POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Machine Learning Techniques For The Estimation Of Soil Moisture From Satellite Data

Master's Thesis in Space Engineering

Author: Marco Varalla

Advisor: Prof. Claudio Maria Prati

Co-advisor: Dr. Alfonso Amendola (Eni S.p.A.)

Co-advisor: Dr. Simone Sala (Eni S.p.A.)

Academic year: 2022-2023

1. Introduction

Soil moisture plays a crucial role as it quantifies the amount of water present within the soil matrix. It is a fundamental aspect in Earth's hydrological cycle and has wide-ranging applications, including agriculture, weather forecasting, and climate modeling.

Historically, the challenge of obtaining precise soil moisture measurements across various terrains, especially in remote or inaccessible areas, has been a significant hurdle.

Currently, there are approximately 71 International Soil Moisture Networks (ISMN) worldwide, with over 2800 operational stations. These stations offer nearly real-time soil moisture measurements at specific locations. However, the distribution of these stations globally is uneven, leading to data gaps, especially in regions with limited or sparse coverage of these measurement stations.

To address these challenges, researchers have suggested the use of satellite imagery to estimate soil moisture on regional and global scales, aiming to overcome distribution limitations. Advanced satellite technologies, including integrated microwave radiometers and radar systems, now make it possible to assess soil moisture non-invasive.

In soil moisture analysis, machine learning algorithms hold promise for addressing significant challenges. Their capacity to handle non-linear and multi-dimensional data equips them well for modeling the complex interactions among variables that affect soil moisture levels. By training on historical datasets from the ISMN hub and high-resolution satellite observations from the Sentinel Program, machine learning has the potential to develop a nuanced understanding of the intricate correlation between remote sensing data and actual soil moisture conditions.

This study's primary goal is to explore and harness the potential of artificial intelligence in advancing soil moisture estimation using satellite-derived data. Additionally, this investigation will delve into the efficacy of machine learning in processing and integrating heterogeneous satellite data sources, with the ultimate aspiration of refining the comprehension of soil moisture dynamics.

2. Study Area

The research was conducted in the region covered by the Texas Soil Observation Network (TxSON), which encompasses 40 monitoring stations in a 1500 km^2 grid near Fredericksburg, TX, between the Pedernales and Colorado rivers. These stations monitor various site characteristics, including weather and soil conditions.

The study area experiences a semi-arid climate with an average annual rainfall of approximately 30 inches. Summers are hot, often exceeding 32°C, while winters are relatively mild, averaging around 15°C. The landscape consists of rolling hills, rocky terrains, and intermittent river valleys, leading to diverse ecosystems and vegetation patterns.

The selection of TxSON as the study site was based on an analysis of various networks within the International Soil Moisture Network, considering factors like the number of satellite observations and station-to-area ratio. After evaluation, TxSON was chosen due to its homogeneous characterization of the territory. Texas' sparsely vegetated and arid lands make it a promising location for gaining insights through the integration of in-situ and satellite data.

Some stations have been organized into groups, listed in Table 1 and visible in figure 1.

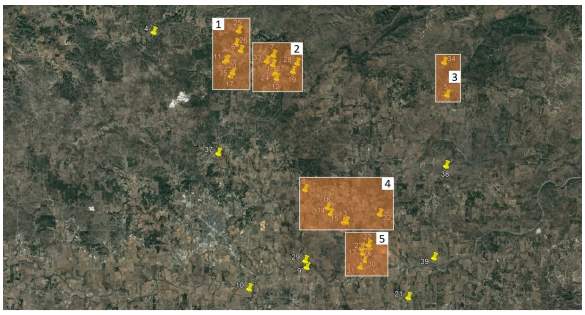


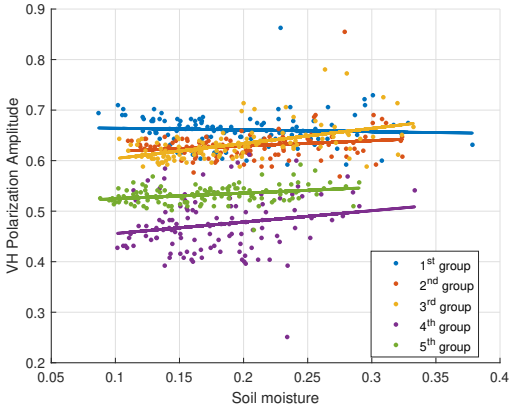
Figure 1: TxSON's Google Maps View

GROUP	STATION
1	6 - 11 - 17 - 20 - 25 - 26 - 35
2	1 - 7 - 12 - 19 - 24 - 27 - 28 - 33
3	5 - 34
4	2 - 13 - 16 - 22 - 23 - 30 - 31
5	3 - 8 - 14 - 15 - 18 - 32

