www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 15(4), Nov 2024, 61-71, ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2024.v15i4.502

Study of Machine Learning Techniques for the Estimation of Soil Moisture in Agriculture

Noel A. Zavala-Díaz[1], *Juan C. Olivares-Rojas*[1], *Jonathan Zavala-Díaz*[1], *Enrique Reyes-Archundia*[1], *Adriana Téllez-Anguiano*[1], *Gerardo M. Chávez-Campos*[1], *Arturo Méndez-Patiño*[1].
[1] División de Estudios de Posgrado e Investigación, Tecnológico Nacional de México / I. T. de Morelia, Morelia, Michoacán, México.
{m21121662, juan.or, d19123006, enrique.ra, adriana.ta, gerardo.cc, artu-ro.mp}@morelia.tecnm.mx.

Abstract. Soil moisture is crucial in various fields and monitoring it to guide irrigation is challenging. Machine learning has emerged as a promising tool to predict soil moisture levels accurately. This study evaluates machine learning techniques for this task, training models with meteorological variables and direct soil moisture measurements. Four machine learning algorithms were implemented, highlighting the Gradient Boosting Regressor as the most effective. In addition, a processed data set that combines meteorological and soil moisture measurements is presented, hoping it will be helpful for future research. This approach seeks to improve the compression and predictability of soil moisture, which is crucial for agricultural planning and water management in agriculture

Keywords: Soil moisture, machine learning, regression models.

Article Info
Received Jul 14, 2024.
Accepted Sep 11, 2024.

1 Introduction

Soil moisture content is vital to various fields, including biology, hydrology, agronomy, engineering, ecology, and soil geology. Its monitoring is increasingly extensive, especially with increased precision irrigation infrastructure and control systems investments. However, monitoring soil moisture to guide irrigation presents significant challenges. Irrigators must carefully select the appropriate equipment for their irrigation system and the specific characteristics of their land plot (Rasheed et al., 2022). Soil moisture is crucial in providing water for agriculture, being its primary resource. Despite its importance, direct field measurement faces significant challenges, underscoring the need to accurately predict it to support agricultural planning activities and relevant research (Palominos-Rizzo et al., 2022).

Machine learning has led to the development of innovative algorithms capable of accurately predicting soil moisture levels, which can be used in irrigation activities or other purposes (Rasheed et al., 2022). Currently, some works apply machine learning in soil moisture prediction. In (Lee et al., 2019), they estimate soil moisture using deep learning based on satellite data. The authors (Ahmad et al., 2010) introduce a new regression technique called Support Vector Machine (SVM) to estimate soil moisture using remote sensing data. In the work (Prasad et al., 2018), hybrid models that combine Extreme Learning Machines (ELM) with data intelligence are developed and examined to perform monthly soil moisture predictions.

In (A et al., 2022), they estimate soil moisture based on remote sensing data and deep learning. This study uses machine learning techniques to estimate soil moisture, motivated by finding a correlation between meteorological data and soil moisture measurements, given the absence of abundant data from our sensors, to provide a solid basis for another research. Models were trained with meteorological variables and direct soil moisture measurements. We implemented four machine learning algorithms: Random Forest Regressor, K-Nearest Neighbors, Gradient Boosting Regressor, and Multiple Linear Regression.

When evaluating these models with predictions, we found that the Gradient Boosting Regressor demonstrated lower mean squared error and mean absolute error than the other models, especially when tested in time intervals different from the training period. Finally, this model was applied to recent data from the Instituto Tecnológico de Morelia meteorological station, obtaining consistent soil moisture estimates that align with the expected behavior over time.

Among the contributions of this study, we highlight the presentation of a dataset created by C. K. Gasch et al (C. K. Gasch et al., 2017). This dataset was processed to be presented in a CSV file, extracted from its original TXT format. We identified the sensor locations with the least missing data and used the nearest neighbor interpolation method, given its temporal structure, to fill in the missing values. Additionally, we enriched this dataset with information from a meteorological station near the locations where the soil moisture measurements were taken. In this way, we created a dataset that integrates meteorological and soil moisture measurements. This dataset will serve as a basis for future research to identify patterns or conduct temporal analyses.

2 Theoretical framework

2.1 Soil moisture

Agriculture and water are deeply intertwined, with water being an essential factor in agricultural production. Agricultural methods influence the hydrological cycle through evapotranspiration, aquifer recharge, and surface water flow. Adequate soil moisture is crucial for various biological and physical processes, including seed germination, vegetative development, nutrient cycling, and soil biodiversity conservation. Measuring soil moisture is essential to assessing water availability for agriculture and understanding soil health and its capacity to retain water, which is vital for maintaining a sustainable agroecosystem (Kashyap & Kumar, 2021). Soil moisture is a crucial factor in agriculture, as it directly influences crop growth and the sustainability of agricultural ecosystems. This moisture depends on irrigation practices and soil management and is closely linked to various climatic variables.

