GEOGRAPHICAL REPORTS
OF TOKYO METROPOLITAN UNIVERSITY 57 (2022) 1–11

COMPARISON BETWEEN RANDOM FOREST AND MULTIPLE LINEAR REGRESSION TO CREATE DIGITAL MAPS OF SOIL CHEMICAL PROPERTIES IN THE THUNG KULA RONGHAI REGION, THAILAND

Sasirin SRISOMKIEW, Masayuki KAWAHIGASHI and Pitayakon LIMTONG*

Abstract Using machine learning (ML) algorithms to digital soil mapping (DSM) allows the elucidation of relationships between soil properties and environmental variables enabling the precise prediction of soil nutrient levels. The accuracy of the predicted values using the random forest (RF) algorithm, which is the most popular ML algorithm for digital soil mapping, and multiple linear regression (MLR) were compared to create digital maps of soil chemical properties in the Thung Kula Ronghai (TKR) region, Thailand. The spectral indices including moisture stress index (MSI), normalized difference water index (NDWI), saturation index (SI), brightness index (BI), and coloration index (CI) obtained from remote sensing (RS) data were found to be more effective for predicting the various soil properties than the topographic indices derived from the DEM in the plain area. The MLR and RF models successfully predicted soil chemical properties with good predictive accuracy. The results indicated that the RF model has a slightly higher accuracy than the MLR model. However, the MLR model is superior in interpreting the relationship with the model equations.

Keywords: environmental variables, predictor variable, spatial distribution, spectral indices, topographic indices

1. Introduction

Understanding of soil nutrients and their spatial distribution is a key element to sustainable land management. However, conducting laboratory tests in a large number of samples is time consuming and expensive. In addition, the conventional soil survey and mapping techniques are often coarse spatial resolution lacking details, because soil data are mapped as polygons reflecting soil properties according to topographical maps (Lagacherie *et al*. 2020). Thus, the use of new methods that quickly allow us to obtain soil properties at a low cost and less time consumption. Soil science research has applied machine learning (ML) algorithms and digital soil mapping (DSM) techniques to predict and map soil nutrients for various purposes and landscapes. The DSM techniques were designed to overcome the limitations of conventional soil mapping methods in creating seamless soil information by predicting soil properties with high probability (Minasny and McBratney 2016). DSM,

* Land Development Department, Ministry of Agriculture and Cooperatives, Bangkok, Thailand.

which secondary (non-soil) data sources into the mapping process, has been recognized as a potential methodology to create updatable, accurate, and high-resolution soil maps (Emadi *et al*. 2020). Apart from its time and cost reductions, the DSM technique effectively predicts the spatial variability of soil properties by developing models that take soil environmental variables into account. The spatial distribution of data cannot be achieved with the conventional mapping approaches. The majority of the environmental variables used for DSM are spectral indices primarily obtained from remotely sensed satellite images and topographic indices derived from the digital elevation model (DEM). However, despite the development of DSM techniques using ML in many parts of the world, this method has not been used to produce maps of soil properties in Thailand. Therefore, the main objectives of this study were to compare random forest (RF), the most popular ML algorithm for digital soil mapping, and the MLR algorithm to map the spatial distribution of soil chemical properties in the Thung Kula Ronghai (TKR) region of Thailand. These algorithms were compared based on three factors: (1) the accuracy of the models, (2) the predictor variable selection, and (3) the spatial distribution characteristics of soil properties.

2. Dataset and Methodology

Study area and soil data

The TKR region is located in the center of the large basin-shaped Korat Plateau in northeastern Thailand (Fig. 1), covering a total area of 3,370 km^2. The elevation levels range from 108 m to 148 m above sea level, with less than 2% of the slope. According to the Köppen climate classification, the area is characterized by a tropical savanna climate. The average annual temperature is approximately 26.7 °C, with a small variation between summer and winter. Furthermore, the average annual precipitation is 1,300 mm. According to the United States Department of Agriculture soil taxonomy, the major soil types in the TKR region include Ultisols, Alfisols, and Entisols. A total of 186 soil samples were collected from cultivated areas of the 22-soil series during the dry season (March–May 2016).

