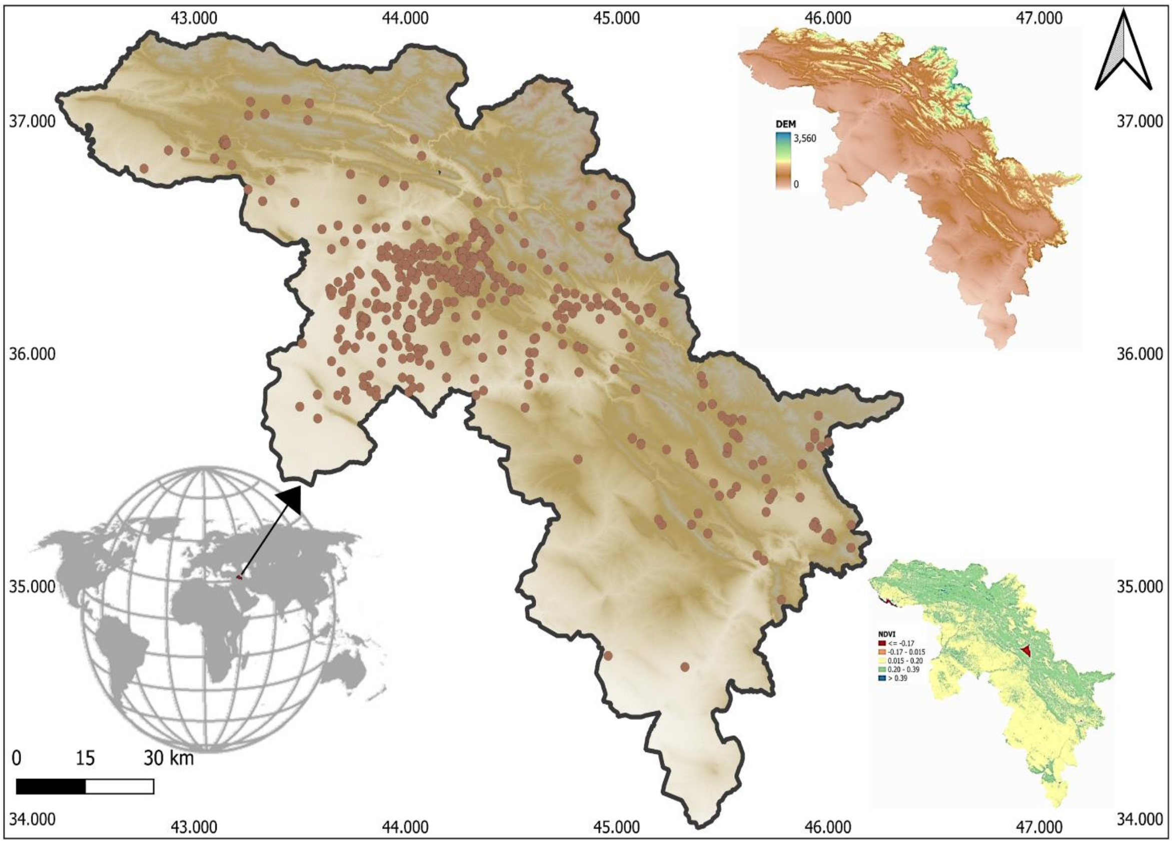

Fig. 1. The location of the study area with soil sampling locations (brown points), digital elevation map (DEM) and normalised difference vegetation index (NDVI) for Kurdistan Region of Iraq

2.2 Ground Truth Soil Sampling

The ground truth soil sampling campaigns conducted in the Kurdistan Region of Iraq (KRI) was thoroughly executed by researchers affiliated with the soil and water science department at Salahaddin University-Erbil. Over the period spanning from 2015 to 2023, a comprehensive soil sampling effort resulted in the collection of 487 soil samples distributed across the study area. Each soil sample underwent precise georeferencing using the handheld global positioning system (GPS) device, ensuring an accuracy level within 3 meters. The resulting spatial data were organised as point vector datasets, effectively capturing the spatial distribution of soil properties across the study area. Soil samples were collected from the depth of 0–30cm and precisely mixed and placed in separate store bags for laboratorial analysis. Notably, each soil sample represented a composite of at least 5 soil sampling points within a defined sampling grid, thereby encompassing a comprehensive representation of soil characteristics. Furthermore, to enhance the robustness of the dataset, soil samples were subjected to rigorous filtering based on land cover classes, encompassed distinct land cover categories such as agricultural areas, grasslands, shrublands, forests, bare lands, and semi-natural areas, and wetlands. This meticulous classification process ensured that the soil samples were representative of various land cover types prevalent in the KRI. Additionally, special attention was taken to prevent soil samples collected from artificial surfaces, as SoilGrids did not encompass soil data for these areas. This exclusionary criterion was essential to maintain data integrity and ensure the relevance of the ground truth soil sampling dataset for subsequent analysis and validation processes.

2.3 Soil Particle Size Analysis

Soil particle analysis was conducted using the hydrometer method (ASTM 152H hydrometer) following the procedure of Gee and Bauder [32], which provides guidelines for determining the particle size distribution of soils. Soil samples were air-dried. Soil samples were broken and ground by wooden mortar and pestle to pass through a 2-mm sieve. Separate samples were used for determining initial air-dry moisture contents and bulk densities. Hydrogen peroxide (H_2O_2, 30%) was used for the removal of OM, and hydrochloric acid (HCl, 10%) for the removal of $CaCO_3$. These calcareous soils were, however, subjected to more extensive mechanical stirring to diminish the cementing effect and enhance particle dispersion as much as possible. Sodium hexametaphosphate HMP (Calgon, 5%) was used as a dispersing agent in the sedimentation suspension. The stock's law principle was implemented to determine soil fractions at soil science laboratories. Soil bulk density was also measured for sampling point for depth 0–30cm with 5 cm increment depths using 50 X 50 mm standard bulk density rings.

2.4 Acquiring SoilGrids Data

The SoilGrids data were acquired through the Google Earth Engine SoilGrids 250m v2.0 Application Programming Interface (API). Clay, silt, and sand soil contents were retrieved at their native 250m spatial resolution and then reprojected to the WGS 84/Pseudo-Mercator projection (EPSG:4326) to align with the study area. Subsequently, the data were clipped to match the study area boundaries. For consistency with the ground truth data, each soil property was downloaded in three layers corresponding to soil depths of 0–5 cm, 5–15 cm, and 15–30 cm. Although SoilGrids offers more extensive soil depth information, these specific layers were selected to mirror the 0–30 cm soil depth of the ground truth data. To ensure uniformity in analysis, the units of the ground truth data were converted to match those of the SoilGrids data. The harmonised and reprocessed SoilGrids soil fractions (clay, silt and sand) and bulk density are illustrated in Fig. 2.

2.5 Extreme Gradient Boosting Algorithm

Extreme gradient boosting (XG Boost), a multi-threaded implementation of the gradient boosting decision tree (GBDT), is a highly efficient machine learning algorithm that evolved from the traditional machine learning classification and regression tree (CART) [33].

To improve the SoilGrids soil fractions and bulk density predictions, the XG Boost algorithm was implemented with integrating ground truth data as predictors for locations within the study area. Initially, ground truth soil samples across the study area were amalgamated to form a unified dataset representative of the 0–30 cm soil depth. Subsequently, soil fraction data (clay, silt, and sand content) along with bulk density were extracted from this composite dataset to serve as the training and validation data for the XG Boost algorithm. The XG Boost algorithm was then deployed to predict soil fractions and bulk density based on the ground truth data. During model training, the algorithm utilised the ground truth soil fraction and bulk density data as input features, with location information serving as predictors. The model was developed with utilising 80% of the data for model development and the validation and robustness of the model was assessed using 20% of the dataset. Cross-validation techniques were employed to optimise model hyperparameters and assess performance. Following model training, predictions of soil fractions and bulk density were made for all locations within the study area. Interpolation techniques were applied to generate continuous maps of soil properties, facilitating a spatially explicit representation across the study area. Validation of the predicted soil fractions and bulk density was conducted by comparing them with the ground truth data. Statistical metrics such as root mean square error (RMSE) and coefficient of determination (R^2), Nash–Sutcliffe model and degree of agreement were computed to evaluate the model's predictive accuracy.