2.2 Machine Learning Techniques

Machine learning techniques, such as Random Forest Regressor, K-Nearest Neighbors, Gradient Boosting Regressor, and Multiple Linear Regression, are powerful tools for predicting values in various contexts. These algorithms can model complex relationships between variables and generate accurate predictions about future values. The authors (Khalyasmaa et al., 2019) present the application of a specific machine learning method, Random Forest Regressor, to generate accurate daily forecasts of solar energy generation using historical measurement data and meteorological data from open sources provided by meteorological services. In (Gajula et al., 2021), a proposed method uses the K-Nearest Neighbors (KNN) algorithm to assess soil quality and predict the most suitable crops. This approach considers temperature and soil quality as input variables for the algorithm. The article (Ponraj & Vigneswaran, 2020) uses machine learning models to predict reference evapotranspiration, thus facilitating irrigation planning. Daily meteorological data, including maximum and minimum temperature, relative humidity, solar radiation, soil temperature, and wind speed, were used. The data were processed using Multiple Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor techniques. The results indicated that the model preprocessed with GBR outperformed the other models' accuracy of reference evapotranspiration predictions.

The study (Ponraj & Vigneswaran, 2020) aimed to predict daily soil moisture at the crop level using meteorological information through multiple linear regression models. It was concluded that these models, by incorporating meteorological variables, effectively estimated soil moisture. This is because moisture tends to replicate seasonal patterns and respond to variations in precipitation.

2.3 Datasets

In (C. K. Gasch et al., 2017), a dataset is obtained from monitoring soil water content and complementary data collected at a 37-hectare zero-tillage experimental farm in the northwest United States. The water content measurements have been taken hourly since 2007 using ECH2O-TE and 5TE sensors distributed across 42 locations, covering five depths (0.3, 0.6, 0.9, 1.2, and 1.5 meters), totaling 210 sensors throughout the RJ Cook agronomic farm. This dataset is available in (C. Gasch & Brown, 2017).

This dataset includes hourly and daily measurements of water content (in m^3/m^3) and soil temperature (in °C) at 42 locations and five depths (0.3, 0.6, 0.9, 1.2, and 1.5 meters) from April 20, 2007, to June 16, 2016. The data are stored in .txt files for each location. The website meteostat.net is a meteorological and climatic database that provides detailed data from thousands of weather stations and locations worldwide. Fortunately, it has a station in Pullman, very close to R.J. Cook Agronomy Farm, where soil water content measurements and auxiliary data were taken at different depths (*Meteostat*, n.d.).

The website (*Meteostat*, n.d.) offers the opportunity to obtain data in different ways; however, when downloaded over seven days (one week), the obtained data have an hourly frequency, which is like the dataset (C. Gasch & Brown, 2017).

3 Methodology

Fig. 1 shows the methodology employed in this study. First, a dataset that includes soil moisture and meteorological data is consolidated (see Section 3.1). Then, the machine learning techniques to be used are selected. Next, the selected algorithms are trained to estimate soil moisture. Subsequently, tests are conducted, and the effectiveness of the chosen algorithms is evaluated. Finally, the results obtained are analyzed.

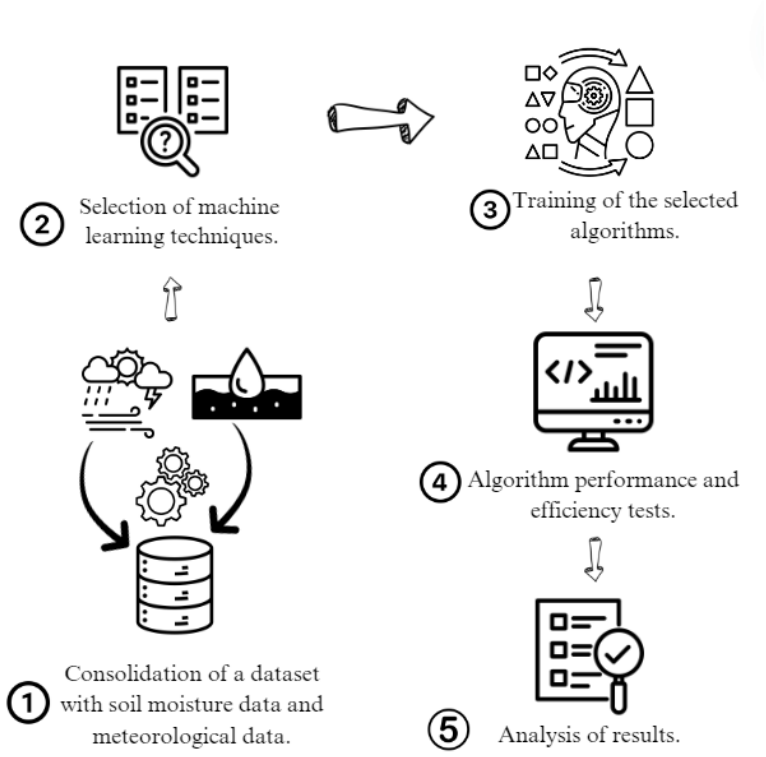

Fig. 1. Methodology

3.1 Consolidating a dataset containing soil moisture and meteorological data

This stage involves consolidating a dataset that includes soil moisture from the dataset (C. Gasch & Brown, 2017) and meteorological data obtained from (*Meteostat*, n.d.). First, Section 3.1.1 shows the process of obtaining the meteorological data dataset.

