SIMULATION OF SENTINEL-2 DATA USING HYPERSPECTRAL DATA FOR BARE SURFACE SOIL MOISTURE ESTIMATION

Jayantrao Mohite, Suryakant Sawant, Ankur Pandit and Srinivasu Pappula

TCS Innovation Labs, Tata Consultancy Services, India

ABSTRACT

The main objective of this study is the assessment of simulated Sentinel-2 data for bare surface soil moisture estimation. Hyperspectral soil moisture data provided by [1] in the spectral range of 450-950 nm over test site at Germany has been utilized in the study. Simulation of 8 bands (B2-B8A) of Sentinel-2 covering the same spectral range as that of hyperspectral camera was carried out using spectral response curve of Sentinel-2. Random Forest Regression based model was developed using all bands. Morover, important features were selected based on %IncMSE. Selected bands include B2, B4, B5 and B7. Evaluation of models developed using all the 8 bands and selected bands was carried out using the testing data. Root Mean Square Error of 1.0180 and R^2 value of 0.9131 was achieved for a model with 8 bands. However, RMSE was reduced to 0.9661 and R^2 was increased to 0.9357 in case of selected 4 bands. Moreover, validation of models developed on simulated data was carried out using Sentinel-2 satellite observations on demo farm, Pune India. Difference between actual and estimated soil moisture was found to be between -2.10 to 3.18.

***Index Terms*—** Hyperspectral Sensing, Soil Moisture, Simulation of Satelllite Data, Sentinel-2

1. INTRODUCTION

In agriculture, soil moisture is one of the driving variables for precision irrigation management [2, 3], drought assessment [4] etc. Various approaches have been established in the past for measurement and monitoring of soil moisture ranging from laboratory-based methods, sensor-based measurement, remote sensing based spatial-temporal soil moisture estimation and monitoring methods etc. Laboratory-based methods such as the Gravimetric method [5] are based on the destruction of soil samples in the laboratory. Hence methods are time-consuming and also can not be reproduced. Various sensors are used for point-based measurement of soil moisture such as Time Domain Reflectometry (TDR) [6], neutron probes [7] and gamma-ray scanners [8]. As the sensors measure the soil moisture at a single point, they can not be used as an indicator for regional-scale studies [9].

Remote Sensing based methods provide an alternative way for the regional-level soil moisture estimation using the Microwave, Optical (Multispectral and Hyperspectral) and Thermal region of the electromagnetic spectrum. Microwave (Active/Passive) remote sensing has been used for soil moisture estimation due to its advantages of cloud penetration, all-weather capabilities etc [10]. However, the availability of high spatial and temporal resolution microwave datasets is a challenge and hence it has limited applicability at field level applications such as irrigation management, water stress detection etc. Alternatively, thermal infrared remote sensing has been also used for soil moisture estimation mainly due to the proven strong relationship between thermal properties of soil and soil moisture [9]. However, the availability of the thermal infrared data is a challenge due to the high costs associated with the sensor. Hyperspectral remote sensing has been extensively used for various applications including soil moisture estimation due to the availability of rich spectral information. Many researchers have successfully demonstrated the use of the optical domain for soil moisture estimation [11, 12, 13]. however the availability of near-real-time data from hyperspectral sensors is a challenge. There are very few operational hyperspectral sensors in the space which are capturing the continuous spatial-temporal data. Contrary to this, data from multiple optical sensors such as Sentinel-2, Landsat-8, MODIS and commercial satellite such as World-view, Quickbird etc are available. Hence we have attempted to simulate the hyperspectral data to Sentinel-2 data for soil moisture estimation. The main objective of this paper is the assessment of simulated Sentinel-2 data for bare surface soil moisture estimation. We have utilized the hyperspectral dataset on soil moisture provided by [1]. Sentinel-2 data has been chosen due to its high spatial-temporal availability over the Indian region. The developed models were then applied to Demo Farm, Pune, India using the Sentinel-2 observations and model evaluation was carried out.

