REMOTE SENSING BASED DIGITAL MAPPING OF SOIL PROPERTIES IN A HIMALAYAN WATERSHED EMPLOYING MACHINE LEARNING TECHNIQUE

Nyenshu Seb Rengma (1), Justin George K (1), Suresh Kumar (1)

1 Agriculture and Soils Department, Indian Institute of Remote Sensing, Indian Space Research
Organisation, Govt. Of India, 4 Kalidas Road, Dehradun District, Uttarakhand, 248001, India
Email: nyenshu@gmail.com; justinagri@gmail.com; suresh_kumar@iirs.gov.in

KEYWORDS: Soil texture, Organic carbon, Support vector regression, Himalaya

ABSTRACT

Spatially distributed soil information is essential for informed decision-making aimed at environmental management and as well as addressing many research issues. To obtain continuous soil data from limited point measurements, digital soil mapping (DSM) techniques using various remote sensing data products are now recognized as effective tools for a wide range of spatial scales and diverse landscapes. The study was conducted in an Indian Himalayan watershed located near Chamba in Tehri Garhwal district of Uttarakhand, for mapping soil organic carbon and textural fractions (sand, silt, and clay) employing support vector regression, a popular machine learning technique. The watershed covers an area of 43 square kilometres and the elevation varies between 667 and 2459 metres above the mean sea level. Soil database comprising laboratory analysis results of 212 surface (0-15 cm) soil samples corresponding to georeferenced sampling locations within the area was used for mapping the spatial distribution using high-resolution remote sensing satellite data (Sentinel-2) and terrain data (CartoDEM). Different spectral indices reflecting the variations in vegetation and parent material were generated from archived long-term Sentinel-2 datasets using Google Earth Engine cloud computing platform and terrain indices were generated using high-resolution CartoDEM. These different environmental variables were used for fitting prediction models using support vector regression (SVR) in the R computing platform. A recursive feature elimination approach with 10 fold cross-validation was used for the selection of significant environmental variables concerning different targeted soil properties. The performance of the models was statistically analyzed and optimal accuracy was obtained by calculating the R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) values. The models could predict sand and silt textural fractions with R-squared values of 0.61 and 0.50 respectively. Whereas clay distribution could be predicted with a higher R-squared value of 0.71, and the MAE and RMSE values of 0.52 and 3.12 respectively, while the soil organic carbon predictions had the least R-squared value of 0.04. The validated models were further used for mapping different soil properties in the watershed. These approaches making use of the increasingly available free remote sensing and terrain data employing machine learning techniques will aid in improving soil information at different scales especially in mountainous terrains with limited resources and much-reduced costs. Ultimately, such rapid and reliable soil information systems with known uncertainties will help us in adopting various land management decisions for improving agricultural productivity, enhanced resource utilization as well as environmental sustainability.

INTRODUCTION

Spatially distributed soil information is essential for informed decision-making aimed at environmental management and as well as addressing many research issues. To address the affair of the resources ranging from local to global scales, environmental scientists and policymakers are seeking soil information that is more specific and detailed with explicit spatial data. These users' needs demand recent and improved spatial soil information, particularly in a digital format that is readily incorporated into geographic information systems (GIS) and can be analyzed with other spatial data (Lagacherie and McBratney, 2007). The approach of the information acquirement, modeling, and mapping of the soil is significant to their response, but conventional soil mapping cannot efficiently provide the amount of data that is required at an optimal cost. The vital reason is well known; when spatial soil information is needed, conventional soil sampling for a wide area and laboratory analyses of the huge soil samples is time-consuming and expensive (Viscarra Rossel and McBratney, 1998). Conventional mapping can produce accurate maps but is labor-intensive and often impractical in large, inaccessible areas (Bui and Moran, 2003). To obtain continuous soil data from limited point measurements, digital soil mapping (DSM) techniques using various remote sensing data products are now recognized as effective tools for a wide range of spatial scales and diverse landscapes.