3.1.1 Meteorological data dataset

The process of obtaining the meteorological data consists of selecting the 7-day measurement period starting from 4/20/2007 within the Pullman station from the meteostat.net repository, then downloading the file by selecting the CSV format. This way, we will obtain weekly information files at hourly intervals. This process is repeated until 6/16/2016, corresponding to the dataset period (C. Gasch & Brown, 2017), with which a new dataset will be consolidated, including soil moisture and meteorological data.

Table 1. Variables found in meteorological dataset, definitions and null values.

Variable	Definition	Null Values
time	Time	0
temp	Temperature	213
dwpt	Dew point	256
rhum	RH	256
prcp	Precipitation	6558
snow	Snow depth	80283
wdir	Direction of the wind	23414
wspd	Average wind speed	414
wpgt	Maximum wind gust	80283
pres	Pressure	1125
tsun	Sun time	80283
coco	Climate condition code	80283
Total		353368

Several methods are available, such as mean, median, or mode imputation, linear or multiple regression, MICE (Multiple Imputation by Chained Equations), matrix factorization, advanced machine learning algorithms, and interpolation. Since the data are ordered in time and have a temporal structure, interpolation methods can predict the missing values based on the existing values.

The "nearest" interpolation method, or nearest neighbor interpolation, is a form of interpolation based on the idea that values close in time (or in sequence) are more similar to each other so that the nearest value will be a good approximation for the missing value. Fig. 2 shows the meteorological variables of this dataset.

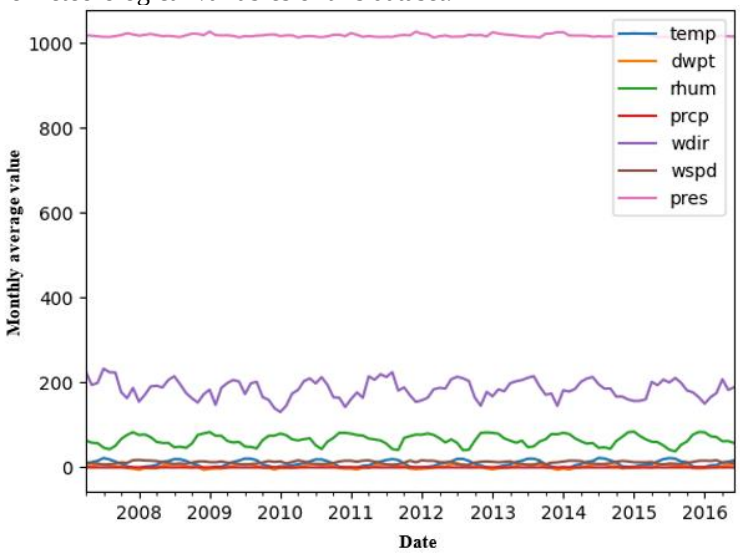

Fig. 2. Monthly average value of the meteorological variable's dataset

3.1.2 Soil moisture dataset

As mentioned earlier, this dataset includes hourly and daily records of soil moisture (expressed in m^3/m^3) and soil temperature (in °C) at 42 different locations and five different depths (0.3, 0.6, 0.9, 1.2, and 1.5 meters). These measurements span from April 20, 2007, to June 16, 2016. The data are stored in text files separated by location. The first challenge was determining which locations offered the best data quality, the one with the least number of null records. This selection was crucial to ensure the reliability of subsequent analyses.

After analyzing the corresponding text files, it was determined that the CAF308.txt file had the least number of null values compared to the files of the other 42 sensor locations, specifically concerning moisture measurements. Table 2 details the dataset's variables from this selected file for this study, including soil moisture and temperature at different depths, along with the number of null values found for each variable. It is noted that the soil moisture measurement at a depth of 30 cm has the fewest null values compared to other depths. The total number of values for these measurements, extending from April 20, 2007, to June 16, 2016, should be 80,283 records.

Table 2. Variables found in soil moisture dataset, definitions and null values.

Variable	Definition	Null Values
H_30cm	Humidity at 30 cm	18806
H_60cm	Humidity at 60 cm	23701
H_90cm	Humidity at 90 cm	22323
H_120cm	Humidity at 120 cm	24540
H_150cm	Humidity at 150 cm	25577
T_30cm	Temperature at 30 cm	18806
T_60cm	Temperature at 60 cm	23705
T_90cm	Temperature at 90 cm	22330
T_120cm	Temperature at 120 cm	24540
T_150cm	Temperature at 150 cm	25578
Total		229906

To address the missing values, the "nearest" interpolation method or nearest neighbor interpolation was used. Since the data are temporally ordered, values close in time tend to be more like each other, making it reasonable to assume that the nearest value is an adequate approximation for the missing value.

Figure 3 presents the graph of the monthly average soil moisture value at a depth of 30 cm, which will be the focus of this study. It covers the period from 2007 to 2016. However, at first glance, a more precise and significant trend can be seen between 2012 and 2016. Therefore, we will focus on this time interval for the subsequent analyses.