2. DATASETS USED AND VALIDATION SITE

2.1. Ground Soil Moisture Data:

The present study has utilized the soil moisture dataset collected by [1] which is freely available and open under GNU GENERAL PUBLIC LICENSE Version 2, June 1991 (https://www.gnu.org/licenses/gpl-2.0.html). For the data collection,a five-day field campaign was carried out in Karlsruhe, Germany during May 2017 [1] . An undisturbed bare soil (type- clayey silt) samples without any vegetation near Waldbronn, Germany was used for the measurements. In-situ soil moisture was measured using the TRIME-PICO time-domain reflectometry (TDR) sensor. The soil moisture variations are ranging from 25% to 42% [14]. While the sensors were placed at various depths ranging from 2 to 18 cm, the sensor data at 2 cm which represents the subsurface soil moisture was available for public use [1]. Hence, we have used the soil moisture data at 2 cm depth in our study.

2.2. Hyperspectral Data:

Hyperspectral image data collection was carried using Cubert UHD 285 hyperspectral snapshot camera in the spectral range of 450-950 nm [14]. The camera was placed on a tripod at a height of 1.7m. The camera records the images with 50 by 50 pixels with a spectral resolution of 4 nm and collects the data in 125 spectral bands. The data available for public use consist of mean spectra of soil surface (mean value of 50 by 50 pixels at a particular time instance per spectral bands) which was calibrated using the spectralon spectrum. We have used the same aggregated data available in 125 spectral bands in our study. The total size of the dataset was 679 spectra.

2.3. Validation Sites used in the study:

Demo farm maintained by our organization named Tata Consultancy Services (TCS) Limited at Pune, located in the Maharashtra state of India. The location map of the Pune demo farm is given in Figure 1. We have deployed the KDS-042 sensor developed by Komoline Aerospace Limited for in-situ soil moisture measurement and monitoring. Data from node 3 was collected at an hourly interval. In addition to this, we have also accessed the cloud-free Sentinel-2 data available during Nov.-Dec. 2017 for the demo farm.

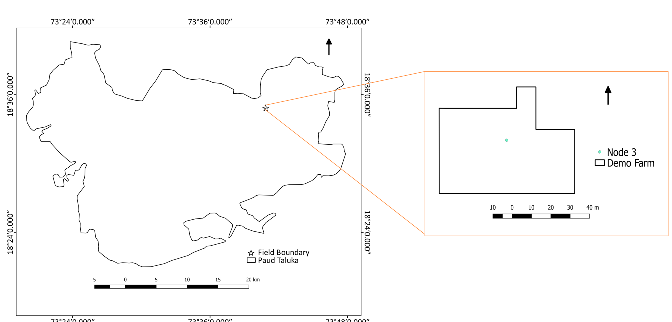

Fig. 1. Validation Site: TCS Demo Farm Located in Pune, India

3. OVERALL ANALYSIS FRAMEWORK

Figure 2 represents the overall analysis approach followed in this study. Key components of the framework include a. simulation of hyperspectral data to Sentinel-2, b. model development, variable selection using Random Forest Regression (RFR) and validation, c. evaluation of the best model using Sentinel-2 data on Demo Farm, Pune, India.

