ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 6, No. 1, 636-650 (2021)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Comparison of Machine Learning Parametric and Non-Parametric Techniques for Determining Soil Moisture: Case Study at Las Palmas Andean Basin

Carlos López-Bermeo*, Mauricio González-Palacio, Lina Sepúlveda-Cano, Rubén Montoya-Ramírez, César Hidalgo-Montoya

Universidad de Medellín, Facultad de Ingenierías, Medellín, 050026, Colombia

ARTICLE INFO

Article history:
Received: 07 October, 2020
Accepted: 09 January, 2021
Online: 05 February, 2021

Keywords:
Soil moisture
Machine Learning
Regression

ABSTRACT

Soil moisture is one of the most important variables to monitor in agriculture. Its analysis gives insights about strategies to utilize better a particular area regarding its use, i.e., pasture for cows (or similar), production forests, or even to answer what crops should be planted. The vertical structure of the soil moisture plays an important role in several physical processes such as vegetation growth, infiltration process, soil – atmosphere interactions, among others. Despite a set of tools are currently being evaluated and used to monitor soil moisture, including satellite images and in-situ sensor, several drawbacks are still persisting. In situ data is expensive for high spatial monitoring and vertical measurements and satellite data have low spatial resolution and only retrieval information of soil moisture for the top few centimeters of the soil. The present work shows an experiment design for collecting soil moisture data in a specific Andean basin with in-situ sensors in different kinds of soils as a promising tool for reproducing soil moisture profiles in areas with scarce information, employing only surface soil moisture and simple soil characteristics. Collected data is used to train machine learning supervised parametric (Multiple Linear Regression - MLR) and non-parametric models (Artificial Neural Networks - ANNs and Support Vector Regression - SVR) for soil moisture estimation in different depths. Conclusions show that parametric methods do not meet goodness of fit assumptions; so, non-parametric methods must be considered, and SVR outperforms parametric methods regarding regression accuracy allowing to reproduce the soil moisture content profiles. The proposed SVR model represents a high potential tool to replicate the soil moisture profiles using only surface information from remote sensing or in-situ data.

1. Introduction and problem statement

This paper is an extension of work originally presented in *CISTI 2020* [1]. Soil moisture is one of the most critical variables to be monitored in soils [2]. Understanding the behavior of the soil moisture makes it possible to determine the soil use [3], i.e., if it is suitable for cropping, for animal's pastures [4], and even, as a critical variable to understand if a particular terrain may be considered to real estate projects [5]. Another important application of soil moisture analysis belongs to risk management, such as landslides prevention and prediction [6]. For hydrologic applications, the soil moisture represents one of the most important variables controlling the interactions between the atmosphere and land through the evapotranspiration and evaporation processes. Additionally, the soil moisture represents a key variable for the infiltration process and direct surface runoff production in hydrological models.

In that way, different knowledge disciplines, such as hydrology [7], environmental management [8], geology [9], topography [10], among others, deal with soil moisture to deeply study its effects in plenty of different applications. Nonetheless, getting information about soil moisture is commonly a difficult task: available sources are scarce and are associated with satellite images, which causes low resolution regarding spatial distribution [11], that is, soil moisture is measured in broad areas and cannot be explicitly determined in points. During the last decades satellite images have emerged as a powerful tool to retrieve information about soil moisture in the superficial soil surface, even though data in other depth horizons is mandatory to model the behavior in a lot of applications.

*Corresponding Author: Carlos López-Bermeo, celopez@udem.edu.co

Given the importance and drawbacks mentioned above, several techniques based on Artificial Intelligence (AI) theory are currently employed to estimate and predict the soil moisture for several engineering and scientific applications. Knowledge-based systems that apply AI are considered as easy and user-friendly tools [12, 13], as they have advantages in terms of neither requiring a pre-defined conceptual relationship between the input-target parameters nor requiring expensive experimental and field measurement apparatus [14].

In this sense, several studies have been proposed to determine the best Machine Learning (ML) or (AI) model to estimate the soil moisture, considering different variables like meteorological data and satellite images, among others. There are two main ML models used: Support Vector Regression (SVR) and Artificial Neural Networks (ANN) in different configurations. It is the case of [15], three approaches are proposed to generate the SVR regression model for soil moisture estimation: in the first approach, the authors use the soil moisture and meteorological data (air temperature, relative humidity, average solar radiation, and soil temperature at five and 10cm) at time step $t-1$ and t for predict soil moisture at $t+4$ and $t+7$, where t is in days. For the second approach, they use only the meteorological data, while in the third approach, they use only soil moisture at the same time step. The results of this study show that the SVR models performed better forecasting than a simple ANN model. In the same line, [16] proposes the use of SVR for soil moisture prediction using remote sensing data coming from 10 sites in the Lower Colorado River Basin (US). The data includes backscatter and incidence angle from Tropical Rainfall Measuring Mission (TRMM), and Normalized Difference Vegetation Index (NDVI) from Advanced Very High-Resolution Radiometer (AVHRR). The authors trained an SVR with five years of data (time series) and tested on three years of data. The results show a Root Mean Square Error (RMSE) less than 2%. Additionally, results are compared with an ANN and a Multivariate Linear Regression Model (MLR), showing that the SVR has a better performance than the other models.

