 DOI: https://doi.org/10.4038/ta.v171i4.48

Research Paper

R-Modelling-Based Quantile Regression Approach for Spatial Prediction and Mapping Soil Properties in Sri Lanka

H.M.A.H. Uduwerella*, A.U. Iddawela, H.K. Kadupitiya

Natural Resources Management Centre, Peradeniya, Sri Lanka

Corresponding author: harshaniahu@gmail.com

ARTICLE INFO

Article history:
Received: 31 May 2023
Revised version received: 08 July 2023
Accepted: 24 August 2023

Citation:
Uduwerella, H.M.A.H., Iddawela, A.U., Kadupitiya, H.K. (2023). R-Modelling-Based quantile regression approach for spatial prediction and mapping soil properties in Sri Lanka. *Tropical Agriculturist, 171* (4), 18-31

Abstract

Reliable soil information is crucial for agriculture planning, decision-making, food insecurity, and environmental impact mitigation. Digital Soil Mapping (DSM) solves data gaps and offers an efficient approach to predict soil variability. This study evaluated the potential of DSM, based on the quantile regression forest (QRF-DSM) approach in the R-modelling language for populating gridded soil properties with sparsely distributed observations (soil profile data from 100 locations) and 67 environmental covariates related to soil forming factors to predict and map bulk density (BD), texture fractions, cation exchange capacity (CEC), soil organic carbon (SOC), and pH of topsoil covering entire Sri Lanka. The optimal covariates for each soil property were selected with the recursive feature elimination (RFE) algorithm, and cross-validation was employed to assess model performance. The selected model showed a positive correlation with both Pearson's product-moment correlation coefficient (r) and coefficient of determination (r2). This study ensured precise predictions for the topsoil layer, achieving low mean absolute error (MAE) and low root mean squared error (RMSE) for soil properties: BD (0.11, 0.15), sand (9.33, 13.23), silt (4.29, 5.65), clay (7.12, 9.84), CEC (3.67, 5.32), SOC (0.48, 0.56), pH (0.63, 0.75), and generating mean and uncertainty maps at a 250 m resolution for Sri Lanka. The predicted gridded soil property maps provide information on spatial variability for diverse applications.

Keywords: **Digital soil mapping, Quantile regression approach, R-modelling, Soil properties**

Introduction

Spatial variability of soil attributes is required for many applications, however, the soil property data is mostly available as point-measured data. The multi-functionality of the soil influenced by its chemical, physical, and biological properties, enables it to provide diverse services in agriculture, forestry, water preservation, resource management, and environmental sustainability (Calzolari *et al*., 2016; Nussbaum *et al*., 2018). The physical and chemical properties of soil influence the conditions and resources available to plants, impacting their response to disturbances (Bueno *et al*., 2013). The physical, chemical, and physicochemical properties of soil, including soil BD, soil reaction (pH), texture, SOC content, and CEC, are essential determinants of nutrient availability in soil (Robertson *et al*., 1999). Spatial variability of soil information supports the implementation of localised optimization for soil management strategies, which is imperative for enhanced resource use efficiency in agriculture (FAO, 2022). Digital Soil Mapping (DSM) or predictive soil mapping provides vital tools to populate soil property maps and develop spatial soil information with available field and laboratory observations (Biswajit & Divya, 2020). The DSM, as a solution for the high costs and long timescales associated with traditional soil mapping, is an evolving field of soil science through innovative research in recent years and has experienced a significant surge in popularity as a means of acquiring comprehensive soil information maps (McBratney *et al*., 2003; Grimm & Behrens, 2010; Minasny & McBratney, 2016; Poggio *et al*., 2021).

The DSM based on the quantile regression forest (QRF-DSM) approach considers the underlying dependence structure of soil properties that produces consistency predictions during multi-property estimations (Kasraei *et al*., 2021). According to Meinshausen (2006), the use of QRF has been tested and suggested particularly for regions where the locations of soil observations are inadequately sparse for the application of other soil mapping approaches. The QRF-DSM is better than other DSM approaches such as regression kriging (RK), universal kriging (UK), sequential Gaussian simulation (SGS), and random forest combined with kriging (RFK). In addition to producing soil property maps, QRF is capable of providing uncertainty and probability distribution (Szatmári & Pásztor, 2019; Angelini *et al*., 2022). This study evaluated the potential of the QRF-DSM approach for populating gridded soil properties with sparsely distributed observations and a large number of covariates to predict and map SOC, pH, CEC, BD, and texture fractions of topsoil across Sri Lanka.

Materials and Methods

The QRF technique, firstly presented by Koenker and Bassett (1978), was utilised in this study to predict soil parameters and evaluate uncertainty levels. By combining observed values with location-coordinates and a number of highly correlated environmental variables, this approach uses a number of tools to populate gridded soil property maps. The QRF statistically estimates the different quantiles (percentiles) of a soil property across a study area, rather than just the mean. The approach involves pre-processing and harmonising soil data, selecting relevant covariates, developing a quantile regression model, fitting the model, predicting and mapping the quantiles, and evaluating the model's performance. Figure 1 illustrates the methodological flow adapted for generating soil property and uncertainty maps using the QRF approach.