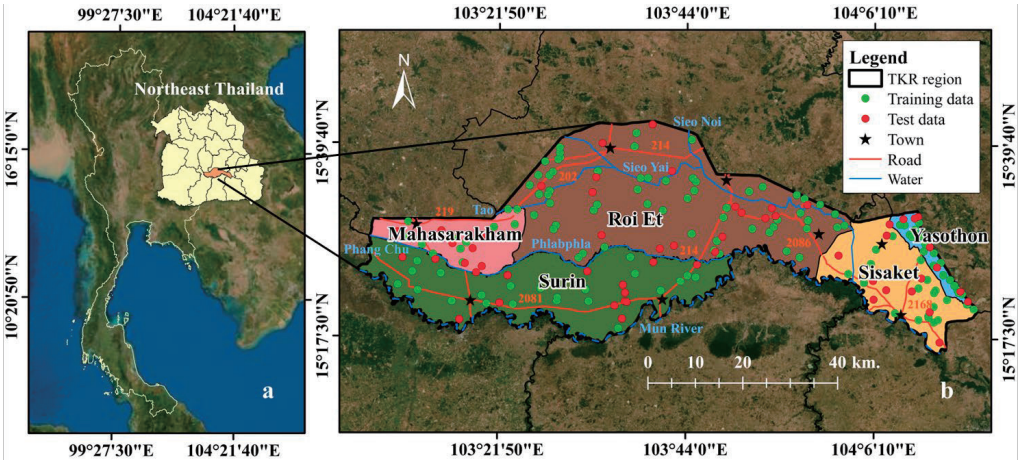

Fig. 1 The TKR region (a) within the northeast region of Thailand and (b) five provinces in the region including soil sampling locations to obtain model equations indicated bygreen dots and to validate the equations shown by red dots. Town, road, and water bodies (dash line is Mun River and solid lines are their tributaries) are also indicated.

The 22-soil series was derived from Paleustults and Paleaquults (Roi Et, Korat, Tha Tum soil series), Natraqualfs (Kula Ronghai soil series), and Quartzipsamments (Nam Pong soil series) (Srisomkiew *et al*. 2021).

Environmental variables

Landsat-8 images with 30 m resolution, which were collected under bare soil conditions in the dry season, enabled us to detect the soil, water, and vegetation indices. The images acquired on April 14th and May 7th, 2016, coincide with the peaks of the dry season with less vegetation, particularly in the paddy fields. The DEM data, with a spatial resolution of 5 m in the TKR region, were generated using aerial photographs from a digital camera onboard the aircraft. Twelve predictor variables derived from Landsat-8 images and the DEM data were used to generate the predictive models. The resolution of the DEM was adjusted to 30 m, similar to the Landsat-8 image, using GRASS GIS software version 7.6.1 (GRASS, 2019). The spectral and topographic indices were calculated using the equations listed in Table 1.

Table 1 Equations of spectral and topographic indices

Index category	Indices	Equations
Spectral indices	Brightness index (BI)	$((R^2 + G^2 + B^2) / 3)^{0.5}$
	Saturation index (SI)	$(R - B) / (R + B)$
	Hue Index (HI)	$(2R - G - B) / (G - B)$
	Coloration index (CI)	$(R - G) / (R + G)$
	Normalized difference vegetation index (NDVI)	$(NIR - R) / (NIR + R)$
	Normalized difference water index (NDWI)	$(G - SWIR1) / (G + NIR)$
	Moisture stress index (MSI)	$SWIR2 / NIR$
Topographic indices	Elevation (ELV)	GDAL contour function in QGIS
	Aspect (ASP)	GDAL aspect function in QGIS
	Slope (SLP)	GDAL slope function in QGIS
	Topographic wetness index (TWI)	ln (*SCA* / tan (Slope))
	Stream power index (SPI)	*SCA* × tan (Slope)

R = Red, *G* = Green, *B* = Blue, *NIR* = Near infrared, *SWIR1* = Short wave infrared 1, *SWIR2* = Short wave infrared 2, *SCA* = specific catchment area.