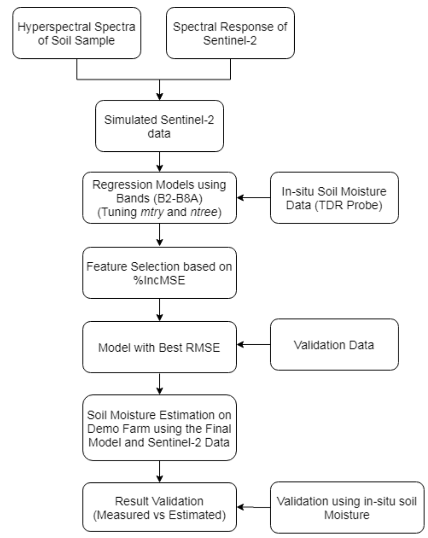

Fig. 2. Overall Analysis Framework

In the first step, we have simulated the Hyperspectral data available in the spectral range of 450-950 nm. The spectral range of hyperspectral data covers the eight bands of Sentinel-2 namely Blue (B2), Green (B3), Red (B4), Red Edge 1 (B5), Red Edge 2 (B6), Red Edge (B7), Near Infrared (B8), Narrow Near Infrared (B8A) with a center wavelength of 490, 560, 665, 705, 740, 785, 842 and 865 nm respectively [15]. The spectral response curve available for Sentinel-2 was used for calculating weighted spectral reflectance. A 'hsdar' package developed in R programming environment has implemented spectral response functions for various satellite sensors such as Sentinel-2, Landsat-8, Quickbird, RapidEye, Worldview-3 etc [16]. The 'spectralResampling' function for Sentinel-2 was used to calculate the weighted average reflectance using the hyperspectral data [16]. Using the function, hyperspectral data in 125 bands was simulated to the 8 bands available in Sentinel-2. As a second step, data in those 8 simulated bands have been fed to Random Forest Regression (RFR). The method has been widely accepted due to its ability to avoid over-fitting and works well in limited and noisy datasets [17].Two parameters of the RFR need to be fixed- 1. The ntree i.e. the number of trees in the forest and 2. mtry i.e. the number of variables available at each split. As there is no standard rule on the optimal number of trees and may vary depending on the data and the problem at hand, we have varied ntree from 50 to 300 with an interval of 50. However, mtry has been tried between 1 to 7 with an interval of 1. Maximum mtry has been restricted to 7 as the total features were 8. We have initially split the data into model development and testing data. The total size of the dataset was 679 samples, out of 544 samples were used for model training, however, 135 samples were used for independent testing. We have separated those 135 samples at the start of the experiment and used for model testing. Model training was carried out using three-fold cross-validation on 544 samples. Average RMSE from three folds was used as a performance metric to decide the best model. The best model is defined as the model with a certain value of mtry and ntree which produced the lowest RMSE. Further, to identify the significance of each feature i.e. band, we have carried out feature selection using RFR. Increasing Mean Squared Error (%IncMSE) is widely used to evaluate the importance score of a given variable in the case of RF. We have ranked the variables based on the %IncMSE value for each variable. The high value of %IncMSE depicts the high significance of that particular variable. The final selection of total significant variables was carried out using the following approach. The highest value of %IncMSE was divided by two. The variables with %IncMSE value greater than half of the highest value of %IncMSE were considered important and used in further analysis. We have again evaluated the performance of these important features using RFR. For the selected bands, we have tuned the RFR for ntree only as the number of features were less compared to the original dataset. As a last step, performance of best model was evaluated on Demo Farm Pune, India using real Sentinel-2 observations.

4. RESULTS AND DISCUSSION

This section contains three subsections. Subsection 4.1 discusses about the results obtained using the all simulated Sentinel-2 bands and further to this, Subsection 4.2 important simulated Sentinel-2 bands obtained using RF. Moreover, Subsection 4.3 provides the results and validation on Demo Farm using real Sentinel-2 data.

4.1. Performance of Simulated Sentinel-2 bands for soil moisture estimation:

We have simulated the soil moisture Hyperspectral data to Sentinel-2 bands using the SRF of Sentinel-2. The simulated Sentinel-2 data in 8 bands (B2-B8A) was used as a feature-set for the estimation of soil moisture using the tuned RF model. Figure 3 shows root mean square error value obtained using the three-fold cross validation for each combination of ntree and mtry (total 42 models). Figure 3 clearly indicates that the parameters such as ntree and mtry affect the performance of RF model and shows the need for parameter tuning. Best results (i.e. lowest RMSE of 1.0396) are obtained at mtry of 5 and ntree of 200. Moreover, lower RMSE has been observed for mtry values between 5-7 and ntree of 200 and 250 i.e. towards higher values of ntree. This shows that standard approaches of selecting mtry for regression problems (number of features divided by 3 i.e 3) may not be always optimal and shows the need for the tuning of the models for wider ranges to get the best performance.