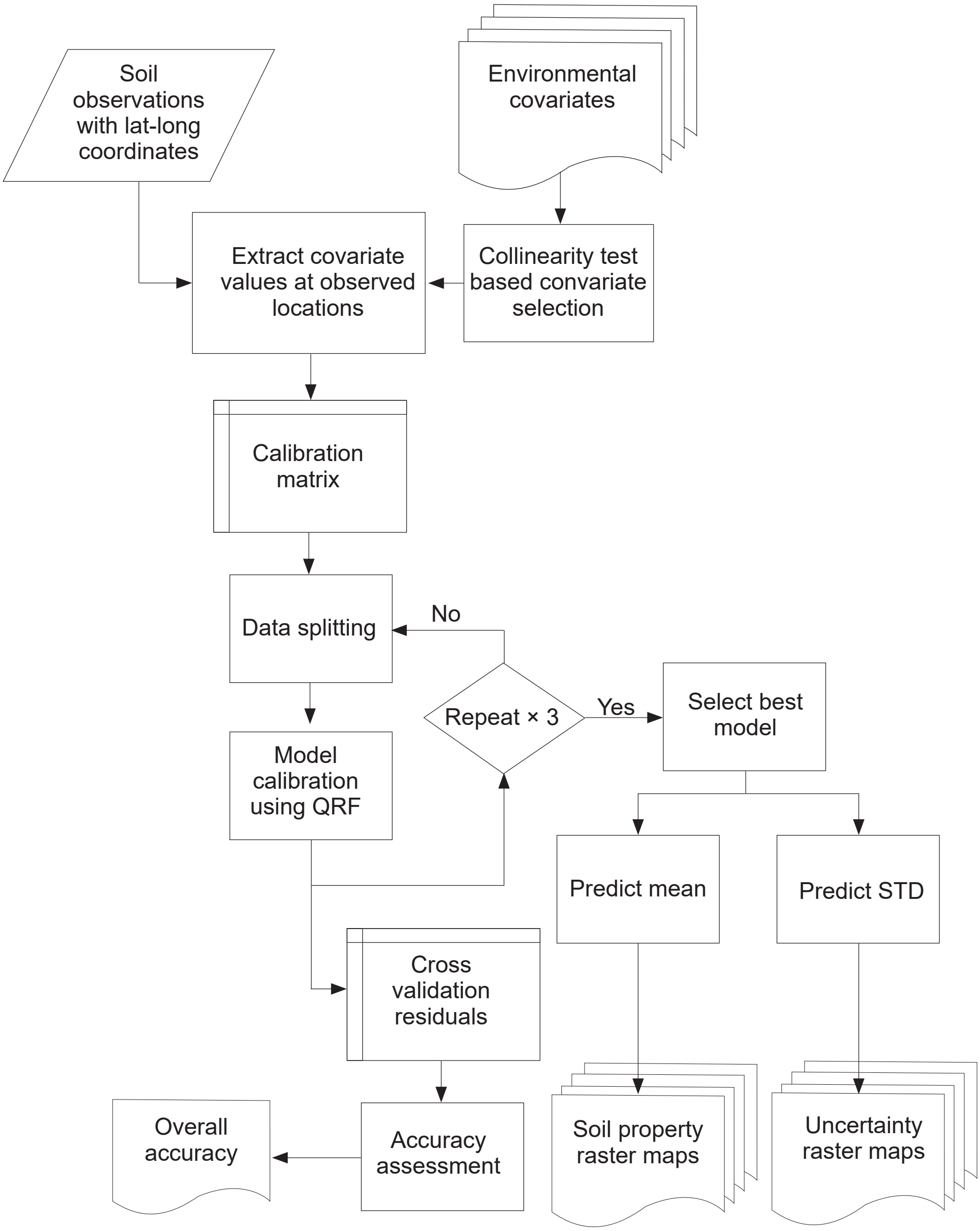

Figure 1. Methodological approach for generating soil property maps and uncertainty maps using the QRF approach

Soil profile data and environmental covariates

In the present study, profile-wise soil property observations from 100 locations across Sri Lanka were used as the input observed data for BD, texture fractions, CEC, SOC, and pH for the QRF modelling approach. The distribution of observation locations is shown in Figure 2. Observed soil property data were extracted from soil profile data sets in the Soil Science Society Soil Factsheet Books (Senarath *et al*., 1998; Dassanayake *et al*., 2001, 2006). The soil profile dataset, along with associated location coordinates, was compiled in a comma-separated value (CSV) file format. This file served as the observed soil property dataset for analysis within the R modelling framework using the R-Studio free software (https://cran.r-project.org/ and https://www.rstudio.com/products/rstudio/download/).

The 'ithir' package (http://R-Forge.R-project.org) in the R environment was employed to harmonise the data pertaining to the top-most layers for developing observed soil property data representing 0–30 cm layers using the mass-preserving spline function (Bishop *et al*., 1999; Malone *et al*., 2009; FAO, 2022). As illustrated by Mcbratney *et al*. (2003), the SCORPAN model was followed to develop relationships with each predicting soil property and its associated environmental factors. Soil, climate, organisms, relief, parent material, age, and spatial location are the seven environmental covariate groups used for the development of empirical quantitative models (Kienast-Brown *et al*., 2017).