Multiple linear regression (MLR)

In the MLR model, soil properties were set as dependent variables, whereas spectral and topographic indices were set as independent variables. The soil datasets obtained from laboratory analyses were randomly split into two sets: a dataset consisting of 70% of the total data (130 out of 186) was used as the training set to generate predictive models for mapping, and the other 30% (56 out of 186) was used to validate the model equations as shown in Fig. 1. The MLR is expressed as:

$$Y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n, \qquad (1)$$

where Y is the dependent variable (soil property); a is the intercept; $b_1, b_2, \dots b_n$ are the coefficients of partial regressions; $x_1, x_2, \dots x_n$ are the predictors or independent variables (spectral and topographic indices), and n is the number of predictors.

Random forest (RF)

In this study, the feature importance (FI) and recursive feature elimination with cross-validation (RFECV) methods were employed for variable selection. In the first step, the variable importance is evaluated to produce a feature of the important score employed in the RF algorithm. The variables were then ranked according to the scores received and the variables with a high score were identified as important variables in the predictive model. In the second step, the RFECV method works by removing the variables with low importance and maintaining the optimal number of variables. The dataset was divided into two groups for model training (n = 130, 70%) and model testing (n = 56, 30%), which are the same dataset used in the MLR model. The k-fold cross validation, where k was set to 10 of the RF model, was then carried out using the training dataset. In the final step, the performance of the final predictive RF model was evaluated using the test dataset (n = 56). The predicted values from the MLR and RF models were used to generate the digital soil maps of the five soil properties using Python 3.8.3 and QGIS software version 3.12.1 (QGIS Development Team, 2020).

Model accuracy assessment

The performance and accuracy of the models were evaluated by the coefficient of determination (R^2), Root Mean Square Error (RMSE), and Normalized Root Mean Square Error (NRMSE).

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}, \qquad (2)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=i}^{N}(y_i - \hat{y})^2}, \qquad (3)$$

$$NRMSE = \frac{RMSE}{\bar{y}}, \qquad (4)$$

where y_i is the observed value of the soil training data, $\hat{y}$ is the predicted value, and $\bar{y}$ is the mean value of y_i. The predictive model's performance and accuracy using the test dataset were estimated in terms of the R^2 value, RMSE, and NRMSE using Python software.

3. Results and Discussion

Model performance of the MLR and RF models

Table 2 presents the MLR model equations for different soil properties. For the NRMSE, in this study, values between 0 and 0.2 indicate a low NRMSE, the values between 0.2 and 0.5 are considered as moderate NRMSE, and values above 0.5 are considered as high NRMSE. The predictive models from MLR for soil pH, OM, and available P showed high R^2 values ranging from 0.71 to 0.79, and low RMSEs and NRMSEs. However, the predictive model equations for EC and K showed low prediction capabilities with low R^2 values of 0.44 and 0.57, respectively.

Table 2 MLR model equations of each soil property (n = 130)

Soil property	Multiple linear regression model equation	R^2	RMSE	NRMSE
pH	$4.44 + 0.08\,SI + 1.21\,NDWI - 1.48\,MSI$	0.76	0.23	0.04
EC	$46.08 + 831.94\,SI - 7.68\,CI - 69.74\,NDWI - 40.63\,MSI + 0.90\,ELV$	0.44	2.04	0.05
OM	$120.92 + 1.98\,BI - 0.39\,SI - 2.81\,NDWI - 3.22\,MSI - 0.87\,ELV$	0.71	0.26	0.03
P	$-45.59 + 71.47\,BI - 28.14\,SI - 4.47\,HI + 4.25\,NDWI + 2.79\,MSI$	0.79	3.07	0.46
K	$-29.13 - 126.87\,SI + 44.49\,HI + 4.93\,NDWI + 33.12\,MSI$	0.57	10.72	0.60

Abbreviations of variables in the model equations are listed in Table 1.

Table 3 shows that the RF model has high predictive capabilities for all soil indicators and higher R^2 values than the MLR model. The RF model had higher performance to predict soil pH with the highest R^2 value (0.84) and lower errors. In particular, the EC and K models showed relatively low predictive capabilities with high error values as compared to the other soil indicators in the RF model.