Digital soil mapping (DSM), also known in the literature as predictive soil mapping has evolved from traditional soil survey with advances in computing and geographic data handling, as well as increased availability of environmental covariate data from digital elevation models, and remotely sensed imageries. It is the computer-assisted production of digital soil maps of soil classes or soil properties based on quantitative relationships between spatially explicit environmental data or covariates and measurements made in the field and laboratory (McBratney et al., 2003; Scull et al., 2003; Lagacherie and McBratney, 2007). The benefit of digital soil mapping is in its quantitative character of soil; external factors dependence that makes soil cartography more scientifically based. Soil is a continuum, where the soil properties at a given location depend on their geographic position and also on the soil properties at neighboring locations. This approach was followed and summarized by McBratney et al. (2003), who identified 7 factors (scorpan) for soil spatial prediction. The proposed scorpan model, where soil (as either soil classes, Sc or soil attributes, Sa) at a point in space and time is an empirical quantitative function of seven environmental covariates. The scorpan in its expanded form is soil, climate, organisms (vegetation or fauna or human activity), relief, parent material, age, and spatial position respectively, which are the soil-forming factors characterizing environmental covariates (Minasny et al. 2013). It is similar to the clorpt (Jenny, 1941) model but intended for quantitative descriptions of relationships between soil and other spatially referenced factors and is used as soil spatial prediction functions rather than explanation (McBratney et al., 2003). The scorpan equation also explicitly incorporates space (x, y coordinates) and time (~t), which indicates that scorpan is a geographic model, where the soil and factors are spatial layers that can be represented in a geographic information system.

The development of machine learning as a branch of artificial intelligence (AI) is very fast. Its usage has spread to various fields. Machine learning (ML) is currently applied to mapping soil properties or classes much in the same way as other unrelated fields of science. Mapping of soil, however, has unique aspects that require adaptations of the ML algorithms. Machine learning in soil science covers a set of data-mining techniques that can recognize patterns in datasets and learn from these to predict quantitative soil variables. Many algorithms are available and robust prediction results are possible (Hastie et al., 2009; Li and Heap, 2008; Li et al., 2011). Digital soil mapping aims to predict the target soil variable at unobserved locations from observations at neighboring locations, preferably with the help of layers of environmental covariates.

The support vector machine (SVM) with its ability to learn complex data-classes, handle large, and low-dimension datasets has been applied to soil mapping (Li and Zhao, 2009; Brungard et al., 2015; Brevik et al., 2016; Heung et al., 2016; Forkuor et al., 2017).In undulating and highly elevated terrain of the Central Himalayas, soil properties are locally variable (Bahuguna and Chaturvedi, 2018) encouraging the regression approach of digital soil mapping. Support vector regression (SVR) algorithm is one of the most promising performances in regression-based digital soil mapping (Ballabio, 2009), which was used to spatially predict and map the soil properties of the unsampled area in this study.

MATERIALS AND METHODS

Study area

The study was conducted in an Indian Himalayan watershed located near Chamba in Tehri Garhwal district of Uttarakhand. The study area covers 4270 ha (43 sq. km) and lies between 78˚27'01'' E - 78˚21'53''E longitudes and 30˚20'15'' N - 30˚25'22'' N latitudes. The region is highly undulating and exhibits the typical mountainous topography of North-Western Himalayas with an elevation of 667 to 2459 meters above mean sea level. Agriculture land is the dominant land-use system in the watershed followed by forest and scrubland. The watershed considered for the study is characterized by deep gorges, slopes, narrow valleys, and rocky escarpments. The entire watershed consists of high hills and ridges which are deeply incised by the streams.