The latest studies show a tendency for the use of deeper ANN like shows [17] in their review. The authors classify the use of ANN in three categories according to the types of training data. For the first category, the ANN is trained with model-generated data. For the second category case, the ANN is trained with in-situ measurements, with the restriction that the spatial scale could mismatch between point-scale measurements. Finally, the third category encloses the ANN trained with global Land Surface Model (LSM) simulations for soil moisture estimation at large scales.

Other useful applications using ANN are related to the infilling missing soil moisture records such as the research presented in [18], who were using five statistical methods, and nine ANN categorized into Feedforward, Dynamic, and Radial Basis Network methods estimate missing soil moisture records. The obtained values were validated against known values for 13 soil moisture monitoring stations for three different soil layer depths in the Yanco region in southeast Australia. The results show that the nonlinear autoregressive neural network performs in similar quality than other typical methods such as the rough sets method, and monthly replacement. Other studies employing ANN for hydrologic variables are [19–22].

Although SVR and ANN are the most used techniques, several works can be found related to soil moisture prediction that use other techniques of AI and ML. Some authors include multiple techniques, like, where Classification and Regression Trees (CART), Boosted Regression Trees (BRT), Random Forest (RF), Multivariate Adaptive Regression Splines (MARS), and Flexible Discriminant Analysis (FDA), are tested, all with promising results. In this sence, [14] using several input variables such as dielectric constant, soil bulk density, clay content, and organic matter for 1155 soil samples, proposed a hybrid Adaptative Neuro-Fuzzy Inference System (ANFIS) - Grey Wolf Optimization (GWO) intelligent model for simulating soil moisture content. The results based on several statistical parameters verified the feasibility of these kinds of models improving accuracy by around 50% when compared with other similar models as ANN, SVR, and standalone ANFIS models.

Despite the increasing usage of Artificial Intelligence (AI) theory tools related to soil moisture estimation, most of the studies have the same thing in common; none of them is focused on the estimate of the soil moisture at different depths. The data estimate corresponds to a depth between 5 and 10 cm approximately or the root zone. Few studies such as [23], employed surface soil moisture (0–5 cm) values and Hydrological Soil Groups (HSGs) information to perform a Statistical Soil Moisture Profile (SSMP) model to transfer the spatial variations of soil moisture profile with the change in soil hydraulic properties. The proposed model was based on correlation techniques with preceding time steps of soil moisture fields. In another study, [24] employed multispectral imagery from Unmanned Aerial Vehicle (UAVs) and a method based on the combination of an evolutionary algorithm and artificial intelligence called genetic programming (GP), to propose a methodology to simulate soil moisture at different levels. The results were compared to ANN and SVR methods.

With this idea in mind, the present work shows a methodology to fit some ML models with in-situ sensor data, which transmit real-time information by using IoT tools and methods to overcome limitations associated with the number of sensors to determine soil moisture at different depths using surface topsoil moisture estimates. The proposed methodology represents a high potential tool for reproducing the soil moisture profiles using only surface information from satellite or in-situ data for scarce information areas. Specifically, we use the most known statistics-based parametric method: Multiple Linear Regression (MLR) and perform the corresponding Analysis of Variance (ANOVA) [25] as well as the Goodness of Fit metrics (normality, independence, homoskedasticity). We also use two widespread ML non-parametric methods: SVR and ANN [26, 27]. All the models are primarily assessed by a case study, with collected data in the Las Palmas basin, at the Andean region, in Medellín, Colombia.

The rest of this paper is organized as follows: section 2 shows a short theoretical framework; section 3 stands the methodology and experiment design; section 4 shows the models fitting; in

section 5, the analysis over results is carried out. Finally, conclusions, future work, acknowledgments, and references are presented.

2. Theoretical framework

In this section, some key concepts will be defined to understand the work better.

2.1. Soil moisture

A given portion of soil is composed of solid particles and the rest of the voids. A part of the holes is occupied by water and the rest by air. The volume occupied by water is measured using the soil moisture content, which is defined as θ as in equation (1) :

$$\theta = \frac{Volume_{water}}{Volume_{total}} \quad (1)$$

The study and knowledge of soil moisture are essential due to its influence on hydrological processes and energy flows on the earth's surface. It is also critical because of its connection with precipitation, runoff generation, nutrient transport, and groundwater [28–30]. Recent research identifies that a surplus or lack of soil moisture can favor floods or droughts occurrence [31]. Likewise, the feedback of soil moisture on evapotranspiration is essential for temperature variation and the appearance and persistence of heatwaves, as well as for precipitation generation and location [32]. Besides, the role of soil moisture in photosynthesis, ecosystem dynamics, soil respiration, and the terrestrial carbon balance is undeniable.