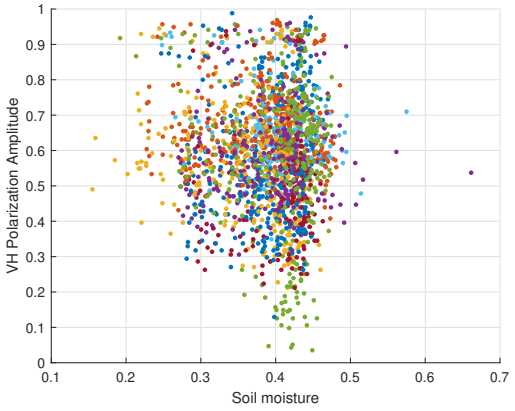
Table 1: Groups

Stations that are not included in these groups have not been taken into consideration. The identified sub-regions for the study are approximately 25 square kilometers in size. Within these sub-regions, significant variations are not anticipated in terms of soil moisture levels, as well as in the magnitudes of SAR images in VH and VV polarizations, and NDVI.

The decision to select the TxSON network as the study site was also guided by the VH-SM and VV-SM correlations observed within each network. Figures 2 and 3 illustrate a comparison between the WegenerNet and TxSON networks. Notably, these figures reveal distinct correlations within the TxSON network, indicating the possibility of positive results. In contrast, the WegenerNet network presents a more complex situation.

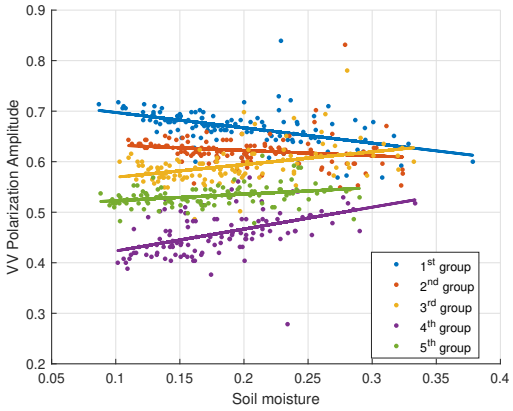


(a) TxSON Network

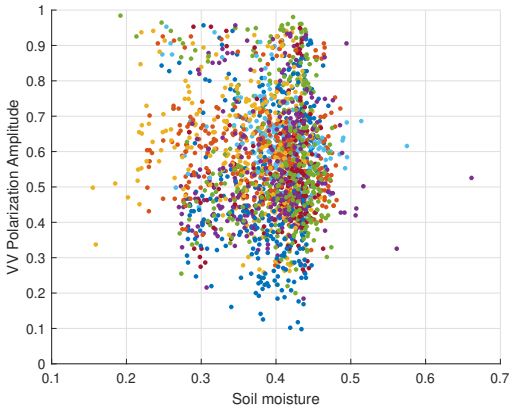


(b) WegenerNet Network

Figure 2: VH Amplitude vs Soil Moisture



(a) TxSON Network



(b) WegenerNet Network

Figure 3: VV Amplitude vs Soil Moisture

3. Dataset

The study utilized VV and VH dual polarization GRD (level-1) radar images to gather comprehensive surface information and interactions, extracted from Sentinel-1. VH polarization involves transmitting radar waves vertically and receiving them horizontally, while VV polarization involves both transmission and reception in a vertical orientation.
Due to data availability limitations for the selected region, only data from Sentinel-1A were accessible, except for a brief time window in June 2019. Consequently, it is decided to utilize Sentinel-1A data from January 1, 2018, to December 31, 2021. This approach resulted in the collection of 115 observations of the region, with an approximate frequency of one observation every 12 days, aligning with the specified revisit time, all of which were associated with relative orbit 107.
Regarding Sentinel-2, only two specific bands were utilized: band-4 (Red; 665 nm) and band-8 (NIR; 865 nm). These bands were employed to calculate the Normalized Difference Vegetation Index (NDVI) using:

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{1}$$

As the name suggests, NDVI is a normalized difference metric used to determine the presence of live vegetation in the observed area. Its range goes from -1 to 1: negative values indicate water, values near zero suggest arid areas, and values near 1 represent lush vegetation.
Using the ISMN, time series data for each of the 40 stations within the TxSON network were obtained. To standardize the data, daily averages were calculated from hourly time series, aggregating 24 data points.
This process was applied to each of the five station groups, ensuring consistency and facilitating comparisons with satellite observations.