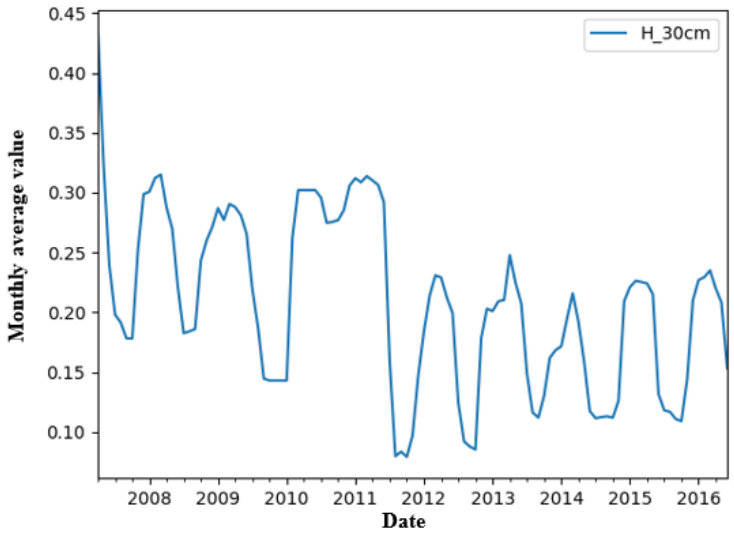

Fig. 2. Monthly average soil moisture values

3.1.3 Soil Moisture and Meteorological Data Dataset

After acquiring the soil moisture and meteorological datasets, we combined them into a single dataset. Then, we created a correlation matrix between the variables to explore possible relationships. Since our hypothesis suggested that relative humidity might correlate with soil moisture, we generated a relative humidity graph and applied a filter to smooth out the noise. We used a moving average with a window size 24, as shown in Fig. 4 c). The window size 24 for the moving average was selected to match the daily periodicity of the hourly collected data, covering a complete cycle. This window size facilitates the smoothing

of daily fluctuations and highlights clear trends in relative humidity, providing a solid basis for analyzing daily effects on soil moisture. Figure 4a shows the filtering with a window of size 6, while in Figure 4b, many variations are still observed compared to the window of size 24 in Figure 4c. In Figure 4d, a slight improvement in the variations is observed with a window size of 48. However, it is decided to use a window size of 24 to match the daily periodicity.

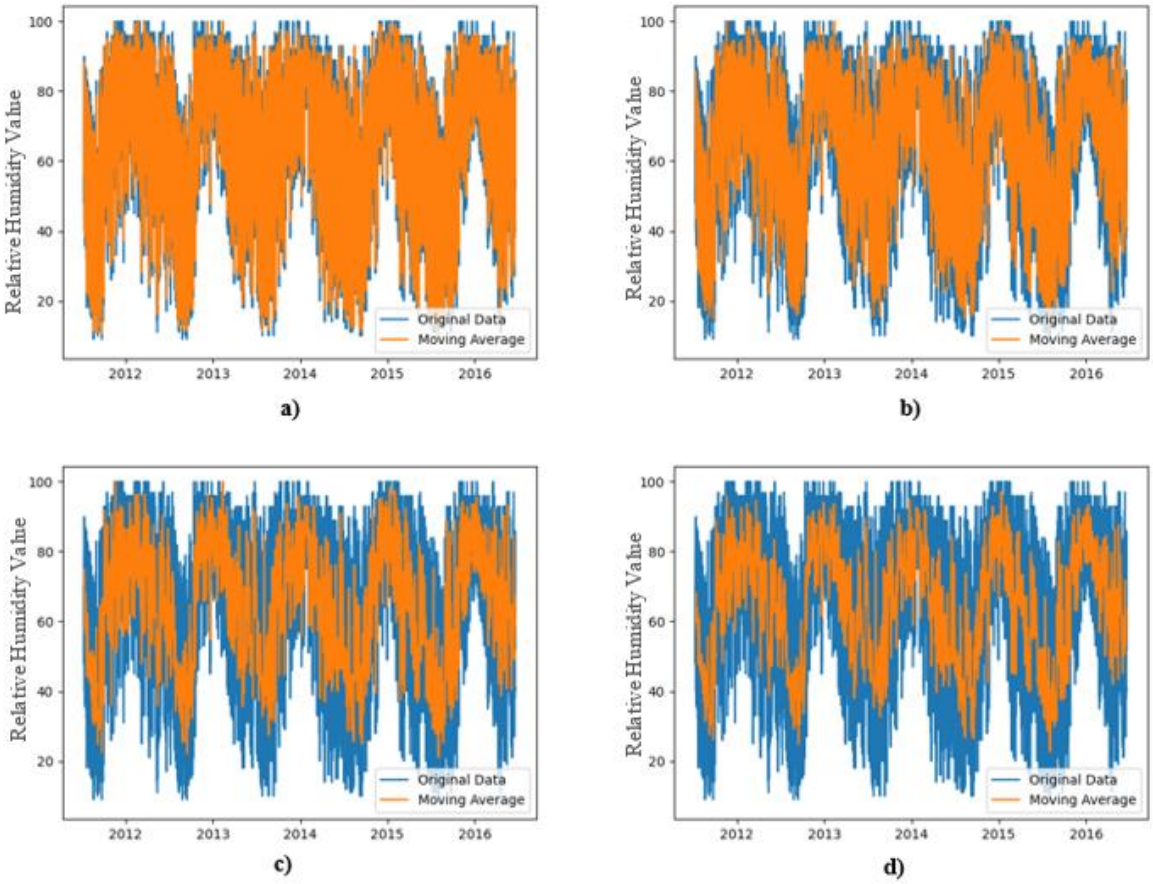

Fig. 4. Relative humidity with and without filter, a) window size of 6, b) window size of 12, c) window size of 24, and d) window size of 48