Therefore, soil moisture variation is critical in hydrological, ecological, and environmental studies [33] and, in particular, as a support for agriculture and biomass production [34]. Combinations of the factors mentioned above cause variations in spatial and temporal soil moisture content, which makes it a significant limiting factor for crop growth. Indeed, low crop yields should be more often related to insufficient soil moisture than insufficient rainfall [35,36], which causes that soil moisture plays a vital role as a critical resource for vegetation growth that supports agricultural production. In this sense, it is identified that soil moisture contributes in a crucial way to understanding the global climate system. For this reason, it has been highlighted by the Global Climate Observing System (GCOS) as one of the "essential climate variables". Therefore, monitoring temporal and spatial soil moisture variability is essential to estimate water availability limits and to quantify its climatic variations sensitivity and human pressures [32].

2.1.1. Soil moisture sensing technologies

Currently, several techniques are used to measure soil moisture, allowing monitor of humidity on a large scale: in-situ sensors (to take humidity measurements at the site where the sensor is located), and remote sensing from towers, aircraft, and satellites, using radiometers in the microwave region, scatterometers, synthetic aperture radars, and radar combinations-radiometers. According to [28], the products obtained with passive and active sensors have different correlation values around the world, when compared with in-situ measurements.

Although there are many possible sources of information, in the case of the tropics and specifically in the study area, there could be limitations for the use of satellite information. Some derive from the possibility or not of having reliable and spatially coherent in situ soil measurements of soil moisture, which makes it challenging to carry out a cohesive evaluation of the accuracy and information content of remote sensing products. Besides, the use of inadequate techniques for downscaling can generate alterations in data quality. Another difficulty could arise due to the impossibility of obtaining quality data with fair spatial and temporal resolution in the study area, primarily due to its geographical location, which may limit some sensors use, especially optical ones. In this sense, in the visual spectrum, one of the significant limitations is the limited surface penetration due to high cloudiness. However, visible and infrared sensors and microwave sensors are not limited by cloud cover and night conditions [35]. Observations can be made at any time of day and are not dependent on sunlight [37].

Soil moisture in-situ sensing, In Colombia and many other tropical regions, there is very little information available on soil moisture taken from in-situ monitoring stations that can be used to understand the behavior of this variable for different applications. Although it is true that in-situ measurements require a costly investment and have low coverage of the territory, they allow the analysis at a local scale (as is the case of a basin). Additionally, they are necessary to validate the data from remote sensors that generally have a low spatial resolution.

The primary function of a soil moisture sensor is to report its current state, by using an electric variable (commonly voltage), to any acquisition system, i.e., datalogger, or IoT end node. The voltage reported is scaled to get the corresponding measurement, which is commonly given in cubic meters of water over cubic meters of soil material (m^3/m^3).

The most used operation principle in soil moisture sensors is the variation of capacitance as the moisture changes. Soil acts as a dielectric material between two plates, and it changes when this variable varies. However, conductivity sensors are also applied to sense moisture. Water content allows flowing electric currents between two electrodes, and it is proportional to soil moisture.

2.2. Parametric and non-parametric methods for data fitting

To fit some models for regression of moisture data and perform the comparison, we have chosen some of the most used methods in the literature. MLR, SVR, and ANN algorithms will be reviewed in this section.

2.2.1. Multiple Linear Regression (MLR)

MLR is one of the most used parametric strategies for data fitting. The general model is shown in equation (2):

$$\hat{Y} = \sum_{i=0}^{n} [\hat{\beta}_i X_i] + \varepsilon \quad (2)$$

Where $\hat{Y}$ is the regression value of the dependent variable, $\hat{\beta}_i$ is a set of n+1 predictors (coefficients) which are determined by using least-squares optimization, X_i are n independent (predicting) variables, and ε is the regression error. If a categorical variable is in the predictors, a set of dummy variables is included, depending

on the number of levels of the factor. The goal, then, is finding the values of each $\hat{\beta}_i$ such that the global error ε is minimal.

Once the predictors are determined, an Analysis of Variance (ANOVA) must be performed [25], to determine if the means of the prediction variables are equal, that is, if all the variables are representative for the analysis.

Finally, to check if the model is statistically valid, some goodness-of-fit tests must be performed: normality, by using the Kolmogorov-Smirnoff test [38]; autocorrelation, by using Durbin-Watson test [37]; and homoskedasticity, by using Breusch Pagan test [39]. Due to the model usage is limited to the goodness of fit, many datasets do not meet this strategy.

2.2.2. Support Vector Regression (SVR) [40, 41]

The goal of this supervised learning strategy is to find one (or more) hyperplanes that separate previously tagged classes. By finding the support vectors, it is possible to predict, in the case of regression, the value of a new continuous variable based on a set of inputs. Figure 1 shows the geometric concept for two input characteristics.