Table 3 Statistical evaluation of the RF model (n = 130)

Soil property	Predictive models		
	R^2	RMSE	NRMSE
pH	0.84	0.11	0.04
EC	0.74	42.2	0.07
OM	0.83	2.22	0.07
P	0.78	2.88	0.07
K	0.76	5.16	0.09

The performances of the MLR and RF models in the prediction of soil properties were validated according to the relationship between the predicted and testing data sets (as shown in Figs. 2 and 3). The scatter plots for pH showed a significant linear correlation between the measured and predicted values distributed on the 1:1 line with a high R^2 and low error in both the MLR and RF models. The MLR and RF models predicted the OM and available P with good accuracy. The available range for OM prediction (0–12 g kg^{-1}) was the same for the RF and MLR model equations. The accuracy of EC and K prediction was lower than the other soil properties in both the RF and MLR models, despite the better predictive accuracy of the RF model than the MLR model.

However, the predicted EC and K data using the MLR and RF models are prone to under- and overestimations, as most of the data are distributed above and below the 1:1 line in the relationship between the predicted and tested values. These results were in agreement with the findings of Silva *et al.* (2017) where the models generated by RF with a large amount of data showed higher performance in predicting soil properties than the MLR model. John *et al.* (2020) also summarized that the RF models provided better results than those produced by simple techniques such as MLR models for estimating various soil properties. The RF model can analyze the linear and nonlinear relationships between soil data and environmental variables, which results in a strong fitting ability and high estimation accuracy (Xie *et al.* 2021).

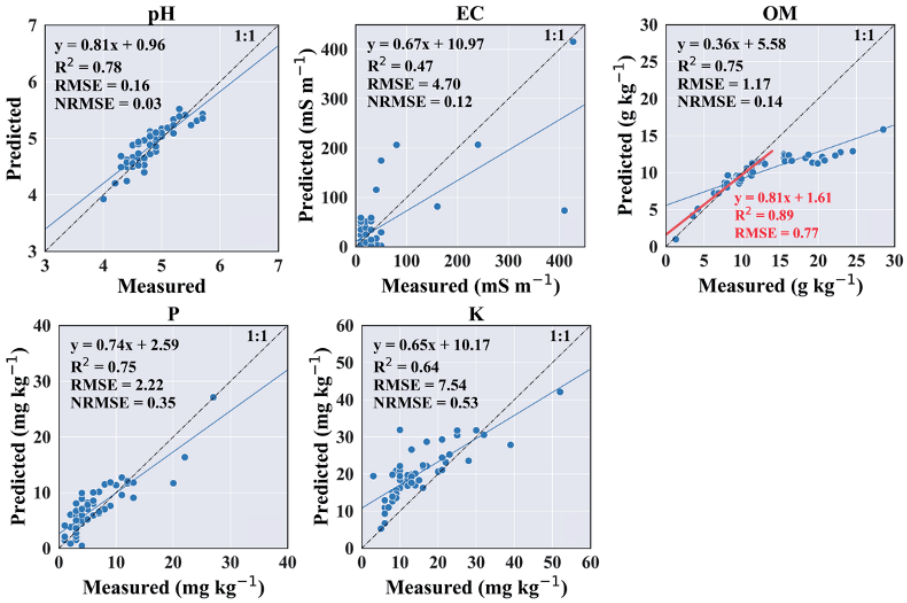

Fig. 2 The relationship between measured and predicted soil properties using the MLR model (n = 56). The blue line is a regression line obtained by the MLR model, and the red line is the regression line of the available range (0–12 g kg^{-1}) for OM prediction with high accuracy (Srisomkiew *et al*. 2021).