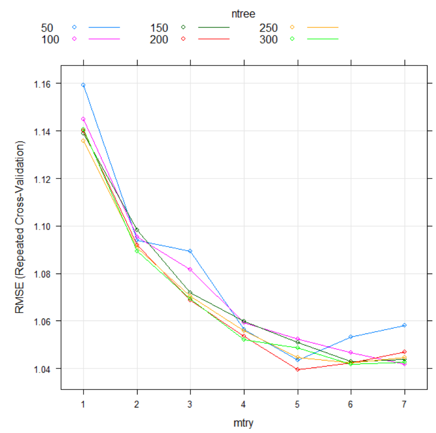

Fig. 3. Model Performance on Training Data using All Bands

Further, we have tested the best model which produced lowest RMSE on test data. The test data was not considered in the model development and kept for independent testing of the model. Figure 4 shows the scatter plot between the actual soil moisture (test data) and predicted soil moisture. Results shows the RMSE value of 1.0180 and R^2 value of 0.9131. This means model with 8 bands was able to explain 91.31% variation of the test data.

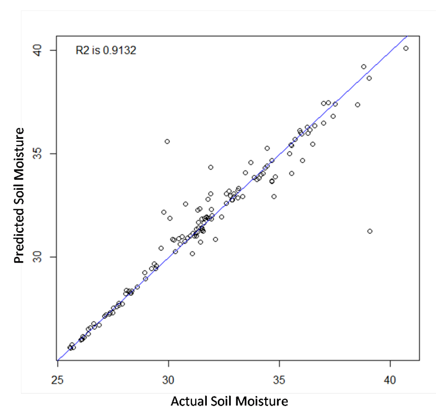

Fig. 4. Performance of Best Model using All Bands on Test Dataset

4.2. Band Importance Evaluation and Selection

Figure 5 shows the band importance score i.e. percentage increase in Mean Squared Error (%IncMSE). The results shows that, bands such as B2, B4, B5 and B7 which represents Blue, Red, Red Edge 1 and Red Edge 3 are showing the high values of %IncMSE which are greater than 10. Higher values of %IncMSE depicts the higher importance of the feature. This means that those bands are contributing more than other bands for soil moisture estimation.

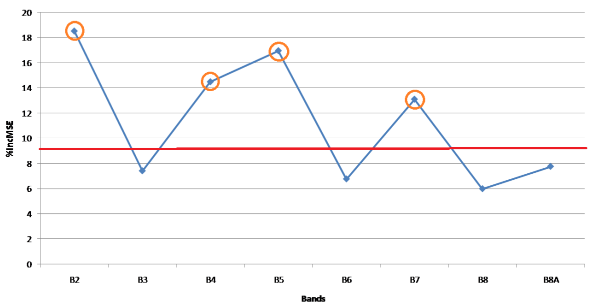

Fig. 5. Importance of Simulated Sentinel-2 Bands

We have attempted to use the important bands (B2, B4, B5 and B7) for soil moisture estimation using the RF model. As the number of features are reduced to 4 we have only tuned the model for ntree. Model configuration such as train (cross-validation), test split, numbers of iterations are kept same as that of earlier scenario. Figure 6 shows the RMSE for various values of ntree. Best results i.e. Lowest RMSE of 1.2471) in case of cross-validation was obtained with highest number of trees i.e 300. In addition to this, as the number of features were reduced, the time required for model training was less in case of selected band scenario which was obvious.