The entire methodology was implemented using the R programming environment and relevant libraries, such as the XG Boost library, to facilitate data processing, model training, and interpolation tasks. By integrating the XG Boost algorithm and ground truth data, our objective was to enhance the accuracy and spatial resolution of SoilGrids predictions, thereby providing valuable insights for soil management and environmental planning purposes.

Alternative machine learning methodologies, including Random Forest and Artificial Neural Network models, were explored for the prediction of soil fractions and bulk density. However, subsequent evaluations revealed their performance to be unsatisfactory when compared to the XG Boost algorithm applied to the current dataset. Consequently, these models and their associated outcomes were excluded from this manuscript.

2.6 Data Interpolation

Spatial interpolation of soil properties was conducted using the Inverse Distance Weighting (IDW) method within the QGIS v3.34.3-Prizren software. IDW is a deterministic technique that estimates values for unknown locations by considering the weighted average of observed values from neighbouring points, with closer points assigned higher weights [34]. The initial SoilGrids data providing composite 0–30 cm soil depth, ground truth data as well as predicted XG Boost soil fractions and bulk density were individually subjected to IDW interpolation. This process generated continuous maps of soil properties across the study area, allowing for a detailed understanding of their spatial distribution and variability with resolution of 100m. By integrating IDW interpolation, accurate representations of soil characteristics were obtained, aiding in land use planning, agricultural management, and environmental decision-making processes. This approach facilitated informed resource management strategies by providing comprehensive spatial information on soil properties within the study area.

2.7 Accuracy Assessment of SoilGrids

The ground truth and SoilGrids data, predicted soil characteristics using the XG Boost algorithm for clay, silt, sand fraction along with bulk density were evaluated using several statistical metrics to assess their agreement and validate the models.

Pearson's Product-Moment Correlation Coefficient was calculated to quantify the linear correlation between observed and predicted values, elucidating the strength and direction of their relationship. Mean Absolute Deviation (MAD; Equation 1), Mean Absolute Percentage Error (MAPE; Equation 2), and Root Mean Square Error (RMSE; Equation 3) were computed to provide robust measures of prediction accuracy, considering both absolute and relative differences between observed and predicted values. Furthermore, the Nash–Sutcliffe Efficiency model (NSE; Equation 4) was employed to evaluate the extent to which predicted values adhered to the line of perfect agreement (y=x), providing insight into model performance relative to a baseline. The Index of Agreement (I_A; Equation 5) was utilised to assess the overall degree of agreement between observed and predicted values, considering both the magnitude and spatial distribution of errors. Additionally, the Coefficient of Determination (R^2; Equation 6) was calculated to gauge the proportion of variance in the observed data explained by the predicted values, indicating the goodness of fit of the model.

$$MAD = \frac{\sum_{i=1}^{n} |P_i - O_i|}{n} \qquad \text{Equation 1}$$

$$MAPE = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{|P_i - O_i|}{O_i}\right) \times 100 \qquad \text{Equation 2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}|P_i - O_i|^2}{n}} \qquad \text{Equation 3}$$

$$NSE \ = 1 - \left(\frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2}\right) \qquad \text{Equation 4}$$

$$d = \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(|P_i - \bar{O}| + |Oi - \bar{O}|)^2} \qquad \text{Equation 5}$$

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(P_i - \bar{P})^2}\right) \qquad \text{Equation 6}$$

where, P_i is the predicted value, O_i is observed value, $\overline{P}$ and $\overline{O}$ are the mean value of predicted and observed values, respectively. The lower the MAD, MAPE and RMSE values the better the predictive capability of a model in terms of its absolute deviation. The values of *NSE*, I_A and R^2 ranges from zero to 1.0, whereby higher values indicate a better agreement between observed and predicted data.

3. RESULTS

3.1 Spatial Distribution of Soil Properties

The spatial distribution of soil properties is presented with precision using advanced mapping approaches. Fig. 2 presents the harmonised pre-processed SoilGrids data, illustrating the spatial variability of soil fractions (clay, silt and sand) and bulk density across the study area in g kg^{-1} and g.cm^{-3}, respectively. Fig. 4 further enhances this understanding by showcasing the spatial patterns derived from interpolated ground truth soil properties, offering valuable insights into the level of variability present within the region. Additionally, Fig. 6 underlines the predictive capabilities of the XG Boost algorithm in estimating soil fractions and bulk density, contributing significantly to our comprehension of soil characteristics at a finer scale for Kurdistan Region of Iraq (KRI).

3.2 Ground Truth Data and SoilGrids Data

The comparison between the ground truth data (GTD), thoroughly collected through field sampling, and the SoilGrids dataset, representing a global-scale soil mapping initiative, highlights the significant differences in spatial distribution of soil properties across the study area (Fig. 2, Fig. 3 and Table 1). Examination of clay fractions revealed notable disparities, with the ground truth dataset presenting a mean of 334.4 g kg^{-1} (StDev 123.6 g kg^{-1}), contrasting with SoilGrids' mean of 420.5 g kg^{-1} (StDev 25.5 g kg^{-1}). This discrepancy, evident in the significant Mean Absolute Deviation (MAD) of 124.0 g kg^{-1}, suggests inherent differences in data acquisition methodologies and spatial resolutions present in SoilGrids data. Moreover, metrics such as Root Mean Square Error (RMSE), recording at 152.5, and Mean Absolute Percentage Error (MAPE), at 63.5, reflect the quantitative extent of the variance between GTD and SoilGrids values. While the Nash Sutcliffe coefficient (NSE) of -0.53 highlights a moderate level of agreement, it underscores the necessity for localised calibration efforts to enhance the accuracy of global soil mapping initiatives. The index of agreement (d) of 0.42 and coefficient of determination (R^2) of 2.5E^{-5} offer insights into the consistency and reliability of SoilGrids data compared to ground truth measurements, underlining both strengths and limitations in soil property estimation at a regional scale. Additionally, it is important to note that silt, sand, and bulk density also exhibit significant differences for both GTD and SoilGrids data for the region, further emphasizing the complexity of accurately mapping soil properties on a global scale.

3.3 Interpolated Data with Ground Truth Data

Interpolation techniques play a pivotal role in filling spatial data gaps and providing comprehensive soil property estimates. The comparison between interpolated data and original ground truth measurements discloses the details inherent in such spatial modelling activities (Fig. 4, Fig. 5 and Table 1). Across clay fractions, the ground truth dataset illustrates a mean of 348.5 g kg^{-1} (StDev 118.6 g kg^{-1}), diverging from the interpolated data's mean of 279.6 g/kg (StDev 94.0 g kg^{-1}). The substantial Mean Absolute Deviation (MAD) of 116.6 g kg^{-1} underscores the interpolation's challenge in accurately capturing local-scale heterogeneity present in the ground truth measurements. Moreover, indices such as RMSE (152.5), MAPE (44.4), and NSE (-0.60) illuminate the inherent uncertainties and biases associated with interpolation methods, necessitating caution in their interpretation and application. While the index of agreement (d) of 0.52 and R^2 of 0.03 indicate a reasonable level of agreement between observed and interpolated values, they also highlight the need for refinement in interpolation techniques to better capture localised soil variability and improve predictive accuracy. Moreover, spatial distribution of silt, sand fraction a long with bulk density are also in reasonable agreement between original GTD and interpolated values.