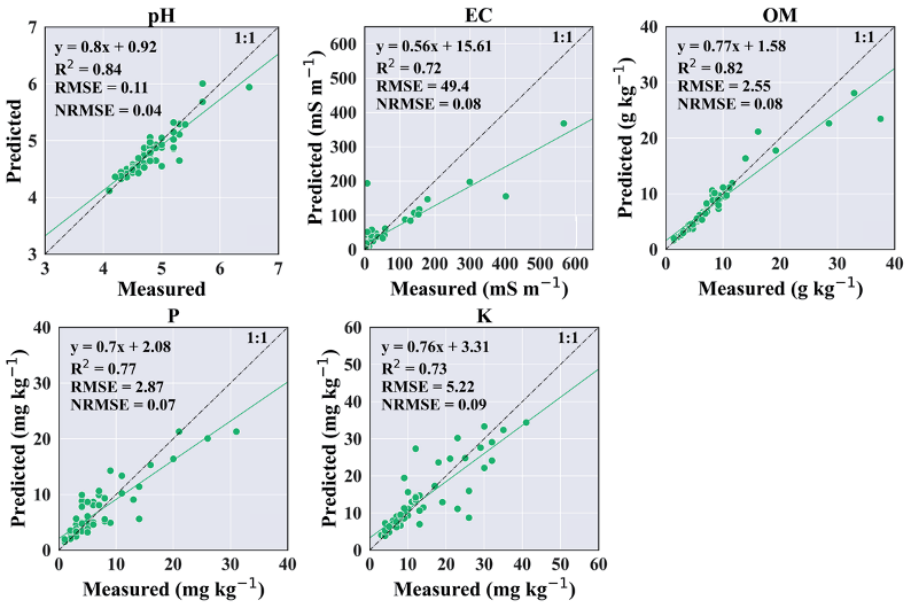

Fig. 3 The relationship between measured and predicted soil properties using the RF model (n = 56). The green line is a regression line obtained by the RF model.

Important predictor variables of the MLR and RF models

The important predictor variables of the spectral and topographic indices derived from the MLR models to predict soil properties are shown in Table 2. Three predictor variables, MSI, NDWI, and SI, explained the soil pH, EC, and OM together with other variables such as BI, CI, and ELV. The results show that NDWI and MSI significantly contribute to the model equations because of their statistically significant coefficients, implying that they are important predictor variables for the prediction of pH, EC, and OM. Angelopoulou *et al*. (2019) suggested that both NDWI and MSI improved the predictive accuracy of soil moisture, pH, EC, and OM content. Morgan *et al*. (2018) reported that NDWI and MSI have been widely used to estimate soil pH, EC, and OM, as well as to evaluate surface soil moisture. To predict the macronutrients (P and K), the spectral indices of BI, SI, HI, MSI, and NDWI were the principal variables in this study. Uerchefani *et al.* (2009) and Kumar (2017) also proposed four spectral indices as suitable variables to predict micronutrients to understand soil nutrient conditions.