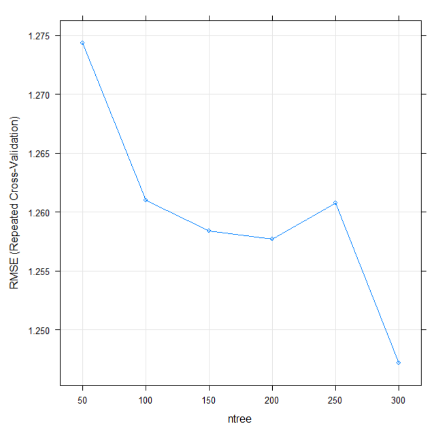

Fig. 6. Model Performance on Training Data using Selected Bands

Moreover, we have estimated the performance of model on testing data (which was not considered in the model training). Figure 7 shows the R^2 value 0.9357. RMSE was reduced to 0.9661. Performance of the best model developed from important 4 bands data was better in terms of R^2 and RMSE on the test dataset. This depicts that, considering data from all the bands is adding some noise in the data which resulted slightly lowering the performance and shows the need for optimum band selection.

Table 1. Comparison of Actual and Predicted Soil Moisture

SN	S-2 Overpass	ASM	PSM	Difference
1	11 Nov. 2017	25.6	27.15	-1.55
2	16 Nov. 2017	26.6	26.89	-0.29
3	21 Nov. 2017	28.1	24.92	3.18
4	01 Dec. 2017	26.4	25.56	0.84
5	06 Dec. 2017	23.5	25.70	-2.20
6	11 Dec. 2017	26.5	28.22	-1.72
7	21 Dec. 2017	27.9	30.00	-2.1

ASM-Actual Soil Moisture, PSM-Predicted Soil Moisture

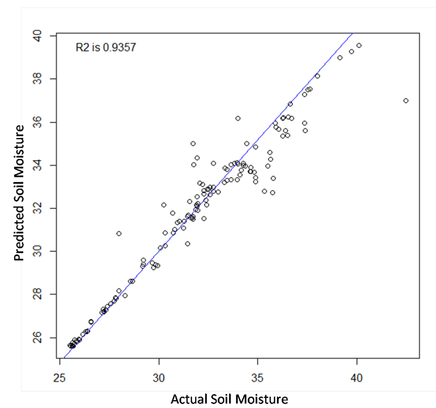

Fig. 7. Performance of Best Model using Selected Bands on Test Dataset

4.3. Validation of the Models on Indian Fields using Real Sentinel-2 Observations:

In-situ soil moisture data was available during Nov-Dec 2017 on selected dates for demo farm located in Pune. We have accessed the Sentinel-2 observations from Google Earth Engine for the dates on which the in-situ soil moisture data was available during the Nov-Dec 2017. The data was available at a node 3 within the field. Hence data for pixel with the soil moisture sensor location was considered for soil moisture estimation. Table 1 shows the comparison between in-situ soil moisture and estimated soil moisture from Sentinel-2 observations. Results show that, difference between actual and estimated soil moisture is between -2.10 to 3.18 %. This depicts that, results are significant and models developed on simulated Sentinel-2 data can be applied on Sentinel-2 satellite observations to estimate bare surface soil moisture.

5. CONCLUSION

Successful evaluation of Simulated Sentinel-2 data for bare surface soil moisture estimation was carried in this study. Performance of RFR models developed using all the 8 bands and selected 4 bands was carried out. Results showed that, RMSE of 1.0180 and R^2 value of 0.9131 was achieved for a model with 8 bands. However, RMSE was reduced to 0.9661 and R^2 was increased to 0.9357 in case of selected 4 bands. Best models i.e. models with lowest RMSE have been found with larger values of mtry than that of standard rules suggested for mtry selection (number of features/3 for regression). In addition to this, there were significant variations in RMSE for different values of ntree. This shows the need for tuning of RF for achieving the best performance. We have also observed that, feature selection using the %IncMSE have shown the improvement in the performance. Quantitative evaluation of the model shows the difference of -2.10 to 3.18 percent between actual and estimated soil moisture on Demo Farm, Pune, India. This shows that, models developed on simulated data were able to estimate the bare surface soil moisture using Sentinel-2 satellite observations.