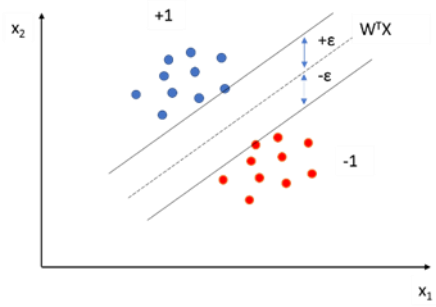


Figure 1: SVR geometric concept

The objective is to find the hyperplane that separates two classes with the maximum margin between them, therefore, finding the model is reduced to an optimization problem, where the coefficients w must be found, as shown in equation (3).

$$\begin{aligned} Min\ Z &= \frac{1}{2}\|W\|^2 \\ St: |y_i - w_i x_i| &\leq \varepsilon \end{aligned} \tag{3}$$

where y_i are the predicted variables, x_i are the independent variables (or features) and ε is the minimum tolerance between the separation hyperplanes and the features. For SVR, a slack variables set is added to equation (3), such that, for each value that falls outside ε, its deviation is recorded to make it minimal. The hyperparameter cost (C) is introduced to penalize said deviations, which is tuned through cross-validation. Thus, equation (3) is re-written, as shown in equation (4) (called the primal problem):

$$\begin{aligned} Min\ Z &= \frac{1}{2}\|W\|^2 \quad + C\sum_{i=0}^{n}|\xi_i| \\ St: |y_i - w_i x_i| &\leq \varepsilon + |\xi_i| \end{aligned} \tag{4}$$

However, in many cases, the classes are not linearly separable, for this reason, it is recommended to use kernel functions, K [41, 42], to perform a transformation that increases the number of dimensions, allowing separating such classes, as shown in equation (5):

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \tag{5}$$

where ϕ is a nonlinear mapping function and $\langle , \rangle$ is the inner product operator.

The dual problem to the primal problem is given by equation (6):

$$\begin{aligned} Min\ & \frac{1}{2}\alpha^T Q\alpha \ - e^T\alpha \\ St: \quad & y^T\alpha = 0 \\ & 0 \leq \ \alpha_i \leq C, i = 1, \dots, n \end{aligned} \tag{6}$$

where e is a vector of all ones, Q is a positive semidefinite matrix with $Q_{ij} = y_i y_j K(x_i, x_j)$ and α_i are the dual coefficients.

Using this model, the prediction function of the equation (7) is obtained:

$$f(x) = \sum_{i=1}^{n/2} y_i \alpha_i K(x_i, x) + b \tag{7}$$

where b is the bias term.

It is important to remark that the standard SVR model in equation (7) considers only real-valued functions [43]. Additionally, it is a discriminant method, i.e. produces a mapping from the data points to the class labels without computing probability distributions, allowing the method to be less computationally expensive, but more sensitive to noise than generative methods [44]. . The main advantage of SVR is that can deal with sparsity, non-linearity, and high dimensionality of the input data [45].

Regarding the SVR hyperparameter tuning (section 4.2), it is performed through cross-validation, and finally, the RMSE that exhibits the best results will be chosen. In the same way that MLR, if there are categorical variables in the set of X, *dummy variables* can be used.

2.2.3. Artificial Neural Networks (ANN). [40, 41]

This bio-inspired strategy is used in a lot of applications, where regression and classification are needed. The structure of a neural network emulates a meta-heuristic concept, based on interactions of neurons within the brain. The most known ANN is known as Multilayer Perceptron, and its architecture is shown in Figure 2. From this figure, a set of m input variables (x) in the Input layer are used to predict the values of n output variables (y) in the output layer. The prediction is achieved by choosing the correct number of neurons in different hidden layers h, which are fully connected with both previous and next layers. The effects of each connection are weighted by a set of constants W, which are determined by an optimization iterative method (commonly Gradient Descend). The main advantages of this kind of ANN (feedforward networks) are that they do not have a priori assumptions about the relationships between the independent and dependent variables [46], their

nonlinear modeling capability [47], and their minimal assumptions needed about the data [48]. In the same way that MLR, if there are categorical variables in the set of *X*, dummy variables can be used.

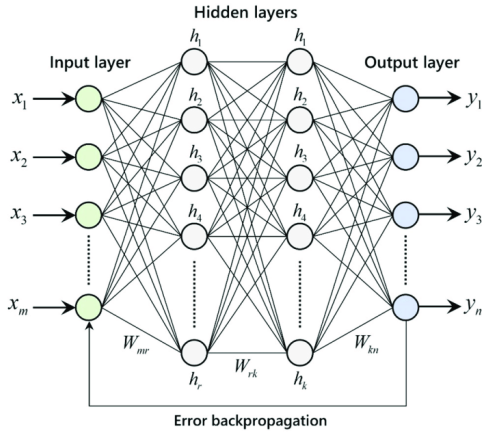


Figure 2: Multilayer Perceptron Architecture with backpropagation feedback [49]

Although using ANN can outperform other methods regarding its accuracy, its application must be extensively checked using cross-validation, since a common phenomenon known as *overfitting* makes that ANN cannot be a general model sometimes.