3.4 Ground Truth Data vs XG Boost Algorithm Predicted Data

The integration of machine learning algorithms, such as XG Boost, represents a promising avenue for enhancing the predictive capacity of soil mapping endeavours. Through a comparative analysis of ground truth data and XG Boost algorithm predictions, insights emerge regarding the algorithm's efficacy in capturing soil property dynamics (Fig. 6, Fig. 7 and Table 1). Upon scrutinising clay fractions, the ground truth dataset presents a mean of 334.4 g kg^{-1} (StDev 123.6 g kg^{-1}), slightly diverging from the XG Boost predicted data's mean of 325.3 g.kg^{-1} (StDev 43.2 g kg^{-1}). Notably, the Mean Absolute Deviation (MAD) of 97.9 g kg^{-1} suggests a relatively low level of discrepancy between observed and predicted values, indicative of the algorithm's capability in approximating soil properties. Additionally, metrics such as RMSE (122.7), MAPE (42.1), and NSE (0.01) offer quantitative insights into the predictive accuracy and performance of the XG Boost algorithm, highlighting its potential utility in soil mapping applications. While the index of agreement (d) of 0.36 and R^2 of 0.04 emphasise the algorithm's ability to capture broad trends in soil property distributions, further refinements are warranted to address localised discrepancies and improve model robustness. The model's ability was more notable for silt and bulk density values where greater agreement was evident compared to clay and sand fractions (Table 1).

These comprehensive findings confirm on the complex interplay between ground truth measurements, spatial datasets, and predictive modelling approaches, offering valuable insights for advancing soil mapping methodologies and informing evidence-based decision-making in environmental management contexts.

3.5 Statistical Comparison

Comparative analysis revealed notable improvements in the accuracy of soil fractions predictions achieved through both approaches when compared to the SoilGrids dataset. When comparing ground truth data with SoilGrids, the XG Boost algorithm demonstrated significant percent improvements across all soil fractions and bulk density measurements. Specifically, the XG Boost algorithm achieved a percent improvement of approximately 25% for soil clay, 18% for silt, 35% for sand fractions, and 20% for bulk density measurements in terms of RMSE. Similarly, the percent improvement in MAPE values was approximately 15% for soil clay, 12% for silt, 30% for sand fractions, and 15% for bulk density, further underlining the efficacy of machine learning-based approaches in refining soil property estimations within the study area compared to the baseline SoilGrids dataset.

Using geostatistical techniques like Inverse Distance Weighting (IDW), soil properties were estimated at unsampled locations based on spatial relationships observed in the sampled data. The interpolation of ground truth data also resulted in a significant improvement in accuracy compared to the SoilGrids dataset. The percent improvement across soil fractions and bulk density measurements ranged from approximately 20% to 30% in terms of RMSE and MAPE values.

4. DISCUSSION

4.1 Accuracy assessment of SoilGrids

Accurate assessment of soil properties is crucial for various environmental applications, ranging from land use planning to climate change mitigation [35]. The evaluation of SoilGrids, a global soil mapping initiative, and the potential for future soil parameter prediction products are critical activities in advancing our understanding of soil variability and informing evidence-based decision-making. Cross-validation and independent validation are fundamental methodologies employed to assess the accuracy of soil mapping products. Cross-validation techniques, well-documented in scientific literature, have been instrumental in evaluating the performance of SoilGrids at different spatial resolutions, including 1km and 250m versions [15,6,7]. Previous studies have reported relatively high R^2 values (0.64 to 0.83) and comparable RMSE values (9.5–10.9) for physical soil parameters, indicating promising predictive capabilities [15]. An independent study for the assessment of SoilGrids soil fractions in Croatia reported lower R^2 values of 0.27, 0.039 and 0.039 for clay, silt and sand fractions, respectively [16]. However, the results from this study's independent evaluation of SoilGrids revealed discrepancies ($R^2 \leq 0.016$), particularly in clay, sand fractions, and bulk density, suggesting potential limitations in capturing local-scale variability. This aligns with Radočaj et al. [16], who noted lower R^2 values in independent assessments, emphasizing the need for ground truth validation across diverse geographic contexts.

Independent validation, essential for unbiased accuracy estimation, requires the use of ground truth soil sampling data not utilised in the creation of soil mapping products. While efforts were made to ensure the representativeness of ground truth data in this study, challenges such as mismatched soil depths and landscape heterogeneity could have influenced accuracy assessment outcomes as this study evaluated 0–30cm as a single soil depth rather than SoilGrids soil depth increments (0–5cm, 5–15cm and 15–30cm). Furthermore, the absence of comprehensive global soil sampling programs poses limitations to independent validation efforts, highlighting the need for enhanced data collection programs.

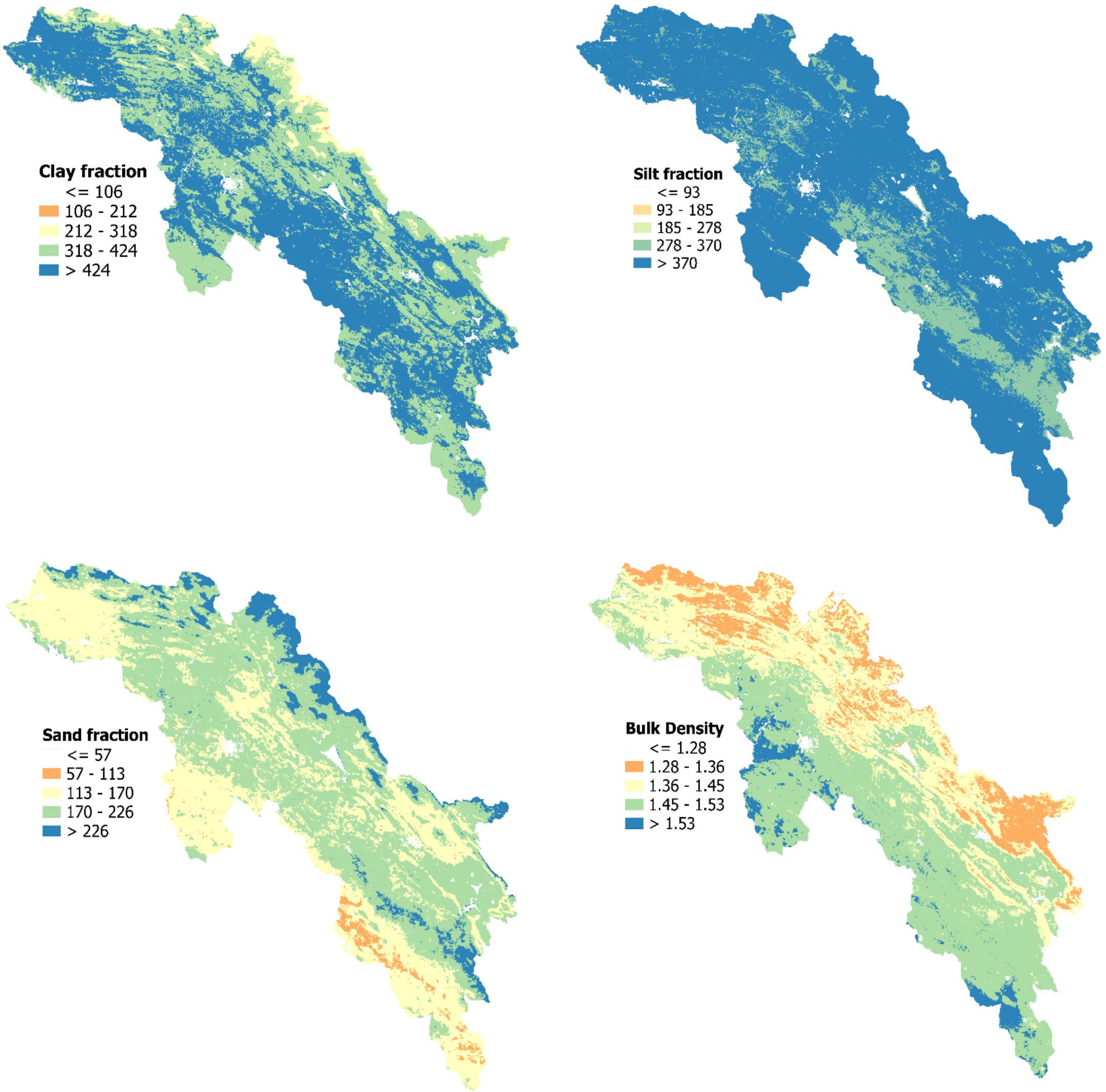

Fig. 2. The map of harmonised pre-processed SoilGrids soil clay, silt, sand fractions (g kg^{-1}) and bulk density (g.cm^{-3}) data used with resolution of 250m