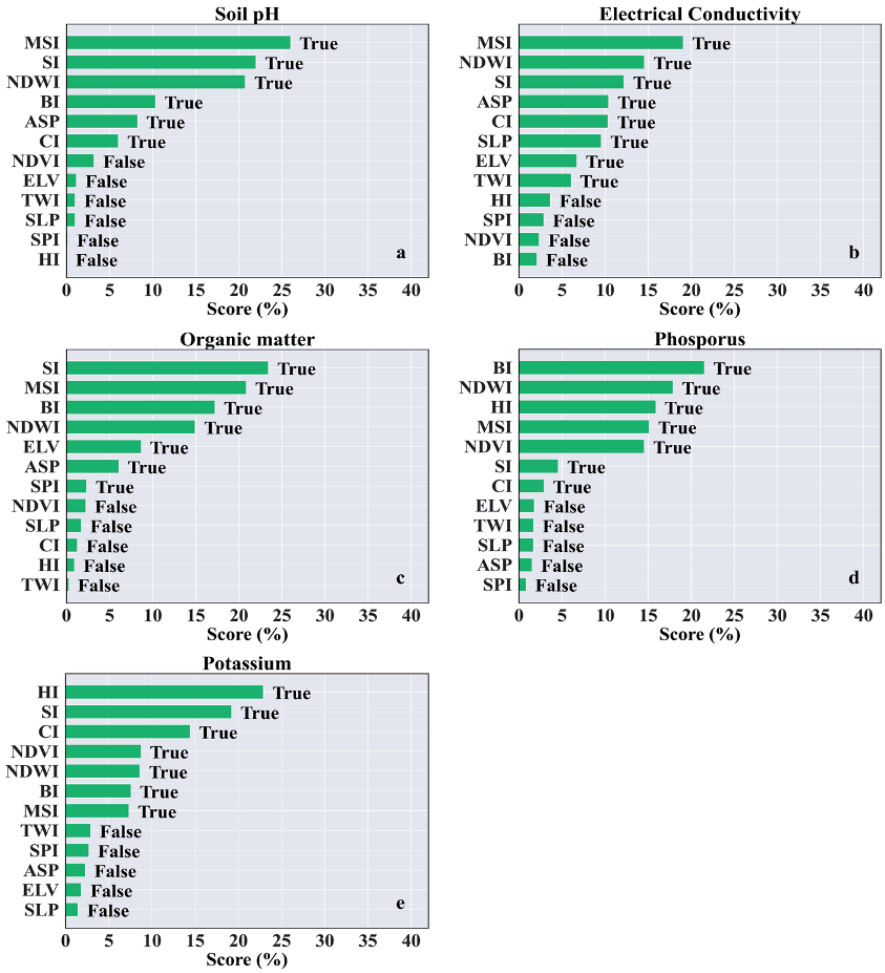

Fig. 4 Importance of predictor variables in the random forest model for: (a) soil pH, (b) electrical conductivity, (c) organic matter, (d) phosphorus, and (e) potassium.

Figure 4 shows the important predictor variables selected through RFECV process (True or False) and FI score (%) and their contributions, in the RF model, to predict soil properties. The results indicate that MSI, SI, NDWI, BI, ASP, and CI are the important variables affecting soil pH. Within the six selected important variables to build the pH model, MSI, SI, and NDWI had high relative importance ranging from 21% to 26% as compared to BI, ASP, and CI which had relatively low importance ranging from 6% to 10%. Based on the FI scores, the relative importance between the selected spectral and topographic variables were not very different in the EC model. The eight predictor variables of the spectral (MSI, NDWI, SI, and CI), and topographic variables (aspect, slope, elevation, and TWI) have a significant influence on the prediction of EC. Among the seven selected predictor variables in the OM model, the spectral variables such as SI, MSI, BI, NDWI showed higher contribution to the model as compared to the topographic variables of ELV, ASP, and SPI. In the P and K models, only spectral variables contributed to the model, while the topographic variables showed lower importance. Although one of the spectral indices of NDVI was just included as an important variable in the P and K model, it was not selected for other soil models. Ryu *et al*. (2020) stated that NDVI has good potential to monitor the variation of the total N, available P, and exchangeable K content in the soil as compared to other soil properties. Findings from John *et al*. (2020) suggest that topographic indices contribute less effectively to predict soil properties owing to a small variation range of topographic parameters in plain lands.

Digital maps of soil properties from MLR and RF models

Figure 5a–e represented the digital maps of soil chemical properties that were produced using predicted data calculated by the MLR model equations. The digital maps using RF models are shown in Figure 5f–j. The grade was indicated by soil nutrient categories expressed as low, medium, and high. The predicted pH values from the MLR model ranged from 3.7 to 7.5, while the value from the RF model ranged from 4.1 to 6.3. The digital map of soil pH indicated that most of the soil in the TKR region was acidic in nature, particularly in the western part of the region, as shown by the red color on the map. The distribution of the EC maps from both the models presented as green pixels can be considered as high salinity areas in the TKR region because of the underestimation of the predictive models. The OM map from the MLR and RF models accurately express the predicted values below 12 g kg^{-1}. The low contents of OM also explained the low soil fertility of the TKR region. The OM maps from the two models showed a similar trend. The digital maps for available P showed a wide range from low to high throughout the TKR region by two models. Besides, the K maps from the MLR and RF models exhibited an identical trend showing deficiency of exchangeable K in the TKR region.

The MLR and RF models have both advantages and disadvantages in predicting soil properties. The advantage of the MLR model is the clear explanation of the relationships between predictor variables and soil properties from model equations. In terms of predictive accuracy, the RF model performed better than the MLR model; however, both models were able to predict soil properties that were statistically reliable and significantly accurate. The MLR model produced consistent and easily interpretable maps as compared to the RF model for most of the soil properties (Jeune *et al*. 2018). Although the RF model has enhanced predictive accuracy, it does not show the relationship between soil properties and predictive variables as the MLR model (Krkač *et al.* 2020; Smith *et al.* 2013).