3. Methodology and experiment design

This section shows the steps to fit the different models. First, a measurement phase is proposed, where both instruments and locations are chosen. Then, with the set of data from in-situ sensors, some models are fitted. Finally, the performance evaluation is carried out, choosing the model which exhibits fewer regression errors.

3.1. Measurement phase

In this section, some considerations regarding the measurement are presented.

3.1.1. Location selection

Soil moisture is highly influenced by climatic characteristics of a region, specifically by rain amount and intensity, evapotranspiration [50], vegetation type, topography [51], soil properties (apparent density, porosity, organic matter content, texture, and structure), among other factors [52].

This is how changes in land use and land cover, in interaction with soil physical characteristics and climate, play a key role in soil moisture variations at different scales [53]. In the Colombian case, the soil moisture variability has been less studied and there is not enough information to understand its spatial and temporal dynamics, and much less has its contribution to water flows regulation been quantified.

This is the case in some areas of the Antioquia region, where precipitation is the only source of soil water and recharge of its surface layers. Therefore, soil moisture in deep layers cannot be replenished with contributions from rain and groundwater. For this reason, soil wet front movement study at different depths is key to understanding its variability and impact on agricultural production and ecosystems in tropical basins.

In different studies [54–56] it has been found that drought conditions affect variations in soil moisture content, while topography and vegetation type are the dominant factors that control the humidity in different soil layers. Specifically, soil water storage is mainly affected by topography at shallow depths, while in deep soil layers it is mainly controlled by vegetation type.

Considering the above, for the present study Las Palmas basin was selected, which has as characteristics of interest that it is in the tropical Andean zone of the municipality of Envigado, Antioquia, Colombia (coordinates: 6 ° 11'26.1 "N 75 ° 31'47.6" W), and which has undergone processes of change in land use and land cover, among which transition from crops and pastures to recreational and commercial uses stands out. There are also some small fragments of forest and secondary vegetation. This basin has an area of 31.31 km^2, with elevations from 2,500 to 2,600 meters above sea level and its predominant climate is cold humid to very humid and with an average annual rainfall of 2781 mm.

The basin also has different soil associations (Figure 3), which determine physical characteristics, as well as different depths in soil profiles [57]. The Tequendamita Association (TE) is found mainly in the Basin, formed by deep to moderately deep soils, well-drained, with medium textures, low to moderate fertility, mild to moderate erosion. A characteristic of these soils is that they are formed by metamorphic rocks (schists, gneisses) covered with volcanic ash. Soil samplings were made, and results were verified with previous studies such as the "General Study of Soils and Zoning of Lands of the Department of Antioquia" [58], and the" Semi-detailed study of soils in zone 13 of the municipality of Envigado for potential use purposes" [59].

Considering that the soil moisture content suitable for plant growth depends on the type of soil [60] and the physical explanation mentioned above related to the main variables affecting the soil moisture variability through the soil layers (vegetation, soil type, topography, among others), such parameters were considered to determine the sites where soil moisture monitoring stations were installed. Therefore, three types of vegetation cover representative of the basin (pastures, crops, and forest) were selected and combined with three different phases of the Tequendamita Association and one phase of the La Ceja Association (see Table 1). It is important to mention that the sensors were located having the characterization of soil profiles in the basin, which allowed determining the average depth of same, in each of the points where the monitoring stations were located. This means that an edaphological criterion was used for the location of each of the sensors. Additional criteria used to guarantee that the data collected were representative are the following: *i)* The points were located near access roads and agricultural areas located toward the higher part of the catchment where installation and security could be achieved with lower cost, *ii)* availability of connection to the telecommunications network to transmit data in real-time, *iii)* accessibility.

3.1.2. IoT end nodes

Soil moisture stations were designed and built using IoT tools and methods. Four stations are installed in the upper part of the basin, where a mobile network operator offers full 2G/3G coverage. At each point, three soil moisture sensors were put into

operation at different depths, depending on land cover (crops, pastures, and forest), and soil physical properties. The first layer, the depths per sensor, according to land cover, are shown in Table 1. Installed sensors are from the brand Meter Group, reference EC-5, measurement range from 0 up to 100%, with a resolution of 0.001 m^3/m^3 of Volumetric Water Content (VWC) and accuracy of ±0.02 m^3/m^3. Sensor output is voltage, from 0 up to 2 Vdc, and is collected by a microcontroller unit ESP32, with an Analog-to-Digital Converter of 12-bit resolution. To ensure that the measurements are stable to be transmitted, the next conditioning protocol is used: *i)* take *n* measures from the sensor, *ii)* arrange samples from the lowest to highest, *ii)* extract the samples in the positions *n/2* and *1+n/2,* and *iii)* take the mean. Engineering units are scaled from ADC units to VWC by using a two-point scheme, by comparing readings of moisture in both water and air and using a Meter Group EM-50 as a calibration pattern. In-situ end-nodes are depicted in Figure 4.