6. REFERENCES

[1] Felix M Riese and Sina Keller, "Hyperspectral benchmark dataset on soil moisture," *doi.org/10.5281/zenodo*, vol. 1227837, 2018.

[2] JM Blonquist Jr, Scott B Jones, and DA Robinson, "Precise irrigation scheduling for turfgrass using a subsurface electromagnetic soil moisture sensor," *Agricultural water management*, vol. 84, no. 1-2, pp. 153–165, 2006.

[3] Jayantrao Mohite, Minal Sawant, Suryakant Sawant, Rajesh Urkude, and Srinivasu Pappula, "Decision support system for near real time irrigation scheduling using proximal sensor networks and human participatory sensing," in *2017 EFITA WCCA CONGRESS*, vol. 43.

[4] José Martínez-Fernández, A González-Zamora, Nilda Sánchez, A Gumuzzio, and CM Herrero-Jiménez, "Satellite soil moisture for agricultural drought monitoring: Assessment of the smos derived soil water deficit index," *Remote Sensing of Environment*, vol. 177, pp. 277–286, 2016.

[5] Edwin T Engman, "Applications of microwave remote sensing of soil moisture for water resources and agriculture," *Remote Sensing of Environment*, vol. 35, no. 2-3, pp. 213–226, 1991.

[6] J Ledieu, P De Ridder, P De Clerck, and S Dautrebande, "A method of measuring soil moisture by time-domain reflectometry," *Journal of Hydrology*, vol. 88, no. 3-4, pp. 319–328, 1986.

[7] DS Chanasyk and Ma A Naeth, "Field measurement of soil moisture using neutron probes," *Canadian Journal of Soil Science*, vol. 76, no. 3, pp. 317–323, 1996.

[8] OOS Bacchi, KJCM REICHARDT, JCM Oliveira, and DR Nielsen, "Gamma-ray beam attenuation as an auxiliary technique for the evaluation of the soil water retention curve," *Scientia Agricola*, vol. 55, no. 3, pp. 498–502, 1998.

[9] Dianjun Zhang and Guoqing Zhou, "Estimation of soil moisture from optical and thermal remote sensing: A review," *Sensors*, vol. 16, no. 8, pp. 1308, 2016.

[10] Ankur Pandit, Suryakant Sawant, Jayantrao Mohite, and Srinivasu Pappula, "A data-driven approach for bare surface soil moisture estimation using sentinel-1 sar data and ground observations," *Geocarto International*, pp. 1–30, 2020.

[11] David B Lobell and Gregory P Asner, "Moisture effects on soil reflectance," *Soil Science Society of America Journal*, vol. 66, no. 3, pp. 722–727, 2002.

[12] Liu Weidong, Frédéric Baret, Gu Xingfa, Tong Qingxi, Zheng Lanfen, and Zhang Bing, "Relating soil surface moisture to reflectance," *Remote sensing of environment*, vol. 81, no. 2-3, pp. 238–246, 2002.

[13] Felix M Riese and Sina Keller, "Fusion of hyper spectral and ground penetrating radar data to estimate soil moisture," in *2018 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE, 2018, pp. 1–5.

[14] Felix M Riese and Sina Keller, "Introducing a framework of self-organizing maps for regression of soil moisture with hyperspectral data," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 6151–6154.

[15] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al., "Sentinel-2: Esa's optical high-resolution mission for gmes operational services," *Remote sensing of Environment*, vol. 120, pp. 25–36, 2012.

[16] LW Lehnert, H Meyer, and J Bendix, "Hsdar: manage, analyse and simulate hyperspectral data in r," *R Package Version 0.4*, vol. 1, 2016.

[17] Bing Wang, Kun Jia, Shunlin Liang, Xianhong Xie, Xiangqin Wei, Xiang Zhao, Yunjun Yao, and Xiaotong Zhang, "Assessment of sentinel-2 msi spectral band reflectances for estimating fractional vegetation cover," *Remote Sensing*, vol. 10, no. 12, pp. 1927, 2018.