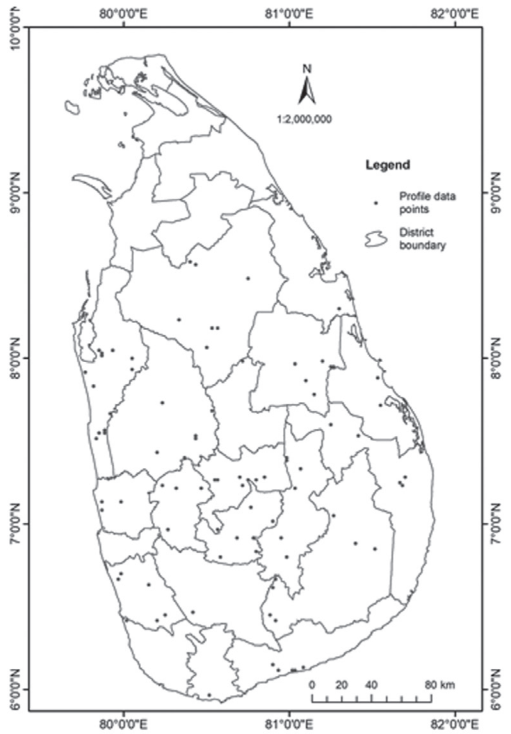

Figure 2. Distribution of Sampling Sites in Sri Lanka

Based on these environmental covariate groups, Angelini *et al*. (2022) and FAO (2022) modelled the spatial soil property predictions using 67 covariate layers, and the same covariate list was used for this study (Table 1).

Table 1. Environmental covariate groups used in the study and sources of data

No	Variable group	Description	Resolution (m)	Source/ Development method
1	Climatic data	Monthly/yearly/seasonal temporal mean and standard deviation of precipitation, temperatures (min, max, mean), evapotranspiration (min, max, mean, range), wind, aridity index. (16 layers)	1000	https://chelsa-climate.org
2	Vegetation indices (NDVI - MOD13Q1)	Mean and standard deviation (8 layers)	250	USGS portal (reverb.echo. nasa.gov)
3	Fraction of photosynthetically active radiation (FPAR - MOD15A2H)	Mean and standard deviation. (8 layers)	500	USGS portal (reverb.echo. nasa.gov)
4	Land surface temperature (LST) day -MOD11A2)	Mean and standard deviation. (8 layers)	1000	USGS portal (reverb.echo. nasa.gov)
5	Normalised difference between LST day and LST night (MOD11A2)	Mean and standard deviation. (8 layers)	1000	USGS portal (reverb.echo. nasa.gov)
6	Short-wave Infrared (SWIR) black-sky albedo for shortwave broadband (MCD43A3)	Mean June-August from 2000-2022	500	USGS portal (reverb.echo. nasa.gov)
7	Dynamic World 10m near-real-time land use/ land cover (LULC)	Mean estimated probability of complete coverage by trees, shrub and scrub, flooded vegetation, grasses and bare. (5 layers)	250	https://code. earthengine. google.com/
8	Terrain & elevation data	Profile curvature, Downslope curvature, Upslope curvature, Deviation from mean value, Elevation, Melton ruggedness, Negative openness, slope, Topographic position index, Terrain wetness index, Multiresolution of valley bottom flatness. (13 layers)	250	https:// opengeohub. org/about-openlandmap/

The remote sensing-based covariate layers were downloaded from the "United States Geological Survey" free data portal (https://lpdaac.usgs.gov, accessed on 20th Feb 2023) via Google Earth Engine (Gorelick *et al*., 2017), clipped using the Sri Lankan boundary vector file, and saved for 250 m resolution and the WGS 84 system.

Covariate selection and model calibration

The process of covariate selection and model calibration for the target soil attribute has consisted of several steps. In the initial step, the stacked covariate layers were overlaid with point data to extract information corresponding to each location and the pixels of the layers. Then, a data frame was created that incorporated to the relevant data on soil properties along with 67 covariates, which were then utilised in further analyses and modelling processes. The RFE algorithm was used to select an optimal set of covariates for regression tree models (Gomes *et al*., 2019; Ramezan, 2022) based on the repeated k-fold cross-validation technique, a statistical approach for evaluating and comparing learning algorithms by randomly splitting the data into training and testing sets (Refaeilzadeh *et al*., 2009). In this study, three iterations of the 10-fold cross-validation procedure were implemented in the R package caret (https://cran.r-project.org/) (Zhang *et al*., 2023).

The regression forest RFE model was calibrated by assigning a set of parameters. A range of feature subset sizes that will be evaluated during the RFE process was set as a sequence (from = 10, to = length (ncovs) - 1, by = 5). Accordingly, the RFE process was performed for subsets starting at 10 and incrementing by 5. Iteratively removed unimportant predictors, refitted the model, assessed performance repeatedly, and finally obtained the resulting information on optimal covariates and their rankings using the model with the best performance.