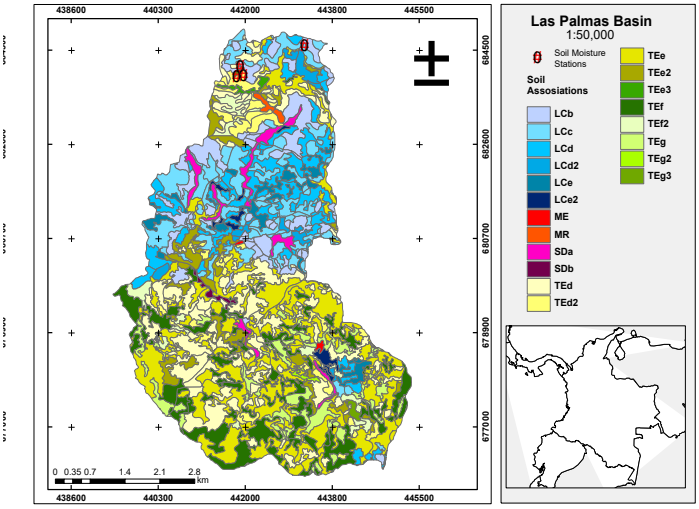
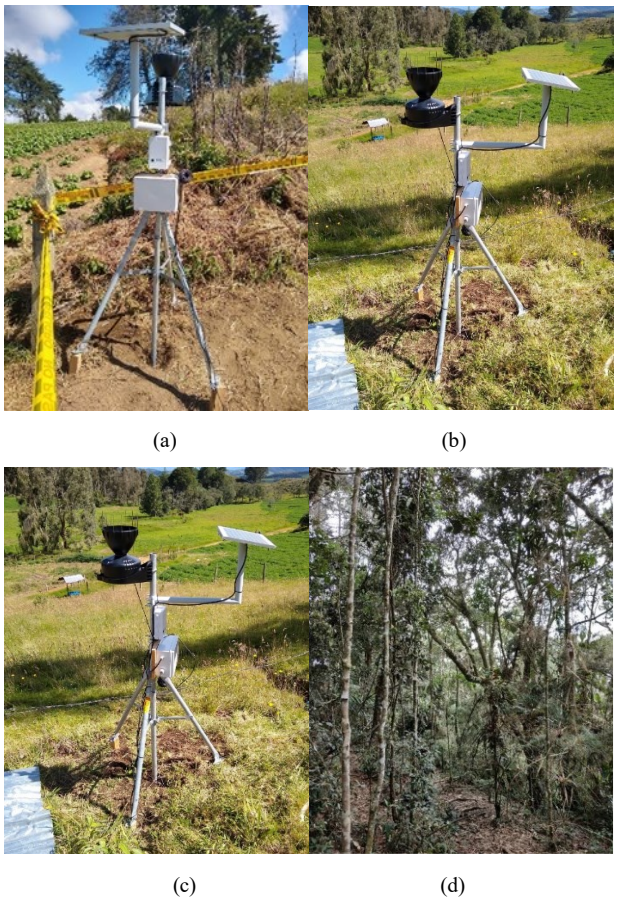


Figure 3: Study area location in Antioquia and Colombia and soil associations present in the Las Palmas basin

After collecting measurements, data is sent to an IoT backend by using the protocol Message Queue Telemetry Transport (MQTT). The chosen IoT platform is Ubidots [61], and it also helps to store information and to provide a frontend user interface to visualize real-time indicators, as shown in Figure 5.

Table 1: Soil moisture sensor depths according to soil use

Land Cover	Soil association	Sensor 1 depth range (cm)	Sensor 2 depth range (cm)	Sensor 3 depth range (cm)
Crop	La Ceja Consociation (LCb)	0-23	23-37	>37
Pasture 1	Tequendamita Association (Ted2)	0-22	22-33	>33
Pasture 2	Tequendamita Association (Ted3)	0-33	33-60	>60
Forest	Tequendamita Association (Ted)	0-16	16-35	>35



(a) (b)

(c) (d)

Figure 4: IoT soil moisture end nodes in (a) crops, (b) pasture type 1, (c) pasture type 2 and (d) forest

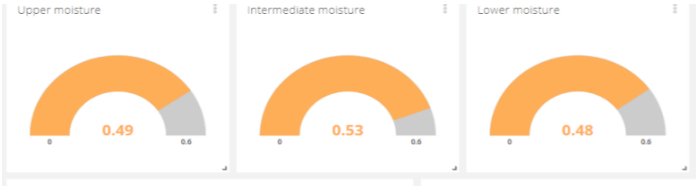


Figure 5: IoT backend / frontend platform [61]

3.2. *Model fitting phase*

First, all data is extracted from the Ubidots platform, using a temporal window from December of 2019 to June 2020, with an equi-temporal sampling rate of 15 minutes. All data is put together into a unique database table, including timestamp, moisture values per sensor, depths per sensor, and soil type. The models are fitted to output a moisture value at a particular depth, regarding the readings of the other two sensors, their corresponding depths, and the soil use.