4. ML Algorithms

The primary goal of the research is to determine the optimal category and structure of machine learning models. The main emphasis is on optimizing and minimizing the root mean square error (RMSE), a crucial metric computed using a specific formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2}$$

Where:

- y_i represents the observed value
- $\hat{y}_i$ represents the predicted value
- n is the total number of data points

Lower RMSE signifies a dependable and precise model, enhancing its versatility for various applications and tasks. To reach a lower value of RMSE, an optimization of the hyperparameters of each model is needed.
Below, the examined models are listed. For each of them, the corresponding MATLAB function performed hyperparameter optimization to minimize the RMSE.

- Linear Regression
- Support Vector Machine
- Random Forest
- Ensemble of Learners

- ◇ Multi-Layer Perceptron
- ◇ Gaussian Process Regression
- ◇ Gaussian Kernel Regression

5. Methodology

5.1. Pre-processing

Satellite data from both Sentinel-1 and Sentinel-2 are processed through distinct workflows, utilizing the ESA SNAP (Sentinel Application Platform) software.
The steps followed for Sentinel-1 are listed below:

- ◇ Apply orbit file
- ◇ Calibration
- ◇ Speckle Filtering
- ◇ Terrain Flattening
- ◇ Terrain Correction
- ◇ Subset
- ◇ Conversion

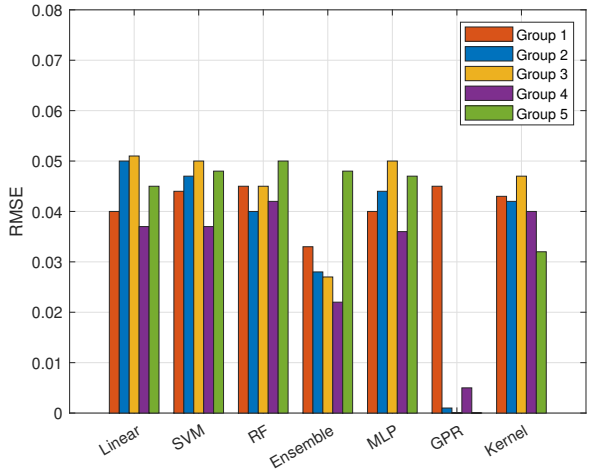
Regarding Sentinel-2 data, the process involved taking Band 4 (red) and Band 8 (near-infrared) into account to calculate the NDVI index (1). Afterward, a "Subset" operation was performed and then the conversion.
Sentinel-1A revisits the same area every 12 days, while the combined Sentinel-2 constellation (Sentinel-2A and Sentinel-2B) revisits every 5 days, although some observations may be discarded due to cloud cover.
This time offset between the two satellite missions is noticeable.
To address this challenge, for each of the 115 Sentinel-1 images, the temporally closest available Sentinel-2 image have been linked.
In this way, 115 images for each feature are available. Within each defined region, VH and VV polarization images, along with NDVI data, are extracted by cropping the primary images into smaller segments using precise station coordinates. Subsequently, median value calculations are performed on these three images. This strategic approach helps mitigate the potential influence of water bodies or urban areas.
Concluding the process, a database represented as a 115x4 matrix for each group is assembled. In this matrix, the first column corresponds to VH, the second to VV, the third to NDVI, and the fourth to soil moisture.