The QRF is a decision tree-based statistical model frequently utilised within the machine learning approach (Meinshausen, 2006). The model building and calibration process includes the formula that specifies the relationship between the independent variables (predictors or covariates) and the dependent variable (soil attribute) to be modelled, training data (a data frame with selected covariates and target soil property), and ‘tuneGrid’ to tune the model's hyperparameter ‘mtry’ value. The parameter ‘mtry’ determines the number of covariates used at each split in regression trees (Angelini *et al*., 2022). In this study, the value of ‘mtry’ was obtained by dividing the number of optimal covariates by 3 and then rounding to the nearest integer. The tuneGrid consists of the ‘mtry’ value, 5 less than that value, and 5 more than that value. The tuning process involves training multiple RF models using each combination of ‘mtry’ specified in the tuneGrid. This allows the model to be trained and evaluated with different values of ‘mtry’ to determine which value leads to the best performance. The hyper-parameters yielding the minimal root mean squared error (RMSE) were ultimately selected. Model tuning was performed with a 10-fold cross-validation procedure with three iterations. The dataset was randomly divided into roughly equal subsets, forming 10 folds, with one part used for testing and the remaining nine parts used for training the model. In each individual splitting step, the training data are employed for model calibration, while the testing data are subsequently utilised with the calibrated model to generate the residuals (FAO, 2022). Validation results were summarised for all folds, and the best model was selected for soil property prediction.

Accuracy assessment

The overall accuracy of model performance was estimated using the model residuals generated during the repeated cross-validation step. Wadoux *et al*. (2022) evaluated accuracy statistics for map quality and found no single measure to explain all aspects. As per the recommendations provided by them, five performance indicators, including mean prediction error (ME), mean absolute error (MAE), root-mean squared error (RMSE), coefficient of determination (r2), and Pearson's product-moment correlation coefficient (r), were used to evaluate the model's performance (Equations 2-5; Angelini *et al*., 2022). A good model should have a high r2 and a low RMSE (Kienast-Brown *et al*., 2017). The indices quantify the error ϵ between the observed (z) value and predicted (ẑ) value of soil property S at location i (Equations 1).

$$e(S_i) = z(S_i) - \hat{z}(S_i) \quad \text{........... (Equation 1)}$$

$$ME = \frac{1}{N}\sum_{i=1}^{N} e(S_i) \quad \text{............ (Equation 2)}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N} [e(S_i)] \quad \text{............ (Equation 3)}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e(S_i)^2} \quad \text{............ (Equation 4)}$$

$$r = \frac{\sum_{i=1}^{N}(z(\mathbf{s}_i)-\overline{z})(\hat{z}(\mathbf{s}_i)-\overline{\hat{z}})}{\sqrt{\sum_{i=1}^{N}(z(\mathbf{s}_i)-\overline{z})^2}\sqrt{(\hat{z}(\mathbf{s}_i)-\overline{\hat{z}})^2}} \quad \text{(Equation 5)}$$

As part of the accuracy assessment, standard deviation maps of the prediction were developed to reveal the spatial distribution of uncertainty.

Prediction of soil properties

After the model calibration, the best set of parameters that fit the model was selected using the whole dataset. Subsequently, the resulting model was used to successfully predict the target soil properties, and finally, the predicted mean maps and standard deviation maps were produced. The final maps were produced using ArcGIS 10.1.3 software.

Results and Discussion

The soil profile data employed in this study were subjected to quality control and harmonised to 0-30 cm in R, and the descriptive statistics that resulted are reported in Figure 3. The number of environmental covariates used was contingent upon soil properties, which were selected using the RFE algorithm: BD (25), sand (25), silt (20), clay (30), CEC (30), SOC (67) and pH (55). The optimal model hyper-parameter 'mtry' values chosen for each soil property were 3, 3, 2, 15, 15, 17, and 18, respectively.

Accuracy Assessment results

The evaluation of model performance was based on an assessment of the observed and predicted values of the soil property. Figure 4 illustrates the scatterplots of the observed data and the corresponding predictions.

The correlation analysis between predicted and observed values for all soil properties demonstrated a positive correlation (Table 2), as indicated by both the r and r^2. The r^2 value of 1 (r ranges from -1 to +1) signifies a complete positive correlation between predicted and observed values, while 0 indicates no correlation between them (Angelini *et al*., 2022; Wadoux *et al*., 2022). The r^2 statistic can be misleading for map accuracy as it measures variance around the linear regression line but ignores deviations from the equality line ($z = \hat{z}$) (Wadoux *et al*., 2022).