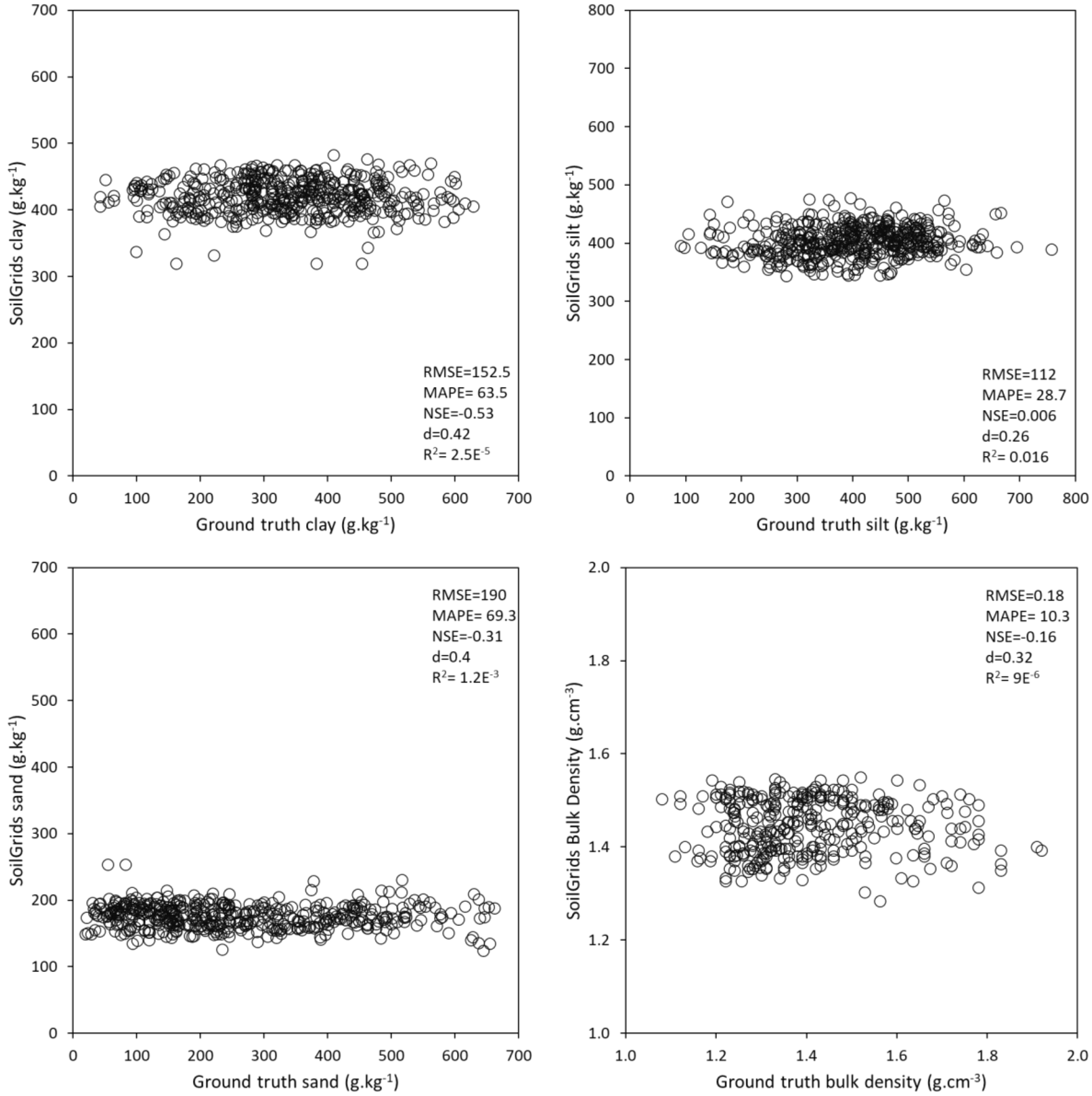

Fig. 3. Ground truth data vs SoilGrids data of soil clay, silt and sand fractions with bulk density

SoilGrids' reliability is important for its applicability in various environmental studies and management practices [19]. Acknowledging the limitations of this study, future research should focus on evaluating SoilGrids' accuracy across different spatial scales, considering factors such as soil types, bio-geolocations, climate classes, and land cover types. Furthermore, conducting digital soil mapping at a national level using comprehensive ground truth soil sampling data could serve as a complementary approach to enhance the reliability of SoilGrids, particularly in regions not adequately represented in its training dataset [5,16]. While SoilGrids offers valuable insights into soil variability on a global scale, its accuracy remains contingent upon the availability and quality of ground truth data. Continued efforts in refining accuracy assessment methodologies and expanding ground truth data coverage will contribute to enhancing the reliability and usability of SoilGrids in addressing various environmental challenges. Acknowledging that factors such as spatial resolution, data coverage, and the representativeness of ground truth data play significant roles in determining the reliability of soil mapping products [12], addressing these challenges requires collaborative efforts between researchers, policymakers, and data providers to improve data quality and enhance the accuracy of soil mapping initiatives [6,5].

The reliability of SoilGrids is crucial for informing land use decisions, agricultural practices, and climate change mitigation strategies in a local scale [6]. Therefore, ensuring the accuracy of soil mapping products is essential for effective resource management and sustainable development in Kurdistan region where the region requires sustainable and productive projects in agriculture, manufacture, and infrastructure sectors. By advancing our understanding of soil properties and their spatial distribution, we can better inform policy decisions and implement sustainable land management practices to mitigate environmental degradation and ensure the long-term health of ecosystems in the region.