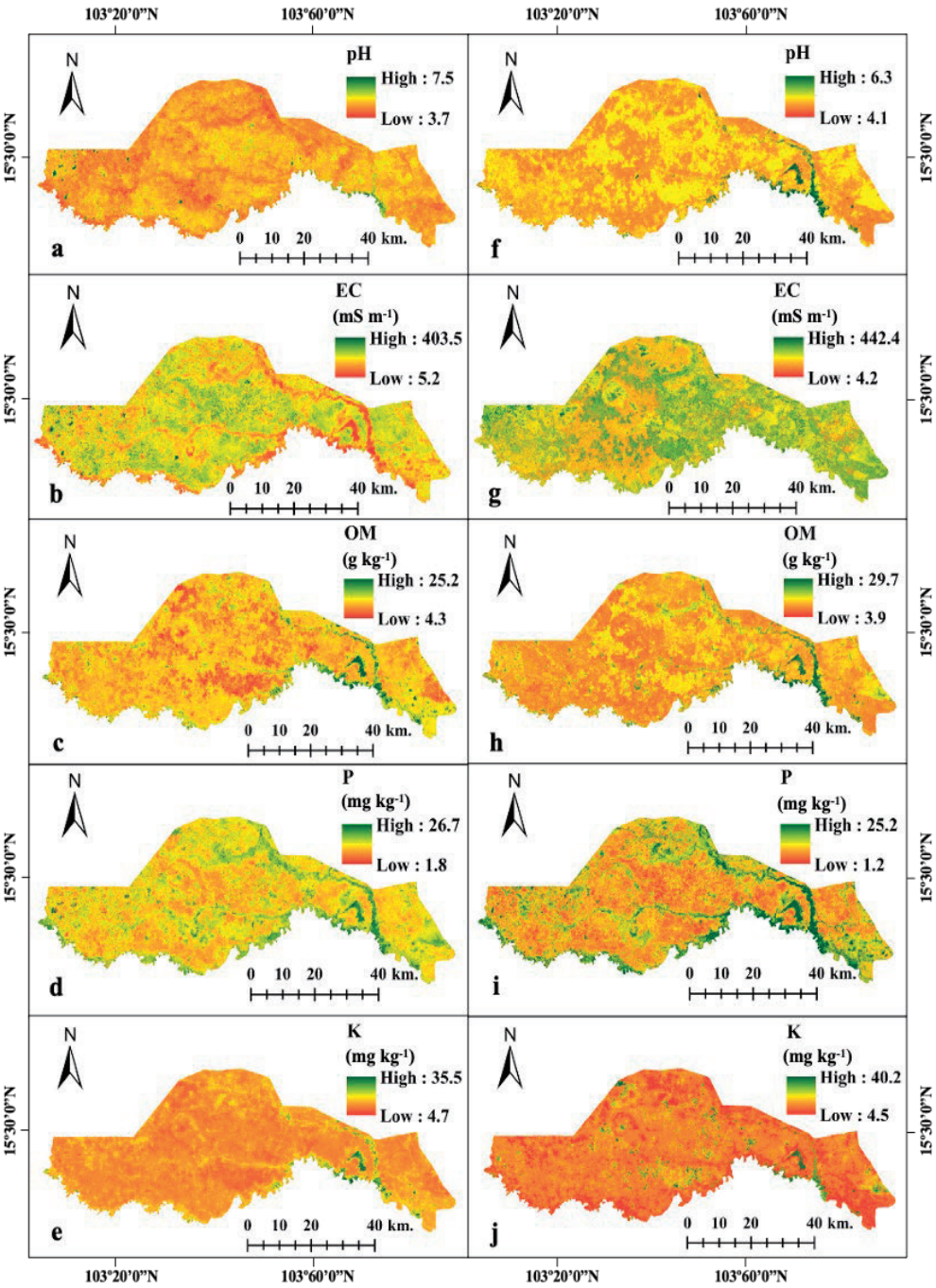

Fig. 5 Digital soil maps of soil properties in the TKR region are shown on the left side (a–e) from the MLR model and the right side (f–j) from the RF model.

4. Conclusions

This study was conducted to estimate soil chemical properties by comparing MLR and RF models in the agricultural land of the TKR region. The results demonstrated that the MLR and RF models effectively create digital maps of soil pH, OM, and available P with statistical reliability and high accuracy. However, these two models exhibited relatively lower EC and K accuracies. The RF model could enhance the accuracy slightly better than the MLR model. However, the spatial distribution of the maps for all soil properties showed a similar trend. In the case of model explanation, the MLR model is superior to the RF model, as it clearly and easily interprets the relationship between the predictor variables and soil properties. In future model implementation, it is recommended to explore the data first to understand the relationship between the dependent and independent variables, followed by the selection of a suitable model. To enhance the model accuracy, it is essential to select appropriate environmental variables to predict soil properties. Additionally, an increase in the number of soil sampling data during the training stage can help to improve the predictive model.