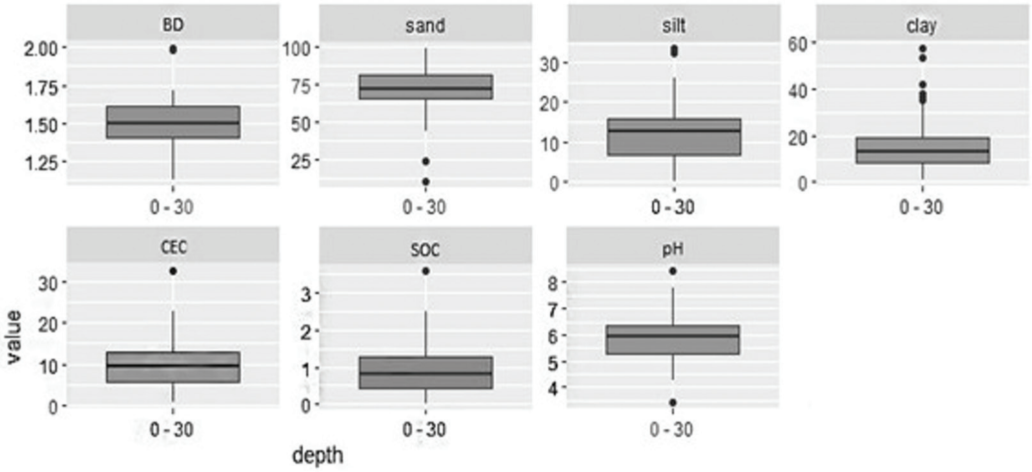

Figure 3. Descriptive statistics of harmonised soil data

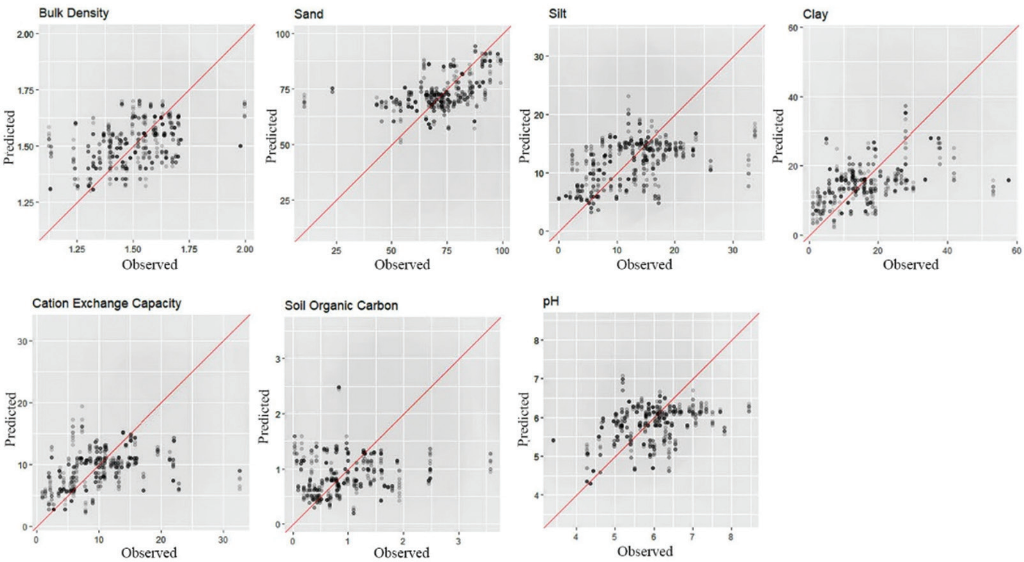

Figure 4. Scatter plots of observed versus predicted soil properties

Table 2. Accuracy indicators

Soil Property	Unit	ME	MAE	RMSE	r	r^2
BD	gcm^{-3}	0.02	0.11	0.15	0.45	0.2
Sand	percent	1.41	9.33	13.23	0.48	0.23
Silt	percent	0.61	4.29	5.65	0.52	0.27
Clay	percent	1.23	7.12	9.84	0.44	0.19
CEC	cmolc/kg	0.77	3.67	5.32	0.41	0.17
SOC	percent	0.07	0.48	0.56	0.43	0.19
pH	NA	0.07	0.63	0.75	0.42	0.18

Average error indices quantify the average deviation between observed and predicted values of each soil property. The model's accuracy in predicting these properties is supported by the MAE values, which are closer to zero for all properties, indicating better accuracy. The ME estimates the prediction bias, while the MAE and RMSE estimate the magnitude of errors (Angelini *et al*., 2022; Wadoux *et al*., 2022). Moreover, the low RMSE values indicated a minimal error in the model's predictions, further emphasising its accuracy and reliability.

Soil property maps for Sri Lanka

Predictive mean (Figure 5) and predicted standard deviation (Figure 6) maps were generated by utilising the chosen models with the above-mentioned parameters. The final maps provided the spatial distribution of key soil properties, enabling a comprehensive understanding of their patterns and variations throughout the study area. The concentration of BD showed a considerable significance in the Kurunegala and Puttalam regions, while it is low in wet zone areas. Typically, soils with increased clay content tend to exhibit lower bulk density measurements (Walpola & Mendis, 2020). When comparing with the resulting clay map, it was evident that areas with low BD are similar to areas with high clay content. Sakin (2012) reported the existence of an inverse correlation between bulk density and the content of organic matter which also aligns with the outcomes of the current study.

Well-reflected changes were observed in the spatial distribution of SOC and pH. The SOC is relatively moderate in the western, central, and southwest areas, with some patches of high values prominently observed. This observation resembles the results reported by Kadupitiya *et al*. (2018). The pH distribution exhibits an inverse pattern compared to that of SOC. The lowest pH level occurred in southwest areas, while patches of high pH values were prominent in Polonnaruwa, Jaffna, and Trincomalee districts. Red-yellow podzolic soils are the dominant soil type in the wet zone and extend into highland areas of the intermediate zone, often exhibiting a pH value below 5.5. Reddish brown earths (RBE), common in arid areas, generally display pH levels 6–7 and are strongly alkaline due to the presence of crystalline limestone (Moormann & Panabokke, 1961). Alfisols are the model soil of the dry zone, while Ultisols dominate the wet zone of Sri Lanka. The heavy rainfall in the wet zone can lead to a decrease in pH due to the leaching of basic cations from Ultisols, and Alfisols show a comparatively higher pH and base saturation (Indraratne, 2020).