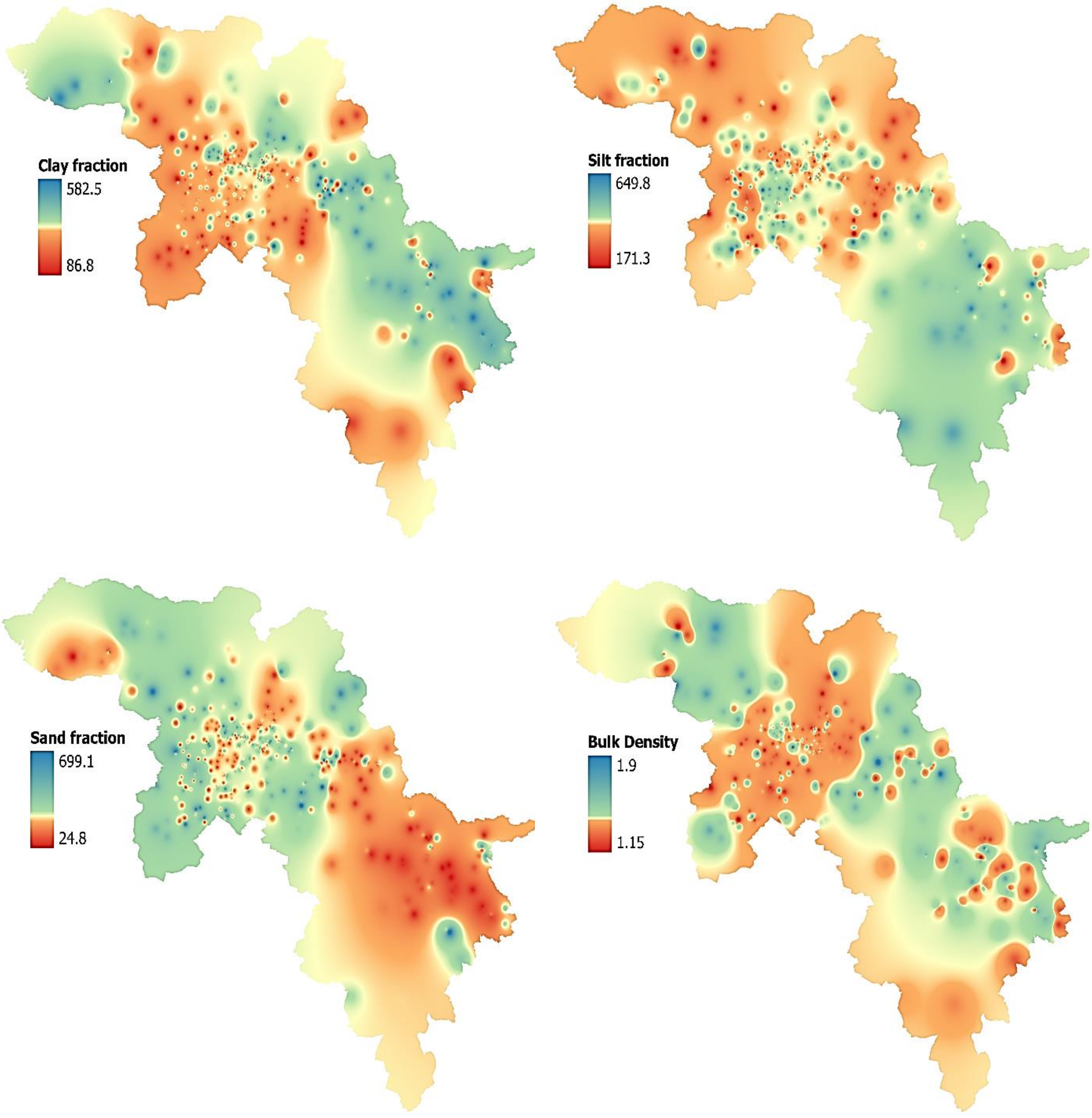

Fig. 4. The map of interpolated ground truth soil clay, silt, sand fractions (g kg^{-1}) and bulk density (g.cm^{-3}) data with resolution of 100m

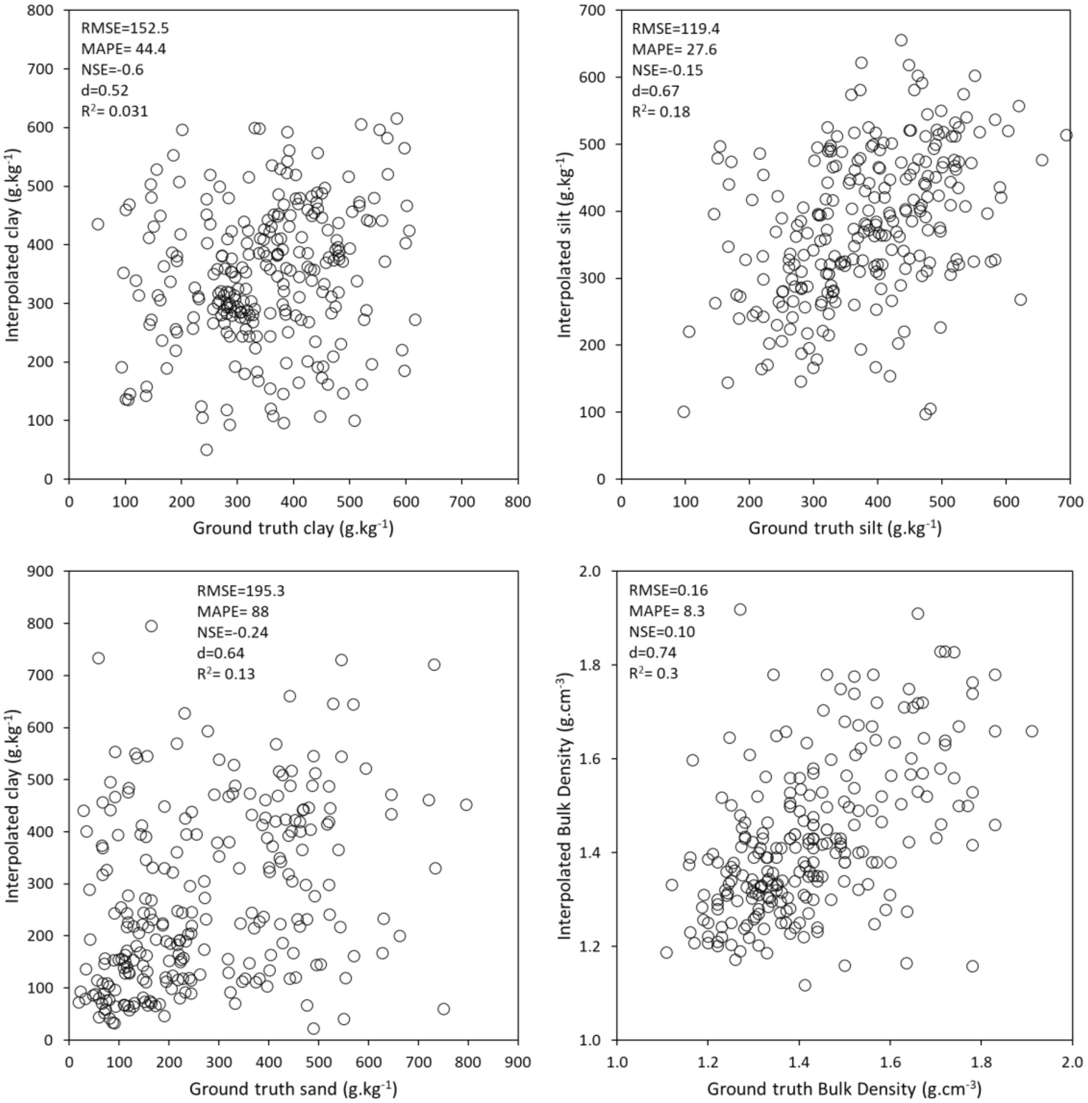

Fig. 5. Ground truth data vs extracted data from interpolated ground truth of soil clay, silt and sand fractions with bulk density

4.2 Independent Validation of SoilGrids

Independent validation serves as a critical step in assessing the accuracy and reliability of SoilGrids products, providing an unbiased estimate of model performance [15,7]. In this study, independent validation involved comparing SoilGrids predictions with ground truth soil sampling data and predicted soil fractions and bulk density using XG Boost algorithm that were not used during the model training process (Skidmore et al., 2002). Challenges arise when conducting independent validation, particularly in regions where SoilGrids was created based on zero soil samples [15,18]. For instance, the absence of soil sampling data in the study area, as documented in the ISRIC WoSIS Soil Profile Database, poses difficulties in accurately validating SoilGrids predictions [6]. Therefore, acknowledging these challenges is essential for transparency in the validation process.

The selection of appropriate ground truth data is paramount to ensure the representativeness of soil variability within the study area [36,37]. Furthermore, discrepancies in spatial resolution between SoilGrids and ground truth data may hinder the assessment of local-scale variations in soil properties [38,7,39]. Then, innovative approaches to address these challenges, such as leveraging supplementary environmental variables or integrating data from alternative soil mapping initiatives could better serve the accuracy of soil mapping [12,40]. Additionally, efforts to expand ground truth datasets through targeted soil sampling campaigns could enhance the accuracy and reliability of independent validation efforts [2,9]. Despite challenges associated with data availability and spatial resolution discrepancies, innovative approaches and expanded ground truth datasets can enhance the reliability of independent validation efforts, ultimately improving our understanding of soil variability and supporting informed decision-making in environmental management for Kurdistan region.