5.2. ML Phase

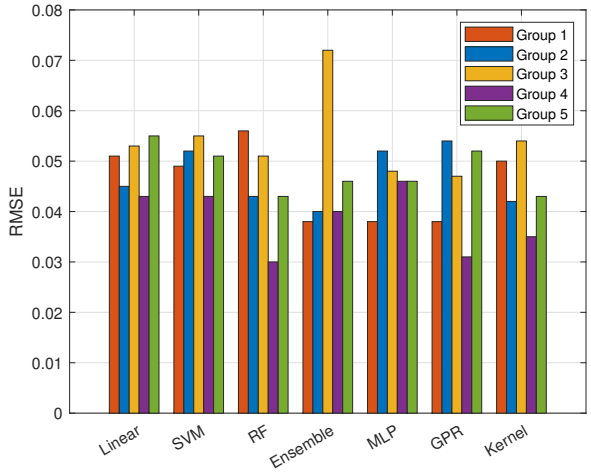
Each of the five databases was divided into separate 'training' and 'inference' subsets, with a distribution of 75% for the first phase and 25% for the second phase. Before this partitioning, the database rows were shuffled to introduce greater randomness. The following step entails the utilization of MATLAB functions, enumerated in Section 4.
After establishing the model hyperparameters, the training database is divided into four blocks. Using a cyclic approach, one of these blocks is systematically excluded while the model is retrained (with unchanged hyperparameters) on the remaining three blocks. This trained model is then tested against the omitted block. By comparing the model's predictions to the actual values, the Root Mean Square Error (RMSE) is calculated, serving as a crucial discriminating metric.
Subsequently, the inference phase occurs, where the algorithm is tested on the portion of data it has not encountered before, comprising the remaining 25%. By comparing the algorithm's predictions with actual data, the Root Mean Square Error (RMSE) value can be computed for each site and configuration.
Improved performance is indicated by reduced RMSE values in this phase. Therefore, determining the optimal architecture is synonymous with identifying the one that exhibits the least difference between predictions and actual values.

6. Results

In conclusion, each of the specified categories has an associated Root Mean Square Error (RMSE) value for both the training and inference phases. In Figure 4, RMSE indices for both the training and inference phases of the models, for each model, and for each site, treated as if they were completely different networks, can be observed.



(a) Training Phase



(b) Inference Phase

Figure 4: ML Models RMSEs

The superiority of the result obtained from the GPR architecture, particularly for site 4, is clearly evident, even when represented graphically. Table 2 displays the results, highlighting the group with the most minimal RMSE, relative at the inference phase, for each configuration.

ML Algorithm	Group	RMSE
Linear	4	0.043
SVM	4	0.043
Random Forest	4	0.030
Ensemble	1	0.038
Multi-Layers	1	0.038
GPR	4	0.031
Kernel	4	0.035

Table 2: Minimal RMSEs

Table 2 highlights the clear advantage of groups 1 and 4.

Nevertheless, it remains crucial to emphasize that attaining these values relies on training and inference architectures with site-specific data. Employing an architecture trained on one area and tested on another fails to produce satisfactory outcomes. In order to predict the soil moisture of a given region, it is necessary to have access to historical data for that region. However, it cannot be assumed that every region exhibits a correlation, as is evident in the cases of groups 3 and 5 of this study.
Therefore, the creation of software that can be used universally in every part of the globe is not possible.

7. Conclusions

In conclusion, the fundamental objective of this research was to analyze the intricate dataset procured from the Sentinel Program's satellites and to create a globally applicable, universally adaptable tool.
It's apparent that these values, while moderately satisfactory, display inherent variability only among individual sites. This underscores the infeasibility of training a model in one location to predict soil moisture in another, different one.
By observing the graphs in Figure 4, it becomes evident that the Random Forest and MLP models are unsuitable for this purpose, most likely due to the limited amount of available data. Conversely, the GPR model, which is generally the most suitable for low data volumes, reaffirms its superiority.
It is evident that the use of machine learning algorithms can produce moderately satisfactory results in the context of soil moisture data training and prediction. However, these achievements are notably limited to a 'local' context, a factor not documented in any existing literature.
Currently, the aspiration of training a single model applicable across diverse global locations remains unattainable, diverging from the initial overarching goal of this endeavor.
Improvements could potentially be achieved by introducing additional variables, such as the satellite incidence angle, or by utilizing in-situ data specifically designed for this purpose. Alternatively, exploring more advanced ML algorithms remains an option. However, it is impor-

tant to emphasize that the focus should always be on achieving a 'local' prediction for a region with existing historical data.
While this may not have been the initial objective of the work, it is nonetheless satisfying to know that it is possible to predict soil moisture in specific areas with precise characteristics for which historical data are available.

Contents

List of Figures

List of Tables