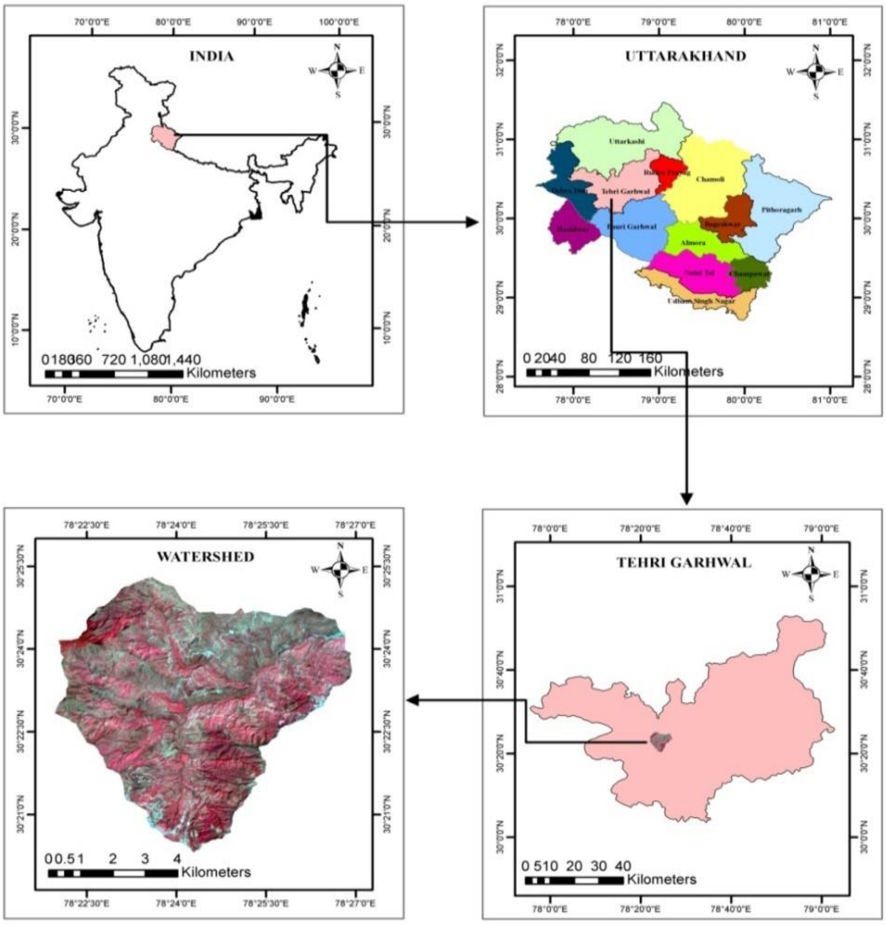


Figure 1: Geographic location of the study area (c CartoDEM Image Copyright 2018 NRSC/ISRO)

Soil database and environmental covariates

Soil database comprising laboratory analysis results of 212 surface (0-15 cm) soil samples corresponding to georeferenced sampling locations within the area was used for mapping the spatial distribution of the soil properties, soil organic carbon, and textural fractions (sand, silt, and clay).

An extensive set of environmental covariates considering selective factors of soil spatial prediction from the scorpan model, primarily remote sensing data are used for predicting soil properties in the digital soil maps. A total of 27 environmental covariates were generated from the 10 metre high-resolution remote sensing satellite data (Sentinel-2) and terrain data (CartoDEM), which were analyzed to understand their relationship with the soil properties. The various spectral indices of Sentinel-2 data were acquired from Google Earth Engine (GEE), an open cloud computing platform. Whereas the terrain parameters were derived from the 10 metre resolution CartoDEM in ArcGIS. The vegetation indices generated include Green normalized difference vegetation index (GNDVI), Green soil adjusted vegetation index (GSAVI), Modified soil adjusted vegetation index 2 (MSAVI 2), Normalized difference vegetation index (NDVI), Optimized soil adjusted vegetation index (OSAVI), Ratio vegetation index (RVI), Renormalized difference vegetation index (RDVI), Soil adjusted vegetation index (SAVI), and Transformed difference vegetation index (TDVI), and Transformed vegetation index (TVI). And the parent material indices generated were Calcareous mineral, clay mineral, ferrous iron, ferrous mineral, ferrous silicate, gypsic mineral, iron oxide, and other spectral indices namely brightness, coloration, and saturation. The terrain parameters and indices generated were aspect, curvature, hillshade, slope, solar radiation, stream power index (SPI), and terrain wetness index (TWI).