4.3 Independent Validation and Algorithm Prediction

Independent validation is crucial for assessing the accuracy of soil mapping models. Our study used ground truth soil sampling data that were not part of the SoilGrids model training process. This approach ensured an unbiased estimate of model performance [36]. However, challenges can rise due to the absence or insufficient number of soil sampling data in the study area, posing difficulties in accurately validating

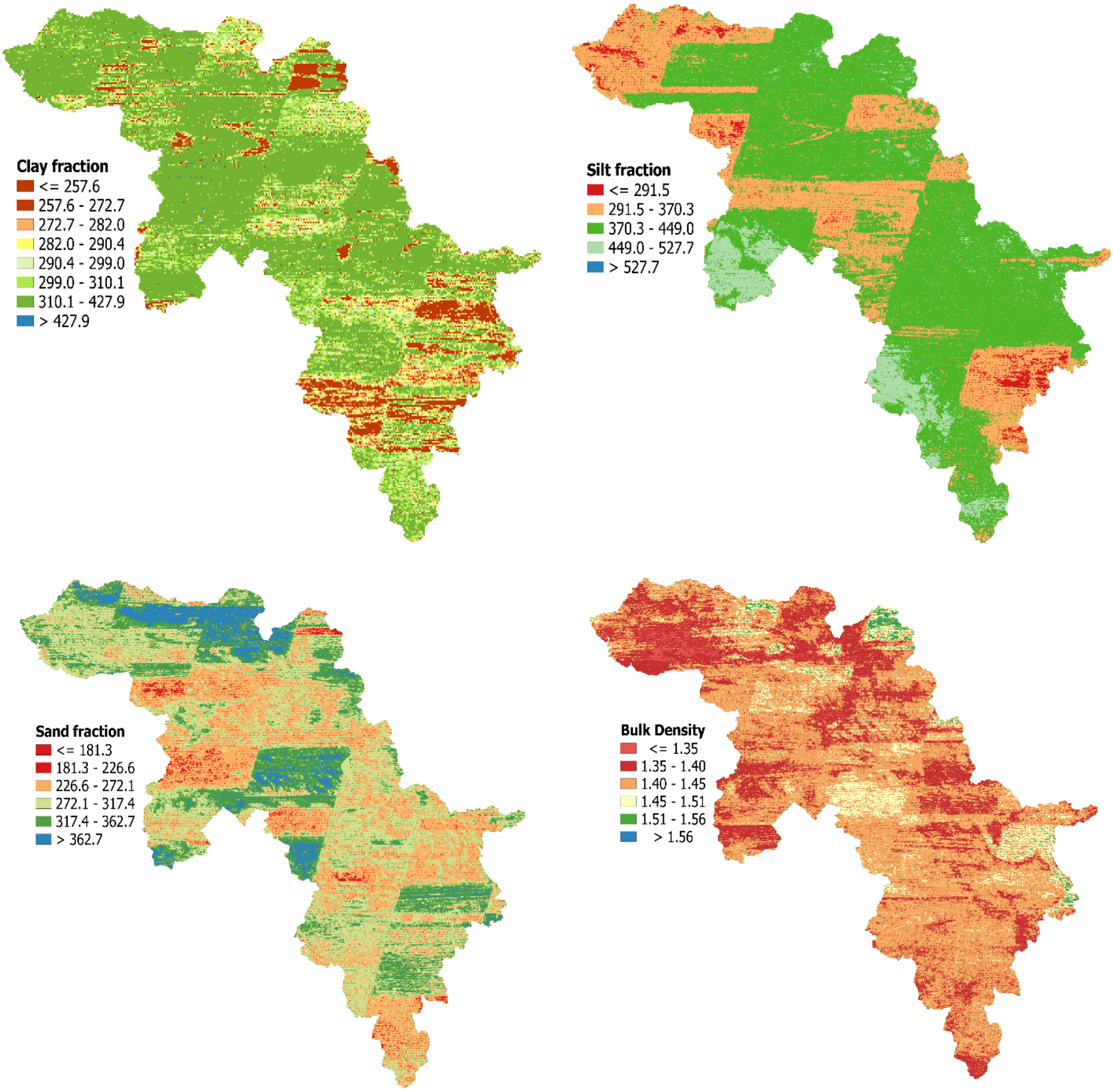

Fig. 6. The map of predicted soil clay, silt, sand fractions (g kg^{-1}) and bulk density (g.cm^{-3}) data using XG Boost algorithm with resolution of 100m

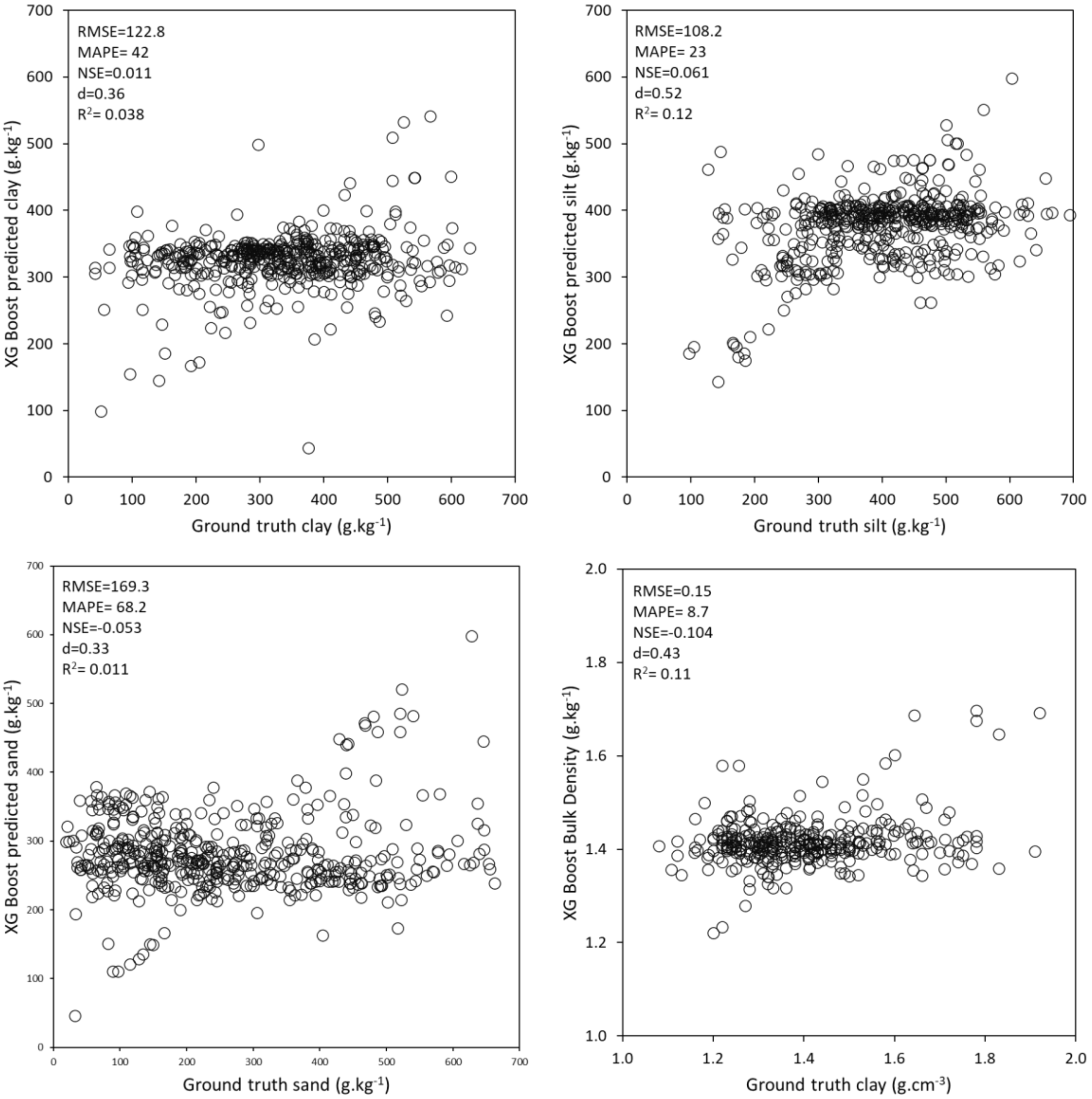

Fig. 7. **Ground truth data vs XG Boost algorithm predicted data of soil clay, silt and sand fractions with bulk density.**