Figure 5 presents the correlation matrix between the variables. Notably, the correlation between Soil Moisture (H_30cm) and Relative Humidity (rhum) is 0.36. However, after applying the moving average filter to Relative Humidity, as shown in Fig. 5, this correlation increases to 0.47.

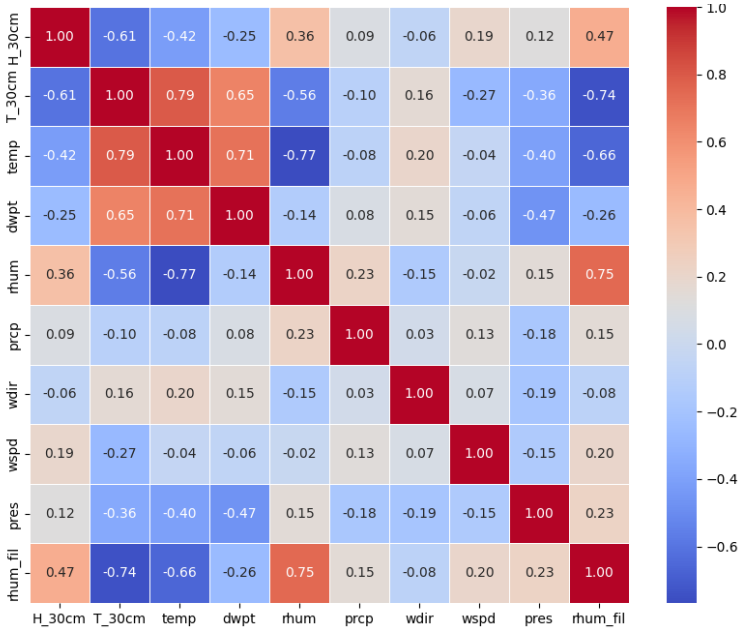

Fig. 5. Variable correlation matrix

4 Results

A comparative analysis of four machine learning techniques (Random Forest Regressor, Gradient Boosting Regressor, K-Nearest Neighbors, and Multiple Linear Regression) was conducted using meteorological variables to generate new soil moisture values for future dates. The selected variables were temperature (temp), dew point (dwpt), relative humidity (rhum), precipitation (prcp), and the corresponding month. The selection of these variables is justified by the following reason: Relative humidity showed a higher correlation with soil moisture, as observed in Fig. 5. Additionally, given the cyclical behavior observed in Figure 4, it was decided to include the month as a feature for model training. The other complementary variables were selected because they are available for future work using meteorological data from the city of Morelia and will be used to generate synthetic soil moisture values from the trained model.

Python and the Scikit-learn library were used to train the machine learning models, including Random Forest Regressor, Gradient Boosting Regressor, K-Nearest Neighbors, and Multiple Linear Regression. Scikit-learn is an open-source tool that offers various supervised and unsupervised algorithms for machine learning. This study used the default hyperparameter values for each model, ensuring a standard and consistent configuration during the training process.

The dataset presented in section 3.1.3, which includes information on soil moisture and meteorological data, was used to train the model. The period from 2011 to 2015 was used for the training process, with the interval from 2015 to 2016 reserved to test the model. During training with the 2011 to 2015 data, 80% of the data was allocated for training and the remaining 20% for testing. After training the models, they were evaluated using data from the 2015 to 2016 interval to test their predictions, as seen in Fig. 6. It was observed that the Gradient Boosting Regressor model showed greater accuracy in fitting the actual values, while the K-Nearest Neighbors model exhibited larger variations concerning the actual soil moisture values. This difference can be verified in Table 3, where the mean squared error (MSE) and mean absolute error (MAE) metrics are presented. A lower error is observed when using the Gradient Boosting Regressor model to estimate soil moisture with data different from those used during training.