To fit all the models, R Studio V 1.2.5019 was used. All the basic statistics measurements were inspected: quartiles, means, medians, and standard deviations (for continuous data), and counters (for categorical data). Regression expressions were adapted with *dummy variables* to allow that the categorical variable *soil type* could involve all the factors with different coefficients. Then, a preprocessing phase for continuous variables was carried out with two different strategies, according to the equations (8) and (9):

$$x_i' = \frac{x_i - \bar{x}}{s^2} \quad (8)$$

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (9)$$

where x_i' is the i^{th} transformed sample, x_i is the i^{th} sample of the continuous variable x from the dataset, $\bar{x}$ is the mean of x, s^2 is the standard deviation of x, $min(x)$ is the minimum value of x, and $max(x)$ is the maximum value of x.

Finally, the model fitting for MLR, SVN, and ANN was performed according to the corresponding method. The database has 77848 observations and is split into two sets, with a pseudo-random strategy: training (70%, that is, 54493 examples) and test (30%, that is, 23355 examples). Moreover, some additional tests were run for MLR, to ensure goodness of fit: Kolmogorov-Smirnov [38], Durbin-Watson [37], and Breusch Pagan [39].

3.3. Results analysis phase

After the models were fitted, the regression was carried out with each model, by using the test set. An initial graphical inspection was performed, and three performance metrics are used: RMSE (Root Mean Square Error), Index of Agreement (Willmott), and R^2. These metrics and the graphical approximation served to determine the accuracy of the models regarding field measurements. The pertinent analysis was developed.

4. Model fitting

For the sake of clearness, we will adopt the following notation: M_i is the *ith* sensor, D_i is the depth of *ith* sensor, i is an integer index to identify the sensor position ($i = 1$ is the superficial sensor, $i = 2$ is the middle sensor, and $i = 3$ is the deepest sensor), and P_1, P_2, and F are dichotomic *dummy variables* for representing the factors *pasture type 1*, *pasture type 2* and *forest*, in the categorical variable *soil type*. All the processes are executed in a laptop Lenovo G40, with a processor Intel Core i7 4500 @ 1.8 GHz and RAM = 12 GB.

A previous graphical inspection is performed to qualitatively behold if there are relationships among different moistures, as shown in Figure 6. Some statistical metrics are presented in Table 2.

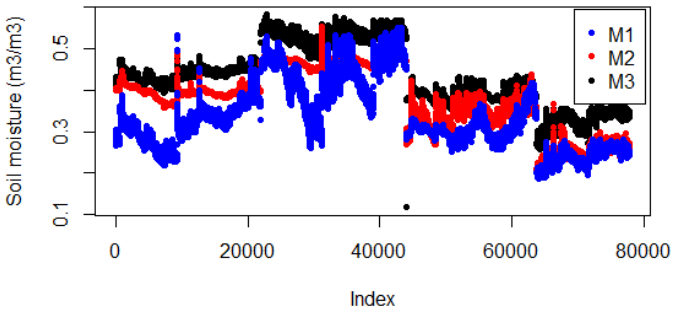


Figure 6. Trends of moisture data. M_1 is the superficial sensor, M_2 is the middle sensor, and M_3 is the deepest sensor. Index interval from 0 up to 21.000 corresponds to crops, from 21.000 to 44.000 corresponds to pasture 1, from 44.000 up to 64.000 corresponds up to pasture 2, and from 64.000 up to 79.000 corresponds to the forest.

From Figure 6, it can be noticed that effectively there is a relationship between the moistures at different depths since when an external variable, i.e., rainfall, causes a perturbation, each moisture varies. Moreover, according to Table 2, medians and means of moistures are arranged from M_1 to M_3 from the least to the greater, indicating that the deep the soil is, the high VWC is. It also can be noticed that the standard deviation of M_3 is the least, supporting that more superficial moisture changes in a wider range than the others, so external variables (rainfall or solar radiation) affect the most to the moistures closer to the soil surface. To validate the hypothesis of correlation among moistures, by using the Pearson correlation coefficient, the next results are obtained: $cor(M_1,M_2) = 0.85$; $cor(M_1,M_3) = 0.91$; $cor(M_2,M_3) = 0.97$.