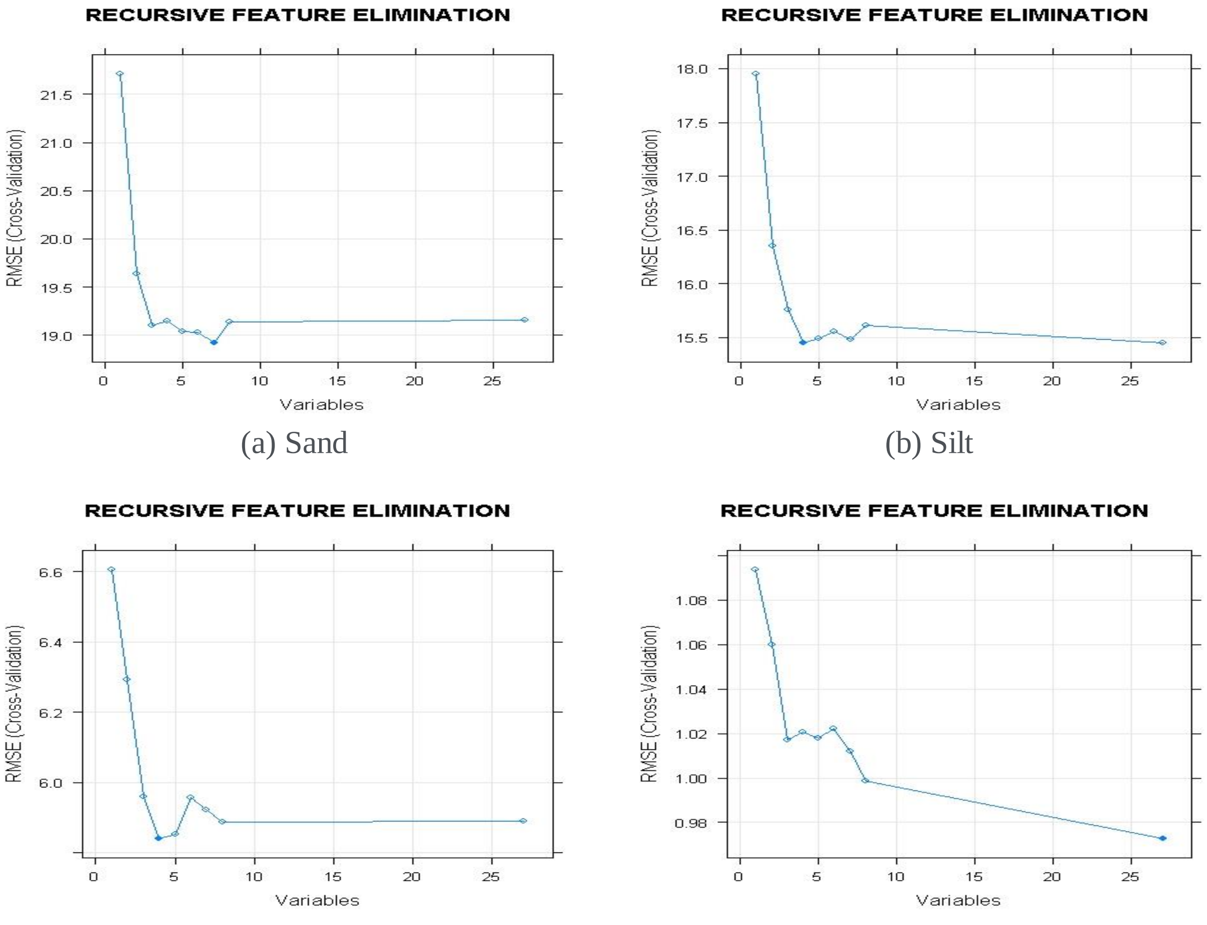
Model calibration and validation

The support vector machine (SVM) algorithm was developed and introduced by Vapnik (1995) as a classifier. The algorithm is used to classify linearly separable classes of objects by finding the hyperplanes that best separate the two classes. It maximizes the margin between the two classes which increases the separability. Those points of the two classes that are closest to each other are called support vectors. Support vector regression has been modified from the SVM and featured with the capability of capturing nonlinear relationships in the feature space and is considered an effective approach to regression analysis. In the case of classification, SVM aims to construct an optimal hyperplane that separates classes creating the widest margin between the data whereas in the case of regression it fits the data and predicts it with minimal empirical risk and complexity of the modeling function.

The SVR models of the soil properties were built in the integrated development environment, RStudio of R using the e1071 package. The SVR models of the soil properties were calibrated with the training dataset (70 percent of the dataset) and validated (evaluation of predictive performance) with the remaining 30 percent of the dataset. While the feature selection was performed by using the Recursive Feature Elimination (RFE) algorithm, implemented on the Caret Package (Kuhn, 2017). RFE is an algorithm that performs a backward selection, which avoids refitting many models at each step of the search (Kuhn and Johnson, 2013). The input data had to be transformed into a higher dimension space using a non-linear function as most data cannot be separated linearly in reality. Kernels, which are mathematical functions were used to transform the data into a higher dimension space in which the data could be more easily separated (Gunn, 1998; Ivanciuc, 2007; Williams, 2011). Perhaps the major task of SVR is to choose one among the many available kernels and accordingly fine-tune several other hyperparameters. These hyperparameters were empirically tuned to achieve good performance of prediction called hyperparameter optimization or model selection. The standard kernels of SVM like linear, polynomial, radial, and sigmoid were used to calibrate the models and were comparatively analyzed.