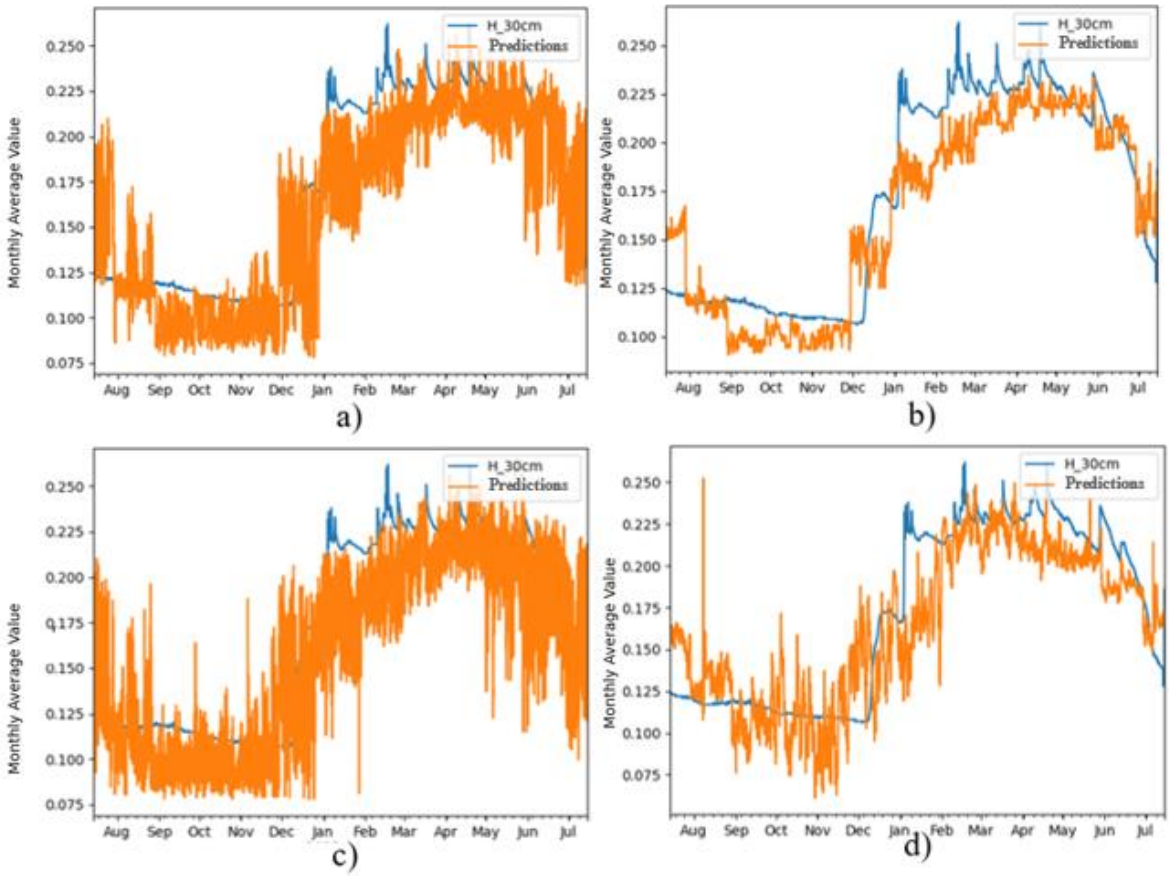

Fig. 6. Soil moisture prediction of the different trained models, a) Random Forest Regressor, b) Gradient Boosting Regressor, c) K-Nearest Neighbors and d) Multiple Linear Regression

To complement the results shown in Figure 6, a moving average filter of size 24 was implemented in the predictions presented to analyze whether applying this filter could reduce the error. Figure 7 shows the actual values, unfiltered predictions, and filtered predictions. Figures 7a and 7c show a notable change when applying the filter to the predictions. In Figures 7b and 7d, less variability is seen after applying the filter.

The mean squared error and the mean absolute error of the filtered predictions were calculated to corroborate the decrease in error. These values marked with '*' are presented in Table 3. It is observed that there is a reduction in error compared to unfiltered predictions. In the case of the Gradient Boosting Regressor model, which demonstrated better results in predicting soil moisture values for the 2015-2016 data set, the mean square error decreased from 0.000460 to 0.000448, and the mean absolute error from 0.017459 to 0.017328. For the K-Nearest Neighbors model, a more significant error reduction was observed when applying the filter to the predictions, going from 0.000814 to 0.000584 in the mean square error and from 0.022255 to 0.019771 in the mean absolute error.