Sand distribution is more prominent in the coastal areas and less prominent in the mid-country. Several patches of high silt content areas can be found within the country, while most of the borderline areas have a low silt percentage. Conversely, the clay content is relatively high in the wet zone regions.

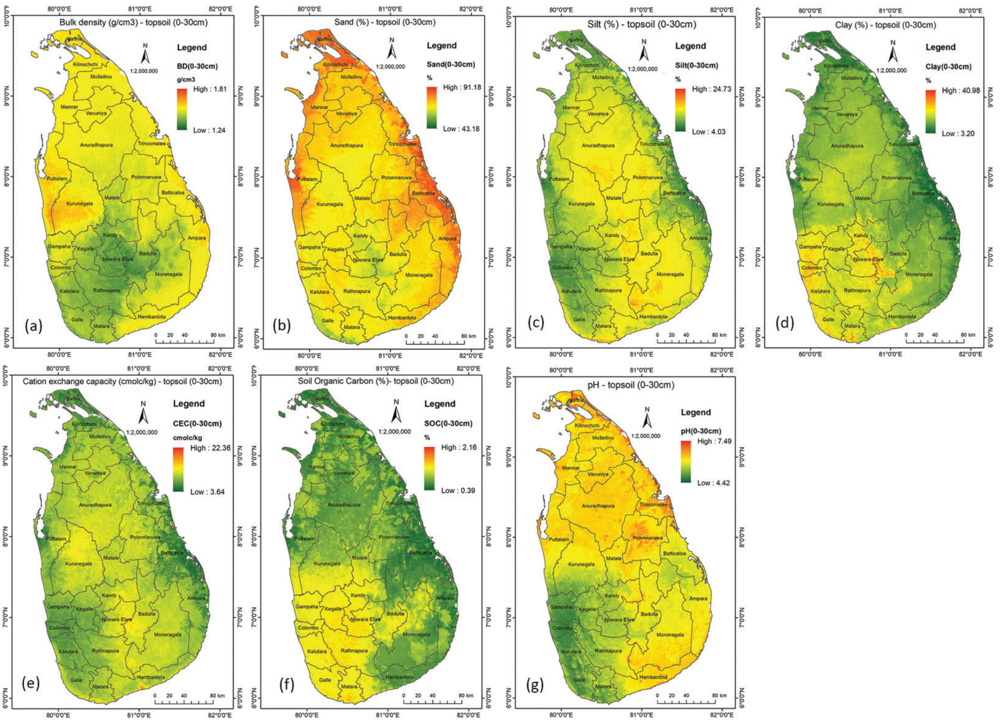

Figure 5. Predicted mean soil properties within the 0-30cm depth interval for Sri Lanka. (a) BD, (b) sand, (c) silt, (d) clay, (e) CEC, (f) SOC, and (g) pH

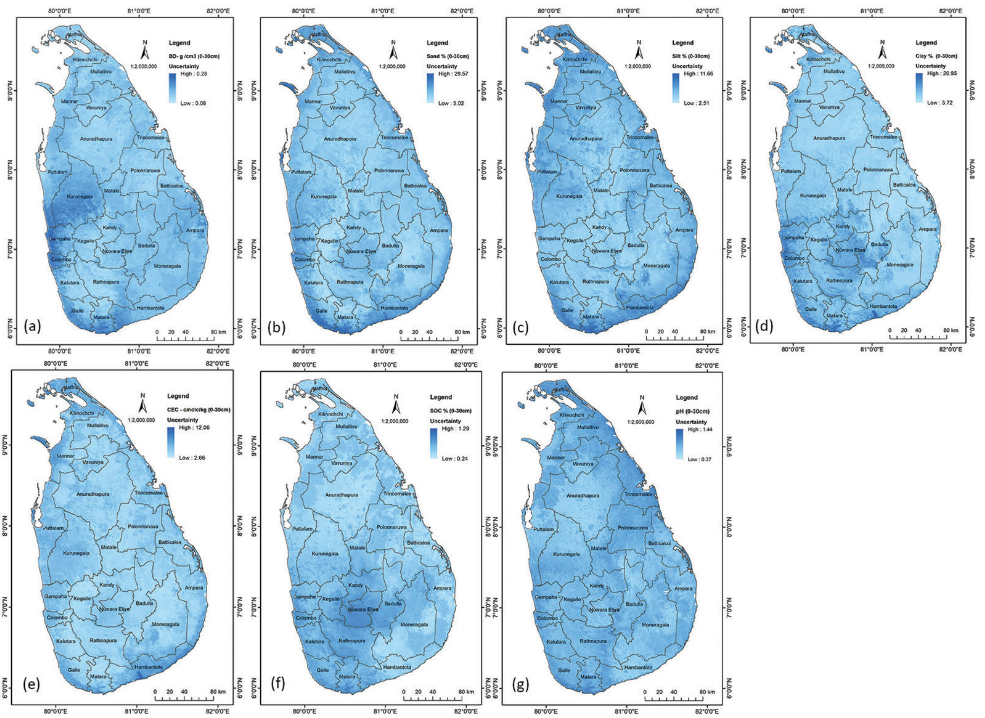

Figure 6. The uncertainty (predicted standard deviation) maps generated by the QRF algorithm (a) BD, (b) sand, (c) silt, (d) clay, (e) CEC, (f) SOC, and (g) pH