RESULTS AND DISCUSSION

The effective SVR models of the soil properties with optimal regression accuracy and generalization performance were obtained by considering the significant variables and fine-tuning the models using the tune function of the e1071 package of R. With feature selection being an indispensable step in machine learning, significant variables or features were selected by the recursive feature elimination (RFE) on 10-fold cross-validation. The acquired optimum number of variables based on RFE are seven, four, four, and twenty seven in the sand, silt, clay, and soil organic carbon respectively (Figure). The selected variables were then used to build the SVR models of the respective target or dependent. The fact that all the twenty seven variables were selected for soil organic carbon was considered and cross-validation of the model with the top 5 variables was performed. The models built with the twenty-seven variables and the top five variables for the target or dependent variable SOC showed similar errors and also the relationship between the observed and predicted soil properties illustrated similar R-squared values. So, the top five variables were considered for building the regression model of soil organic carbon.



(a) Sand

(b) Silt

(c) Clay

(d) Soil organic carbon

Figure 2: Selection of an optimum number of variables based on the least RMSE for 10-fold cross-validation for the SVR model of (a) sand, (b) silt, (c) clay, and (d) soil organic carbon

Table 1: The various features selected for training the SVR model

Soil property	Variables selected
Sand	Calcareous, clay mineral, coloration, ferrous iron, ferrous mineral, ferrous silicate, gypsic mineral
Silt	Calcareous, coloration, ferrous iron, ferrous mineral
Clay	Calcareous, ferrous mineral, iron oxide, RDVI, saturation
SOC	Calcareous, ferrous mineral, iron oxide, RDVI, saturation

While the tune function selects the best model by iteratively running and selecting the most significant hyper-parameters. The hyper-parameters considered in SVR are the kernel, cost, epsilon, and gamma. The four basic kernels (linear, polynomial, radial, and sigmoid) were comparatively analyzed for all the soil properties. The best SVR models of all the soil properties and their respective hyper-parameters selected are given in Table 2. The soil properties sand, clay, and SOC obtained the highest R-squared with linear kernels. While on the contrary, silt illustrated the highest R-squared value between the observed and predicted values with a radial kernel.

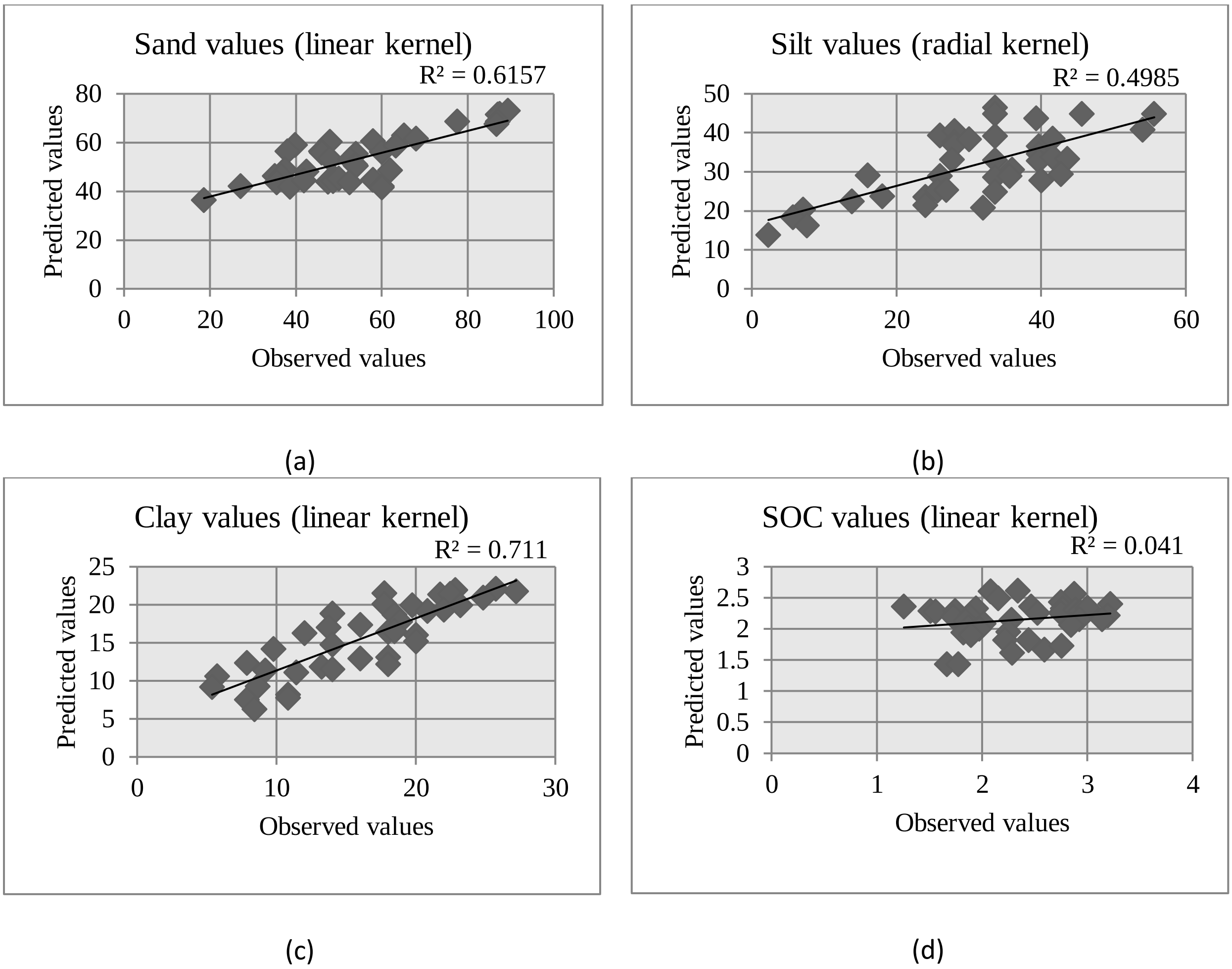
Table 2: Soil properties and their corresponding hyper-parameters selected in the SVR model