Acknowledgments

This research was partly supported by the research budget "Promotion for globalization researches" by Tokyo Metropolitan University. The authors are sincerely thankful to the Land Development Department (LDD), Thailand for providing soil and DEM data, the Ubon Ratchathani Rice Research Center (URRRC), Thailand for providing jasmine rice information as well as the Thai Meteorological Department (TMD) for providing rainfall and temperature data. We also express gratitude to the USGS Earth Resources Observation and Science Center (EROS) for the use of Landsat-8 (OLI) data.

References

Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., and Bochtis, D. 2019. Remote sensing techniques for soil organic carbon estimation: A review. *Remote Sensing* **11**: 1–18.

Emadi, M., Taghizadeh-mehrjardi, R., Cherati, A., and Danesh, M. 2020. Predicting and mapping of soil organic carbon using machine learning algorithms in northern Iran. *Remote Sensing* **12**: 1–29.

GRASS Development Team. 2019. Geographic resources analysis support system (GRASS) software. https://grass.osgeo.org/about/license/ (December 20th, 2019).

Jeune, W., Francelino, M.R., De Souza, E., Fernandes Filho, E.I., Rocha, G.C., 2018. Multinomial logistic regression and random forest classifiers in digital mapping of soil classes in western Haiti. *Revista Brasileira de Ciência do Solo* **42**: 1–20.

John, K., Isong, I. A., Kebonye, N. M., Ayito, E. O., Agyeman, P. C., and Afu, S. M. 2020. Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. *Land* **9**: 1–20.

Krkač, M., Bernat Gazibara, S., Arbanas, Ž., Sečanj, M., and Mihalić Arbanas, S., 2020. A comparative study of random forests and multiple linear regression in the prediction of landslide velocity. *Landslides* **17**: 2515–2531.

Kumar, S., 2017. Geospatial tools and techniques in land resource inventory. In *Sustainable Management of Land Resource.* ed. G.P.O. Reddy, N.G. Patil, A. Chaturvedi, 172–200. New York: Apple Academic Press.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., and Nkuba-Kasanda, L. 2020. Analysing the impact of soil spatial sampling on the performances of Digital Soil Mapping models and their evaluation: A numerical experiment on quantile random forest using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery. *Geoderma* **375**: 1–12.

Minasny, B., and McBratney, A. B. 2016. Digital soil mapping: A brief history and some lessons. *Geoderma* **264**: 301–311.

Morgan, R. S., El-Hady, M. A., and Rahim, I. S. 2018. Soil salinity mapping utilizing sentinel-2 and neural networks. *Indian Journal of Agricultural Research* **52**: 1–6.

QGIS Development Team, 2020. QGIS geographic information system developers manual, Open Source Geospatial Foundation Project. http://www.qgis.org/wiki/Developers_Manual.

Ryu, J. H., Jeong, H., & Cho, J. 2020. Performances of vegetation indices on paddy rice at elevated air temperature, heat stress, and herbicide damage. *Remote Sensing* **12**: 1–25.

Silva, S. H. G., Teixeira, A. F. dos S., Menezes, M. D. de, Guilherme, L. R. G., Moreira, F. M. de S., and Curi, N. 2017. Multiple linear regression and random forest to predict and map soil properties using data from portable X-ray fluorescence spectrometer (PXRF). *Ciência e Agrotecnologia* **41**: 648–664.

Smith, P.F., Ganesh, S., and Liu, P., 2013. A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of Neuroscience Methods* **220**: 85–91.

Srisomkiew, S., Kawahigashi, M., and Limtong, P. 2021. Digital mapping of soil chemical properties with limited data in the Thung Kula Ronghai region, Thailand. *Geoderma* **389**: 1–12.

Uerchefani, D. O., Haou, H. D., Bdeljaoued, S. A., Elaitre, E. D., and Allot, Y. C. 2009. Radiometric indices for monitoring soil surfaces in South Tunisia. *Arid Land Studies* **19**: 73–76.

Xie, X., Wu, T., Zhu, M., Jiang, G., Xu, Y., Wang, X., and Pu, L. 2021. Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land. *Ecological Indicators* **120**: 1–9.