Approximately one-third (33%) of the coastal area is marked by sandy shorelines (Swan, 1983; Chandrajith, 2020). Ultisols are rich in kaolinite and possess a limited cation exchange capacity (Indraratne, 2020). The CEC generally exhibits low values and shows a consistent distribution pattern with low variability. However, certain areas in the mid-country and southern regions exhibit elevated CEC values.

Conclusions

In this study, our assessment focused on the potential of the QRF-DSM approach to address the challenge of populating gridded soil properties in the presence of sparsely observed soil profile data across Sri Lanka. The approach proved to be a valuable tool for predicting significant soil properties such as BD, texture fractions, CEC, SOC, and pH by employing sparsely distributed soil data, a series of environmental covariate layers, and regression tree models within the top soil layer across the entire landscape. This study demonstrated its successful use of the R environment to harmonise data and create 0–30 cm soil layers through a mass-preserving spline function. Its ability to integrate heterogeneous soil data and enhance consistency makes it a valuable approach to soil research. Throughout the process of R modelling, the RFE algorithm emerged as an indispensable tool for identifying and preserving the most significant covariate associated with each soil property. The QRF statistically estimates the different quantiles (percentiles) of a soil property that can capture the entire distribution, providing insights into its variability and uncertainty.

The existing soil property maps can be improved by using recently measured data points, profile information, and incorporating more environmental covariates. This expansion of the QRF method has the potential to provide a more comprehensive understanding of soil characteristics in Sri Lanka, aiding in better land management and decision-making processes.

Acknowledgements

The authors express their gratitude to the AFACI - Soil Project and Asian Soil Partnership for providing funds, technical support, and training on the DSM approach.

References

Angelini, M.E, Luotto, I., Rodriguez Lado, L., Mainka, M., Yigini, Y., & Tong, Y. (2022). *Global soil nutrient and nutrient budgets maps (GSNmap), phase 1*. Technical manual. Global Soil Partnership. FAO, Rome. https://fao-sid.github.io/GSNmap-TM/

Bishop, T.F.A., McBratney, A.B., & Laslett, G.M., (1999). Modelling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma, 91*(1-2), 27-45. https://doi.org/10.1016/S0016-7061(99)00003-8

Biswajit, L., & Divya, R.K. (2022). An introduction to digital soil mapping. In D. R. Naresh, (Ed.), *Advances in agriculture sciences*. (Vol - 25, pp.1-17). AkiNik Publications. https://doi.org/10.22271/ed.book.773

Bueno, C.G., Azorín, J., Gómez-García, D., Alados, C.L., & Badía, D. (2013). Occurrence and intensity of wild boar disturbances, effects on the physical and chemical soil properties of alpine grasslands. Plant and Soil, 373(1–2), 243–256. https://doi.org/10.1007/s11104-013-1784-z

Calzolari, C., Ungaro, F., Filippi, N., Guermandi, M., Malucelli, F., Marchi, N., Staffilani, F., & Tarocco, P. (2016). A methodological framework to assess the multiple contributions of soils to ecosystem services delivery at regional scale. *Geoderma, 261*(1), 190–203. https://doi.org/10.1016/j.geoderma.2015.07.013

Chandrajith, R. (2020). Geology and geomorphology. In R.B. Mapa (ed.), *The soils of Sri Lanka* (pp 23-34). World Soils Book Series. Springer, Cham. https://doi.org/10.1007/978-3-030-44144-9_3

Dassanayake, A. R., de Silva, G. G. R., Mapa, R. B. & Kumaragamage, D. (2006). *Benchmark soils of the Dry Zone of Sri Lanka*. Soil Science Society of Sri Lanka.

Dassanayake, A. R., Somasiri, L. L. W., & Mapa, R. B. (2001). *Benchmark soils of the Intermediate Zone*, Soil Science Society of Sri Lanka.