Soil properties	Kernel	Cost	Epsilon	Gamma
Sand	Linear	8	1	0.14
Silt	Radial	4	0.6	0.25
Clay	Linear	4	0.1	0.25
SOC	Linear	4	0.8	0.2

The R squared value of the validation dataset from the support vector regression model is calculated for each soil property. In the regression model, R-squared (R^2) represents the correlation coefficient between the independent variables and the target or dependent variable. The mean absolute error (MAE), and the root mean squared error (RMSE) were considered in assessing the accuracy performance of the regression predictive model. They were calculated using the Metrics package of R.

Table 3: Cross-validation of the predictive soil properties model with different kernels

Soil properties	Kernels	R-squared	MAE	RMSE
Sand	Linear	0.61	0.95	11.20
	Polynomial	0.38	4.02	11.46
	Radial	0.51	1.53	12.86
	Sigmoid	0.007	4.39	154.68
Silt	Linear	0.46	1.90	9.51
	Polynomial	0.35	1.08	10.51
	Radial	0.50	1.01	8.72
	Sigmoid	0.03	124.04	232.02
Clay	Linear	0.71	0.52	3.12
	Polynomial	0.54	0.48	3.25
	Radial	0.65	0.27	3.28
	Sigmoid	0.12	10.34	51.36
SOC	Linear	0.04	0.16	0.56
	Polynomial	0.001	1.80	2.94
	Radial	0.001	0.82	1.06
	Sigmoid	0.001	11.62	22.01



(a)

(b)

(c)

(d)

Figure 3: The relationship between observed and predicted values of the soil properties illustrated by the R-squared values in (sand), (b) silt, (clay), and (d) SOC

It is seen that the highest R-squared value between the predicted and observed values was obtained in clay with 0.71, while the least in soil organic carbon with 0.04. This may be due to the high dependency of soil organic carbon on the relief factor, low stabilization at a higher altitude, and higher density at lower altitudes with higher stabilization (Sheikh et al., 2009). Whereas the spatial variability of soil texture, which is the physical property of soil has a close association to parent materials that are not highly varied as the relief. The box plot in Figure 4 clearly depicts that SOC in the collected soil samples had outliers which may be due to the contributing factor of relief as described, while the soil textures do not have such outliers.