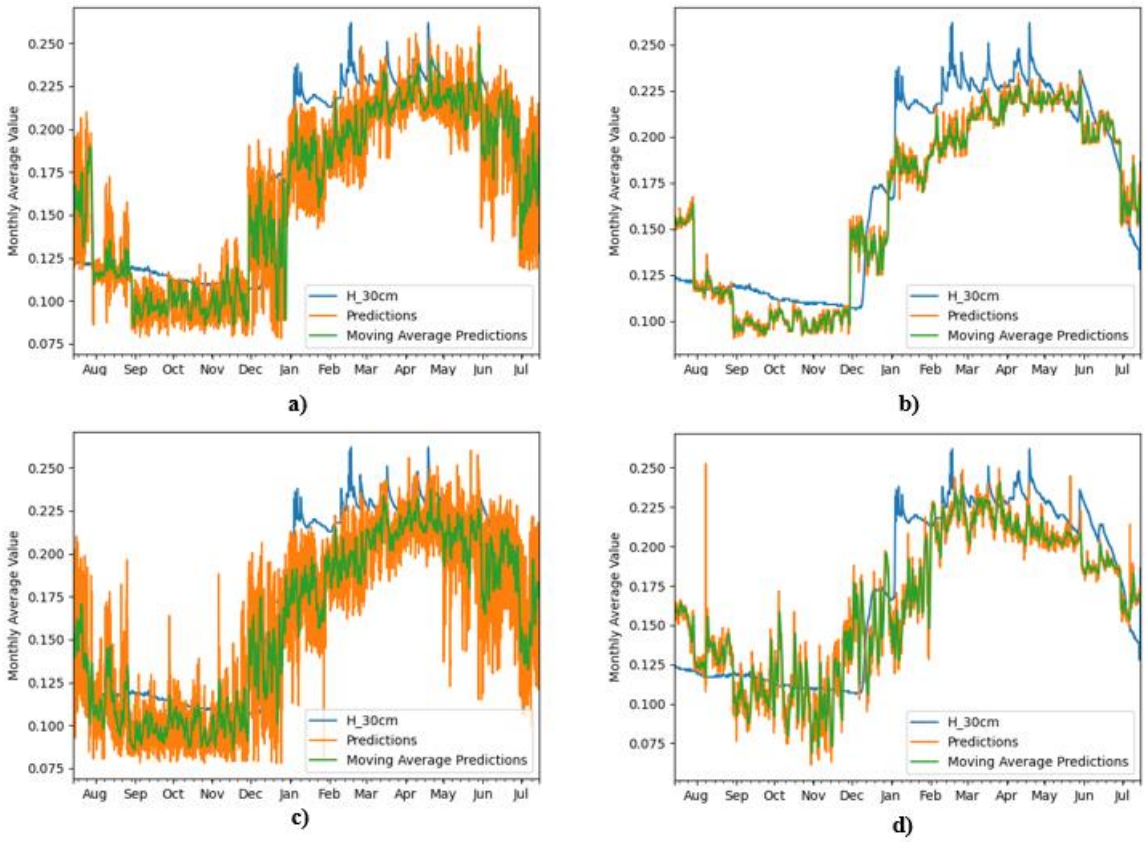

Fig. 7. Soil moisture prediction of the different trained models with filter, a) Random Forest Regressor, b) Gradient Boosting Regressor, c) K-Nearest Neighbors, and d) Multiple Linear Regression

Table 3. Mean squared error and Mean Absolute Error of the trained models.

Variable	Training ECM	Dataset 2015-2016 ECM	Training EAM	Dataset 2015-2016 EAM
Random Forest Regressor	**0.000401**	0.000736 0.000597*	**0.013026**	0.020937 0.019243*
Gradient Boosting Regressor	0.000574	**0.000460** **0.000448***	0.018599	**0.017459** **0.017328***
K-Nearest Neighbors	0.000525	0.000814 0.000584*	0.015645	0.022255 0.019771*
Regresión lineal múltiple	0.001033	0.000718 0.000681*	0.025296	0.021316 0.020781*

* error values for predictions applying 24-window moving media filter

Once we obtained the best-evaluated model, Random Forest Regressor, we made soil moisture predictions using meteorological data from the meteorological station at the Instituto Tecnológico de Morelia. For this, we needed the date from which the month was extracted, as well as the variables of temperature (temp), dew point (dwpt), relative humidity (rhum), and precipitation (prcp). Figure 8 shows the values generated by our model for the time interval from January 2021 to May 2023. A logical and consistent behavior is observed, in line with what was expected.

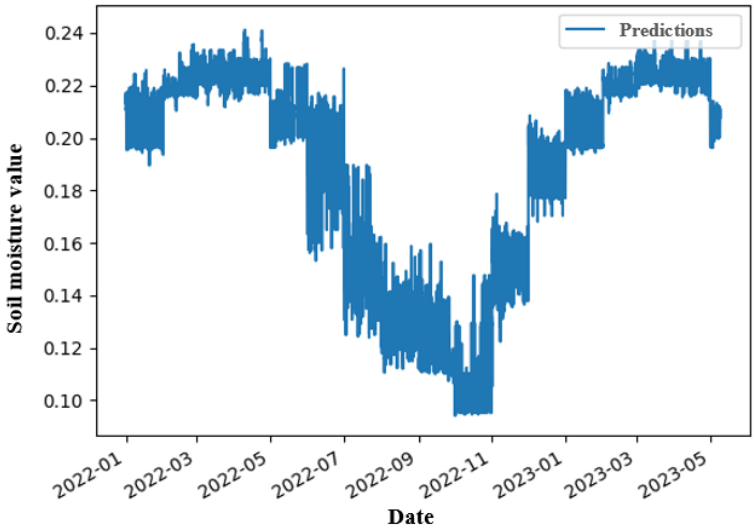

Fig. 8. Soil moisture prediction for ITM meteorological dataset

5 Conclusions

This work demonstrates the applicability of machine learning techniques for soil moisture estimation, with potential for future research and applications in the agricultural field. Various machine-learning techniques for estimating soil moisture in agriculture were explored and compared. The results obtained from the analysis revealed their potential to provide accurate and valuable estimates in agricultural planning and decision-making.