FAO. (2022). *Country guidelines and technical specifications for global soil nutrient and nutrient budget maps – GSNmap: Phase 1*. Rome. https://doi.org/10.4060/cc1717en

Gomes, L.C., Faria, R.M., de Souza, E., Veloso, G.V., Schaefer, C.E.G., & Filho, E.I.F. (2019). Modelling and mapping soil organic carbon stocks in Brazil. *Geoderma, 340*, 337–350. https://doi.org/10.1016/j.geoderma.2019.01.007

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment, 202*, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Grimm, R., & Behrens, T., (2010). Uncertainty analysis of sample locations within digital soil mapping approaches. *Geoderma, 155*(3), 154-163. https://doi.org/10.1016/j.geoderma.2009.05.006

Indraratne, S.P. (2020). Soil mineralogy. In R.B. Mapa (ed.), *The soils of Sri Lanka* (pp 23-34). World Soils Book Series. Springer, Cham. https://doi.org/10.1007/978-3-030-44144-9_4

Kadupitiya, H. K., De Silva, S. H. S. A., Dilshani, D. G. S. & Weerasinghe, P. (2018). Distribution of soil organic carbon in Sri Lanka: A GIS based modelling approach. *Tropical Agriculturist, 166* (3), 91-106. http://doa.nsf.ac.lk/handle/1/2453

Kasraei, B., Heung, B., Saurette, D.D., Schmidt, M.G., Bulmer, C.E., & Bethel, W. (2021). Quantile regression as a generic approach for estimating uncertainty of digital soil maps produced from machine-learning. *Environmental Modelling & Software, 144*, 105139. https://doi.org/10.1016/j.envsoft.2021.105139

Kienast-Brown, S., Libohova, Z., & Boettinger, J. (2017). Digital soil mapping. In C. Ditzler, K. Scheffe, and H.C. Monger (Eds.), *Soil survey manual, Chapter 5* (pp. 295-354). USDA Handbook 18. Government Printing Office, Washington, D.C. https://

www.nrcs.usda.gov/sites/default/files/2022-09/SSM-ch5.pdf

Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica, 46*(1), 33-50. https://doi.org/10.2307/1913643

Malone, B., McBratney, A., Minasny, B., & Laslett, G. (2009). Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma, 154*(1–2), 138–152. https://doi.org/10.1016/j.geoderma.2009.10.007

McBratney, A., Mendonça Santos, M., & Minasny, B. (2003). On digital soil mapping. *Geoderma, 117*(1–2), 3–52. https://doi.org/10.1016/s0016-7061(03)00223-4

Meinshausen, N. (2006). Quantile Regression Forests. *Journal of Machine Learning Research, 7*(35), 983-999. https://jmlr.org/papers/v7/meinshausen06a.html

Minasny, B., & McBratney, A. (2016). Digital soil mapping: A brief history and some lessons. *Geoderma, 264*, 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017

Moormann, F.R. & Panabokke, C.R. (1961). *Soils of Ceylon: A new approach to the identification and classification of the most important soil groups of Ceylon*. Government press, Ceylon. https://edepot.wur.nl/482354

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M. E., & Papritz, A. (2018). Evaluation of digital soil mapping approaches with large sets of environmental covariates. *SOIL, 4*(1), 1–22. https://doi.org/10.5194/soil-4-1-2018

Poggio, L., de Sousa, L. M., Batjes, N.H., Heuvelink, G.B.M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL, 7*(1), 217–240. https://doi.org/10.5194/soil-7-217-2021

Ramezan, C.A. (2022). Transferability of Recursive Feature Elimination (RFE)-derived feature sets for support vector machine land cover classification. *Remote Sensing, 14*(24), 6218. https://doi.org/10.3390/rs14246218

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. In L. Liu & M.T. Özsu (Eds.) *Encyclopedia of database systems* (pp 532-538). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_565

Robertson, G.P., Sollins, P., Ellis, B.G., & Lajtha, K. (1999). Exchangeable ions, pH, and cation exchange capacity. In G. P. Robertson, D. C. Colemen, C. S. Bledsoe & P. Sollins (Eds.) *Standard soil methods for long-term ecological research* (pp 106-114). Oxford University Press New York. https://andrewsforest.oregonstate.edu/publications/2709

Sakin, E. (2012). Organic carbon organic matter and bulk density relationships in arid-semi arid soils in Southeast Anatolia region. *African Journal of Biotechnology, 11*(6), 1373-1377. https://doi.org/10.5897/ajb11.2297

Senarath, A., Dassanayake, A.R., & Mapa, R.B. (1998). *Benchmark Soils of the Wet Zone*, Soil Science Society of Sri Lanka.

Swan, B. (1983). *An introduction to the coastal geomorphology of Sri Lanka*. National Museums of Sri Lanka.

Szatmári, G., & Pásztor, L. (2019). Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. *Geoderma, 337*, 1329–1340. https://doi.org/10.1016/j.geoderma.2018.09.008

Wadoux, A.M.C., Walvoort, D.J., & Brus, D.J. (2022). An integrated approach for the evaluation of quantitative soil maps through Taylor and solar diagrams. *Geoderma, 405*, 115332. https://doi.org/10.1016/j.geoderma.2021.115332

Walpola, B.C., & Mendis, A.P.I. (2020). Impacts of land use changes on selected soil physical and chemical characteristics under pineapple cropping systems in Matara District, the Low Country Wet Zone of Sri Lanka. *Sri Lankan Journal of Agriculture and Ecosystems, 2*(2), 103–121. https://doi.org/10.4038/sljae.v2i2.41

Zhang, X., Xue, J., Xiao, Y., Shi, Z., & Chen, S. (2023). Towards optimal variable selection methods for soil property prediction using a regional soil Vis-NIR spectral library. *Remote Sensing, 15*(2), 465. https://doi.org/10.3390/rs15020465