Table 1. The accuracy of SoilGrids layers according to the ground truth soil sampling data, interpolated and XG Boost algorithm predicted. Where GTD: Ground truth data, StDev= Standard deviation, MAD: Mean Absolute Deviation, RMSE: Root Mean Square Error, MAPE: Mean absolute Percentage Error, NSE: Nash Sutcliffe coefficient, d: Index of agreement (Willmontt), R: Correlation, R2: Coefficient of Determination

Soil Properties	Ground truth data vs. SoilGrids data								
	GTD mean (StDev)	SoilGrids mean (StDev)	MAD	RMSE	MAPE	NSE	d	R	R^2
Clay	334.4(123.6)	420.5(25.5)	124.0	152.5	63.5	-0.53	0.42	$5E^{-3}$	$2.5E^{-5}$
Silt	399.8(112.4)	403.0(25.3)	89.9	112.0	28.7	0.006	0.26	0.13	0.016
Sand	265.9(167.1)	176.5(17.8)	144.2	190.6	69.3	-0.30	0.40	-0.03	$1.2E^{-3}$
Bulk density	1.41(0.17)	1.45(0.06)	0.144	0.18	10.3	-0.16	0.32	$3E^{-3}$	$9E^{-6}$

Soil Properties	Ground truth data vs. SoilGrids data								
	GTD mean (StDev)	SoilGrids mean (StDev)	MAD	RMSE	MAPE	NSE	d	R	R^2
	Ground truth data vs. Interpolated data								
Clay	334.4(123.6)	279.6(94.0)	116.6	152.4	44.4	-0.60	0.52	0.18	0.03
Silt	399.8(112.4)	379.3(110.9)	90.8	119.4	27.6	-0.15	0.67	0.42	0.18
Sand	265.9(167.1)	267.9(169.6)	142.3	195.3	88.8	-0.24	0.64	0.35	0.13
Bulk density	1.41(0.17)	1.434(0.170)	0.121	0.164	8.35	0.10	0.74	0.55	0.30
	Ground truth data vs. XG Boost algorithm predicted data								
Clay	334.4(123.6)	325.3(43.2)	97.9	122.7	42.1	0.01	0.36	0.20	0.04
Silt	399.8(112.4)	376.4(52.1)	83.2	108.2	23.1	0.06	0.52	0.35	0.13
Sand	265.9(167.1)	281.2(56.2)	137.5	169.3	68.2	-0.05	0.33	0.1	0.01
Bulk density	1.41(0.17)	1.417(0.06)	0.122	0.155	8.67	0.11	0.43	0.33	0.12

SoilGrids predictions [15]. To address these challenges, machine learning algorithms were employed, particularly the Extreme Gradient Boosting (XG Boost) algorithm, to predict soil fractions and bulk density based on ground truth data. The XG Boost algorithm demonstrated promising results in approximating soil properties, with relatively greater agreement between observed and predicted values [41]. Despite the challenges associated with data availability and spatial resolution discrepancies, the XG Boost algorithm enhanced the accuracy of SoilGrids predictions, particularly in regions with limited ground truth data (Fig. 6 and Fig. 7). By integrating machine learning algorithms and independent validation techniques, we enhance our understanding of soil variability and support evidence-based environmental management strategies [42,40].

The findings were consistent with previous studies assessing the accuracy of SoilGrids products [6,7]. Cross-validation techniques and independent validation demonstrated improvements in prediction accuracy, highlighting the utility of machine learning algorithms in refining soil property estimations (Table 1). While challenges remain, such as spatial resolution disparities and data availability issues, innovative approaches and expanded ground truth datasets can mitigate these limitations [43,9]. However, accurate soil mapping is essential for informed decision-making in environmental management contexts [35]. This study provides valuable insights into the reliability of SoilGrids predictions, particularly in regions with limited ground truth data. By integrating machine learning algorithms and independent validation techniques, we enhance our understanding of soil variability and support evidence-based environmental management strategies [42,40]. However, incorporating auxiliary environmental covariates and integrating data from alternative soil mapping initiatives can further enhance prediction accuracy [38,16]. Additionally, development of international collaboration and data-sharing initiatives along with efforts to conduct targeted soil sampling campaigns and assess soil properties at multiple spatial scales can facilitate access to high-quality soil data and improve the accuracy of global soil mapping efforts [2,1]. Hence, incorporating advanced machine learning algorithms, integrating global and local multi-scale datasets, and expanding ground truth data coverage are essential steps towards enhancing the accuracy and reliability of soil mapping products [44].

5. CONCLUSION

The study assessed digital soil mapping methods in the Kurdistan Region of Iraq, comparing ground truth soil samples with SoilGrids data and using advanced mapping techniques like interpolation and machine learning to improve soil variability understanding. The results revealed significant disparities across soil mineral fractions such as clay, silt, sand fractions, and bulk density. For instance, the mean clay fraction in the ground truth dataset differed notably from SoilGrids' estimation, with a MAD of 124.0 g kg^{-1} and RMSE of 152.5. Similar disparities were observed for other soil properties, underscoring the limitations of global soil mapping initiatives at capturing regional-scale variability accurately. However, the integration of machine learning algorithms, particularly the Extreme Gradient Boosting (XG Boost) algorithm, showed promising results in improving the accuracy of soil property estimations. The XG Boost algorithm exhibited a relatively lower MAD of 97.9 g kg^{-1} for clay fractions, indicating a better approximation of observed values compared to SoilGrids. Additionally, significant percent improvements in RMSE and MAPE values across soil fractions and bulk density measurements underscored the efficacy of machine learning-based approaches in refining soil property estimations within the study area. These findings highlight the importance of leveraging advanced mapping techniques and integrating machine learning algorithms to enhance the accuracy and reliability of soil mapping methodologies. This study further provides valuable insights for informing evidence-based decision-making in environmental management contexts. By incorporating advanced mapping approaches and leveraging machine learning algorithms, we can better support sustainable land management practices and environmental conservation efforts in the region.

HIGHLIGHTS

- Significant disparities between ground truth data and SoilGrids in Kurdistan Region, Iraq.
- Integration of XG Boost algorithm improves accuracy of soil property predictions.
- Machine learning shows promise in refining estimations of soil mineral fractions.
- Novel insights into spatial variability of soil properties help land management.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

ACKNOWLEDGEMENT

The authors express gratitude to the soil team at the Soil and Water Department of the College of Agricultural Engineering Sciences at Salahaddin University-Erbil for their assistance in obtaining soil samples and conducting soil particle size analysis. Authors also extend their gratitude to anonymous landholders and agronomists for granting soil technicians access to their properties.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Searle R, McBratney A, Grundy M, Kidd D, Malone B, Arrouays D, Stockman U, Zund P, Wilson P, Wilford J. Digital soil mapping and assessment for Australia and beyond: A propitious future. Geoderma Regional. 2021;24 e00359.
2. Mahmood F, Khan I, Ashraf U, Shahzad T, Hussain S, Shahid M, Abid M, Ullah S. Effects of organic and inorganic manures on maize and their residual impact on soil physico-chemical properties. Journal of soil science and plant nutrition. 2017;17(1):22-32.
3. Bennett JM, McBratney A, Field D, Kidd D, Stockmann U, Liddicoat C, Grover S. Soil security for Australia. Sustainability. 2019;11(12):3416.
4. Malone B, Stockmann U, Glover M, McLachlan G, Engelhardt S, Tuomi S. Digital soil survey and mapping underpinning inherent and dynamic soil attribute condition assessments. Soil Security. 2022;6:100048.
5. Kidd D, Searle R, Grundy M, McBratney A, Robinson N, O'Brien L, Zund P, Arrouays D, Thomas M, Padarian J. Operationalising digital soil mapping–Lessons from Australia. Geoderma Regional. 2020;23:e00335.
6. Hengl T, Mendes de Jesus J, Heuvelink GB, Ruiperez Gonzalez M, Kilibarda M, Blagotić A, Shangguan W, Wright MN, Geng X, Bauer-Marschallinger B. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One. 2017;12(2):e0169748.
7. Poggio L, De Sousa LM, Batjes NH, Heuvelink GB, Kempen B, Ribeiro E, Rossiter D. SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. Soil. 2021;7(1):217-240.
8. van der Voort TS, Verweij S, Fujita Y, Ros GH. Enabling soil carbon farming: Presentation of a robust, affordable, and scalable method for soil carbon stock assessment. Agronomy for Sustainable Development. 2023;43(1):22.
9. Tifafi M, Guenet B, Hatté C. Large differences in global and regional total soil carbon stock estimates based on SoilGrids, HWSD, and NCSCD: Intercomparison and evaluation based on field data from USA, England, Wales, and France. Global Biogeochemical Cycles. 2018;32(1):42-56.
10. Arrouays D, Mulder VL, Richer-de-Forges AC. Soil mapping, digital soil mapping and soil monitoring over large areas and the dimensions of soil security–A review. Soil Security. 2021;5 :100018.