The models trained using Random Forest Regressor, Gradient Boosting Regressor, K-Nearest Neighbors, and Multiple Linear Regression demonstrated their effectiveness in predicting soil moisture using meteorological variables such as temperature, dew point, relative humidity, precipitation, and the corresponding month. Among these models, the Gradient Boosting Regressor stood out for its lower mean square error and mean absolute error, which suggests its greater predictive capacity compared to the other techniques evaluated. By applying a moving means filter with a window size of 24, a decrease in both the mean squared error and the mean absolute error could be observed and measured. Using a humidity prediction model and the appropriate type of filter can help improve predictions.

Furthermore, it was observed that including the month as a characteristic in the training of the models contributed significantly to improving their performance, which indicates the importance of considering seasonal variability in the estimation of soil moisture.

The data processing and analysis performed in this study also provided important insights into the availability of information and the feasibility of machine learning techniques in agricultural settings where sensor data may be limited or expensive to acquire.

Another contribution of this work is the presentation and processing of the data set created by C. K. Gasch et al. This data set has been transformed into a more accessible format, facilitating its use and analysis for future soil moisture estimation and precision agriculture research. By identifying and addressing missing data using nearest-neighbor interpolation techniques and enriching the data set with additional meteorological information, we have created a solid foundation for more detailed and comprehensive analyses.

References

A, Y., Wang, G., Hu, P., Lai, X., Xue, B., & Fang, Q. (2022). Root-zone soil moisture estimation based on remote sensing data and deep learning. *Environmental Research*, *212*, 113278. https://doi.org/10.1016/j.envres.2022.113278.

Ahmad, S., Kalra, A., & Stephen, H. (2010). Estimating soil moisture using remote sensing data: A machine learning approach. *Advances in Water Resources*, *33*(1), 69–80. https://doi.org/10.1016/j.advwatres.2009.10.008

Gajula, A. kumar, Singamsetty, J., Dodda, V. C., & Kuruguntla, L. (2021). Prediction of crop and yield in agriculture using machine learning technique. *2021 12th International Conference on Computing Communication and*

Networking Technologies (ICCCNT), 1–5. https://doi.org/10.1109/ICCCNT51525.2021.9579843

Gasch, C., & Brown, D. (2017). *Data from: A field-scale sensor network data set for monitoring and modeling the spatial and temporal variation of soil moisture in a dryland agricultural field*. Ag Data Commons. https://doi.org/https://doi.org/10.15482/USDA.ADC/1349683

Gasch, C. K., Brown, D. J., Campbell, C. S., Cobos, D. R., Brooks, E. S., Chahal, M., & Poggio, M. (2017). A Field-Scale Sensor Network Data Set for Monitoring and Modeling the Spatial and Temporal Variation of Soil Water Content in a Dryland Agricultural Field. *Water Resources Research*, *53*(12), 10878–10887. https://doi.org/10.1002/2017WR021307

Kashyap, B., & Kumar, R. (2021). Sensing Methodologies in Agriculture for Soil Moisture and Nutrient Monitoring. *IEEE Access*, *9*, 14095–14121. https://doi.org/10.1109/ACCESS.2021.3052478

Khalyasmaa, A., Eroshenko, S. A., Chakravarthy, T. P., Gasi, V. G., Bollu, S. K. Y., Caire, R., Atluri, S. K. R., & Karrolla, S. (2019). Prediction of Solar Power Generation Based on Random Forest Regressor Model. *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, 0780–0785. https://doi.org/10.1109/SIBIRCON48586.2019.8958063

Lee, C. suk, Sohn, E., Park, J. D., & Jang, J.-D. (2019). Estimation of soil moisture using deep learning based on satellite data: a case study of South Korea. *GIScience & Remote Sensing*, *56*(1), 43–67. https://doi.org/10.1080/15481603.2018.1489943

Meteostat. (n.d.). Retrieved April 11, 2024, from https://meteostat.net/es/station/KPUW0?t=2007-04-20/2007-04-27

Palominos-Rizzo, T., Villatoro-Sánchez, M., Alvarado-Hernández, A., Cortés-Granados, V., & Paguada-Pérez, D. (2022). Estimación de la humedad del suelo mediante regresiones lineales múltiples en Llano Brenes, Costa Rica. *Agronomía Mesoamericana*, 47872. https://doi.org/10.15517/am.v33i2.47872

Ponraj, A. S., & Vigneswaran, T. (2020). Daily evapotranspiration prediction using gradient boost regression model for irrigation planning. *The Journal of Supercomputing*, *76*(8), 5732–5744. https://doi.org/10.1007/s11227-019-02965-9

Prasad, R., Deo, R. C., Li, Y., & Maraseni, T. (2018). Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. *Geoderma*, *330*, 136–161. https://doi.org/10.1016/j.geoderma.2018.05.035

Rasheed, M. W., Tang, J., Sarwar, A., Shah, S., Saddique, N., Khan, M. U., Imran Khan, M., Nawaz, S., Shamshiri, R. R., Aziz, M., & Sultan, M. (2022). Soil Moisture Measuring Techniques and Factors Affecting the Moisture Dynamics: A Comprehensive Review. *Sustainability*, *14*(18), 11538. https://doi.org/10.3390/su141811538