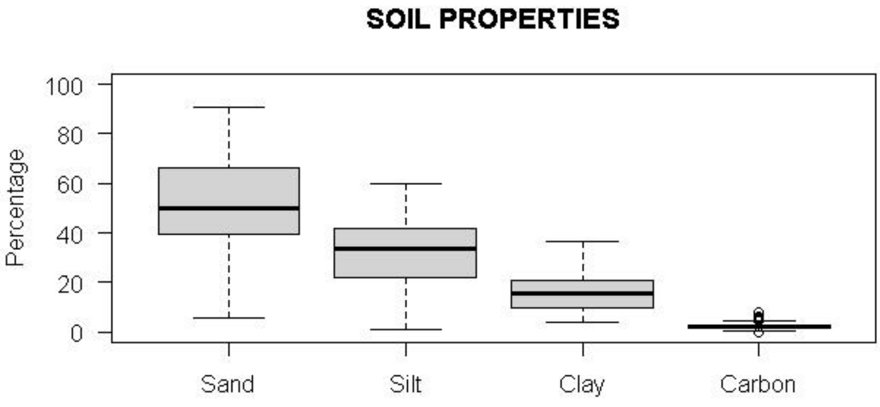


Figure 4: Box plot of soil properties

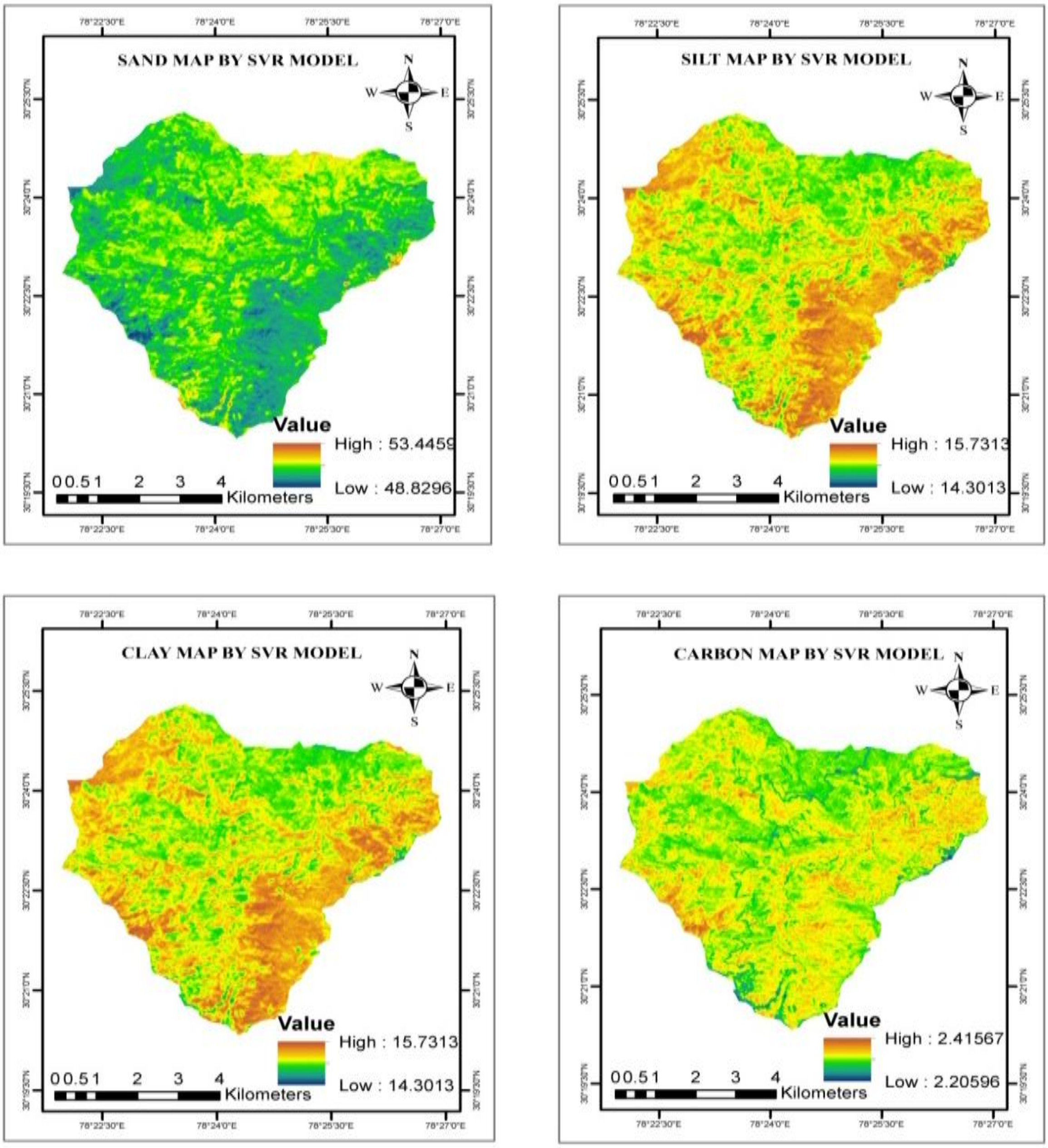


Figure 5: Spatially predicted maps of the soil properties by the SVR model

CONCLUSION

The relatively consistent primary objective of digital soil mapping is to spatially map the soil distribution and to delineate uniform management areas which are useful for decision-makers and resource conservation point of view. This study demonstrates an approach for spatially mapping the soil properties in the Himalayan watershed with laboratory analyzed soil data and remotely sensed data by employing a machine learning algorithm.