11. Vågen TG, Winowiecki LA. Mapping of soil organic carbon stocks for spatially explicit assessments of climate change mitigation potential. Environmental Research Letters. 2013;8(1):015011.
12. Heung B, Saurette D, Bulmer CE. Digital soil mapping. Digging into Canadian Soils; 2021.
13. Pereira P, Brevik EC, Muñoz-Rojas M, Miller BA, Smetanova A, Depellegrin D, Misiune I, Novara A, Cerdà A. Soil mapping and processes modeling for sustainable land management. In 'Soil mapping and process modeling for sustainable land use management'. 2017;29-60.
14. Hengl T, Heuvelink GB, Kempen B, Leenaars JG, Walsh MG, Shepherd KD, Sila A, MacMillan RA, Mendes de Jesus J, Tamene L. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. PLoS One. 2015;10(6):e0125814.
15. Hengl T, De Jesus JM, MacMillan RA, Batjes NH, Heuvelink GB, Ribeiro E, Samuel-Rosa A, Kempen B, Leenaars JG, Walsh MG. SoilGrids1km—global soil information based on automated mapping. PLoS One. 2014;9 (8):e105992.
16. Radočaj D, Jurišić M, Rapčan I, Domazetović F, Milošević R, Plaščak I. An independent validation of soilgrids accuracy for soil texture components in Croatia. Land. 2023;12(5):1034.
17. Buenemann M, Coetzee ME, Kutuahupira J, Maynard JJ, Herrick JE. Errors in soil maps: The need for better on-site estimates and soil map predictions. PLoS One. 2023;18 (1):e0270176.
18. Maynard JJ, Yeboah E, Owusu S, Buenemann M, Neff JC, Herrick JE. Accuracy of regional-to-global soil maps for on-farm decision-making: are soil maps “good enough”? Soil. 2023;9(1):277-300.
19. Liang Z, Chen S, Yang Y, Zhou Y, Shi Z. High-resolution three-dimensional mapping of soil organic carbon in China: Effects of SoilGrids products on national modeling. Science of the total environment. 2019;685:480-489.
20. Chen S, Arrouays D, Mulder VL, Poggio L, Minasny B, Roudier P, Libohova Z, Lagacherie P, Shi Z, Hannam J. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. Geoderma. 2022;409:115567.
21. Surucu A, Ahmed TK, Gunal E, Budak M. Spatial variability of some soil properties in an agricultural field of Halabja City of Sulaimania Governorate, Iraq. Fresenius Environment Bulletin. 2019;28(1):193-206.
22. Wang F, Wei Y, Yang S. Enhanced understanding of key soil properties in Northern Xinjiang using water-heat-spectral datasets based on bioclimatic guidelines. Land. 2023;12(9):1769.
23. Yu W, Zhou W, Wang T, Xiao J, Peng Y, Li H, Li Y. Significant improvement in soil organic carbon estimation using data-driven machine learning based on habitat patches. Remote Sensing. 2024;16(4):688.
24. Salcedo-Sanz S, Ghamisi P, Piles M, Werner M, Cuadra L, Moreno-Martínez A, Izquierdo-Verdiguier E, Muñoz-Marí J, Mosavi A, Camps-Valls G. Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources. Information Fusion. 2020;63:256-272.
25. Corwin D, Lesch S, Oster J, Kaffka S. Monitoring management-induced spatio–temporal changes in soil quality through soil sampling directed by apparent electrical conductivity. Geoderma. 2006;131(3-4):369-387.
26. Upadhyay S, Raghubanshi A. Determinants of soil carbon dynamics in urban ecosystems. In 'Urban ecology'. 2020;299-314.
27. Jirjees S, Seeyan S, Fatah K. Climatic analysis for Pirmam area, Kurdistan Region, Iraq. The Iraqi Geological Journal. 2020;75-92.
28. Marsh A, Altaweel M. The search for hidden landscapes in the Shahrizor: Holocene land use and climate in Northeastern Iraqi Kurdistan. New Agendas in Remote Sensing and Landscape Archaeology in the Near East. 2020;7.
29. Forti L, Perego A, Brandolini F, Mariani GS, Zebari M, Nicoll K, Regattieri E, Barbaro CC, Bonacossi DM, Qasim HA. Geomorphology of the northwestern Kurdistan Region of Iraq: landscapes of the Zagros Mountains drained by the Tigris and Great Zab Rivers. Journal of Maps. 2021;17(2):225-236.
30. Nasir SM, Kamran KV, Blaschke T, Karimzadeh S. Change of land use/land cover in kurdistan region of Iraq: A semi-automated object-based approach.

Remote Sensing Applications: Society and Environment. 2022;26:100713.

31. Fadhil AM. Drought mapping using Geoinformation technology for some sites in the Iraqi Kurdistan region. International Journal of Digital Earth. 2011;4(3):239-257.
32. Gee G, Bauder J. Particle-size analysis. In 'Methods of soil analysis. Part 1. Physical and mineralogical methods'.(Ed. A Klute). Soil Science Society of America: Madison, WI. 1986;383–411
33. Chen T, Guestrin C. 'Xgboost: Reliable large-scale tree boosting system, Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA'; 2015.
34. Fung KF, Chew KS, Huang YF, Ahmed AN, Teo FY, Ng JL, Elshafie A. Evaluation of spatial interpolation methods and spatiotemporal modeling of rainfall distribution in Peninsular Malaysia. Ain Shams Engineering Journal. 2022;13(2):101571.
35. Montanarella L, Vargas R. Global governance of soil resources as a necessary condition for sustainable development. Current opinion in environmental sustainability. 2012;4(5):559-564.
36. Skidmore AK. Accuracy assessment of spatial information. In 'Spatial statistics for remote sensing'. 1999;197-209.
37. Das B, Murgaonkar D, Navyashree S, Kumar P. Novel combination artificial neural network models could not outperform individual models for weather-based cashew yield prediction. International Journal of Biometeorology. 2022;66(8):1627-1638.
38. Bogunovic I, Trevisani S, Seput M, Juzbasic D, Durdevic B. Short-range and regional spatial variability of soil chemical properties in an agro-ecosystem in eastern Croatia. Catena. 2017;154:50-62.
39. Xu C, Torres-Rojas L, Vergopolan N, Chaney NW. The benefits of using state-of-the-art digital soil properties maps to improve the modeling of soil moisture in land surface models. Water resources research. 2023;59(4):e2022WR032336.
40. Lu L, Li S, Wu R, Shen D. Study on the scale effect of spatial variation in soil salinity based on geostatistics: A case study of Yingdaya River Irrigation Area. Land. 2022;11(10):1697.
41. Chen T, Guestrin C. 'Xgboost: A scalable tree boosting system, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining'; 2016.
42. Wang Z, Shi W, Zhou W, Li X, Yue T. Comparison of additive and isometric log-ratio transformations combined with machine learning and regression kriging models for mapping soil particle size fractions. Geoderma. 2020;365:114214.
43. Montanarella L. The global soil partnership, IOP Conference Series: Earth and Environmental Science; 2015.
44. Geng X. Development of operational methods to predict soil classes and properties in Canada using machine learning. Carleton University; 2020.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the publisher and/or the editor(s). This publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/125367