Even the mountainous or hilly terrains have different ecosystems in different parts of the region with a huge variation in the soil-forming factors, which limit one to adopt models used by other researchers in a similar environment. The varying soil-landscape relationships across different topographies make it even more difficult to decide the best model or approach for spatially predicting and mapping soil properties. High soil variation may exist in a small area with varying terrain and different soil-forming factors so mapping for bigger areas is generalized in most cases. These challenges need to be settled with sophisticated technologies and constant refinement of the modeling approach has to be addressed considering the variability of the landscapes.

These approaches making use of the increasingly available free remote sensing and terrain data employing machine learning techniques will aid in improving soil information at different scales especially in mountainous terrains with limited resources and much-reduced costs. Ultimately, such rapid and reliable soil information systems with known uncertainties will help us in adopting various land management decisions for improving agricultural productivity, enhanced resource utilization as well as environmental sustainability.

Acknowledgements

The authors are thankful to the Director, Indian Institute of Remote Sensing (IIRS), ISRO, Dehradun for providing an excellent research environment, and facilities for carrying out the study. The first author is grateful to Dr. Abhijit Rabha (IFS) and all the faculties of IIRS, ISRO for their guidance and support throughout the Post-graduate diploma course.

References

Bahuguna, H. S., Chaturvedi, R. K., & Rajwar, G. S., 2018. Carbon sequestration potential of the forest soils of district Tehri Garhwal, Uttarakhand, India. TROPICAL ECOLOGY, 59(4), 659-678.

Ballabio, C., 2009. Spatial prediction of soil properties in temperate mountain regions using support vector regression. Geoderma, 151(3-4), 338-350.

Brevik, E. C., Calzolari, C., Miller, B. A., Pereira, P., Kabala, C., Baumgarten, A., & Jordán, A., 2016. Soil mapping, classification, and pedologic modeling: History and future directions. Geoderma, 264, 256-274.

Brungard, C. W., Boettinger, J. L., Duniway, M. C., Wills, S. A., & Edwards Jr, T. C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma, 239, 68-83.

Bui, E. N., & Moran, C. J., 2003. A strategy to fill gaps in soil survey over large spatial extents: an example from the Murray–Darling basin of Australia. Geoderma, 111(1-2), 21-44.

Forkuor, G., Hounkpatin, O. K., Welp, G., & Thiel, M., 2017. High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. PloS one, 12(1), e0170478.

Gunn, S. R., 1998. Support vector machines for classification and regression. ISIS technical report, 14(1), 5-16.

Hastie, T., Tibshirani, R., & Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. (pp. 587-604). Springer, New York, NY.

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E., & Schmidt, M. G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma, 265, 62-77.

Ivanciuc, O., 2007. Applications of support vector machines in chemistry. Reviews in computational chemistry, 23, 291.

Jenny, H., 1941. Factors of soil formation. 281 pp. New York, 801.

Kuhn, M., & Johnson, K., 2013. Applied predictive modeling (Vol. 26). New York: Springer.

Kuhn, M., 2017. caret: Classification and regression training (version R package version 6.0–76).

Lagacherie, P., & McBratney, A. B., 2007. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. Developments in soil science, 31, 3-22.

Li, J., & Heap, A. D., 2008. A review of spatial interpolation methods for environmental scientists.

Li, J., Heap, A. D., Potter, A., & Daniell, J. J., 2011. Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling & Software, 26(12), 1647-1659.

Li, X., & Zhao, H., 2009. Weighted random subspace method for high dimensional data classification. Statistics and its Interface, 2(2), 153.

McBratney, A. B., Santos, M. M., & Minasny, B., 2003. On digital soil mapping. Geoderma, 117(1-2), 3-52.

Minasny, B., McBratney, A. B., Malone, B. P., & Wheeler, I., 2013. Digital mapping of soil carbon. In Advances in Agronomy (Vol. 118, pp. 1-47). Academic Press.

Scull, P., Franklin, J., Chadwick, O. A., & McArthur, D., 2003. Predictive soil mapping: a review. Progress in Physical Geography, 27(2), 171-197.

Sheikh, M. A., Kumar, M., & Bussmann, R. W., 2009. Altitudinal variation in soil organic carbon stock in coniferous subtropical and broadleaf temperate forests in Garhwal Himalaya. Carbon balance and management, 4(1), 6.

Vapnik, V., 1995. The nature of statistical learning theory springer new york google scholar. New York.