Table 2: Statistical metrics for soil moistures

	M_1	D_1	M_2	D_2	M_3	D_3
Min	0.047	0.08	0.1918	0.255	0.117	0.33
Q1	0.27	0.11	0.332	0.275	0.382	0.33
Median	0.312	0.115	0.388	0.3	0.425	0.37
Mean	0.3308	0.1215	0.3772	0.3321	0.4322	0.4212
Q3	0.383	0.165	0.457	0.465	0.508	0.6
Max	0.552	0.165	0.553	0.465	0.582	0.6
SD	0.0786	0.029	0.0786	0.085	0.0735	0.113

4.1. MLR

Taking into account that there is a direct correlation among soil moistures, MLR can be fitted according to the process in section 2.2.1. So, the general model in equation (2) is re-written to the particular dataset, as shown in equation (10):

$$\begin{aligned} \widehat{M}_2 &= \hat{\beta}_{0,2} + \hat{\beta}_{1,2}M_1 + \hat{\beta}_{2,2}D_1 + \hat{\beta}_{3,2}D_2 + \hat{\beta}_{4,2}P_1 \\ &\quad + \hat{\beta}_{5,2}P_2 + \hat{\beta}_{6,2}F + \varepsilon_2 \\ \widehat{M}_3 &= \hat{\beta}_{0,3} + \hat{\beta}_{1,3}M_1 + \hat{\beta}_{2,3}D_1 + \hat{\beta}_{3,3}D_3 + \hat{\beta}_{4,3}P_1 \\ &\quad + \hat{\beta}_{5,3}P_2 + \hat{\beta}_{6,3}F + \varepsilon_3 \end{aligned} \quad (10)$$

where all the $\hat{\beta}$ are the model estimators, and ε is the error model. Note that, in the first instance, it is needed to fit two models for M_2 and M_3, depending on the variables associated with the superficial moisture and the *soil type*. However, due to the high correlation between M_2 and M_3, it is essential to find the optimal $\hat{\beta}$ values, the strategy of minimization of *least squares* is applied [62], searching for a hyperplane that minimizes the cumulative errors from measurements and the predicted variable. The R function *lm* is used, getting the results in Table 3, for training data without preprocessing, and with the preprocessing strategies in the equations (8) and (9). From this table, the dichotomic coefficients $\hat{\beta}_{4,2}, \hat{\beta}_{5,2}, \hat{\beta}_{4,3}$, and $\hat{\beta}_{5,3}$ are not determined, due to there are a multicollinearity effect with the other variables; that is, such variables explain the effects thoroughly without needing the inclusion of the variable *soil type* if it is a crop, pasture 1, or pasture 2. The multicollinearity effect is tested by using the *alias* method proposed in [63]. This method shows how a collinear variable is explained by the others, as shown in Table 4.

According to Table 4, for example, the coefficient $\hat{\beta}_{5,2}$ (which corresponds to the binary variable P_2, pasture type 2) is explained by the others, being D_1 (superficial moisture sensor depth), the variable that explains the most the behavior of P_2. A similar analysis can be performed for other variables.

Table 3: Regressor values for the model in equation (8). P-values less than 0.05 indicate that the regressor is significative

Raw data			
Coefficient	Estimate	Std. Error	p-value
$\hat{\beta}_{0,2}$	0.6537	0.0019	<2 x 10^{-16}
$\hat{\beta}_{1,2}$	0.2615	0.1395	<2 x 10^{-16}
$\hat{\beta}_{2,2}$	-17.6344	0.0691	<2 x 10^{-16}
$\hat{\beta}_{3,2}$	5.6271	0.0203	<2 x 10^{-16}
$\hat{\beta}_{4,2}$	*N/A*	*N/A*	*N/A*
$\hat{\beta}_{5,2}$	*N/A*	*N/A*	*N/A*
$\hat{\beta}_{6,2}$	-0.4906	0.0015	<2 x 10^{-16}
$\hat{\beta}_{0,3}$	0.5464	0.0014	<2 x 10^{-16}
$\hat{\beta}_{1,3}$	0.3519	0.0010	<2 x 10^{-16}
$\hat{\beta}_{2,3}$	-8.8458	0.0356	<2 x 10^{-16}
$\hat{\beta}_{3,3}$	2.1546	0.0074	<2 x 10^{-16}
$\hat{\beta}_{4,3}$	*N/A*	*N/A*	*N/A*
$\hat{\beta}_{5,3}$	*N/A*	*N/A*	*N/A*
$\hat{\beta}_{6,3}$	-0.3496	0.0011	<2 x 10^{-16}
Max/Min preprocessing			
Coefficient	Estimate	Std. Error	p-value
$\hat{\beta}_{0,2}$	0.2479	0.0008	<2 x 10^{-16}
$\hat{\beta}_{1,2}$	0.3660	0.0019	<2 x 10^{-16}
$\hat{\beta}_{2,2}$	-4.1683	0.0162	<2 x 10^{-16}
$\hat{\beta}_{3,2}$	3.2853	0.0118	<2 x 10^{-16}
$\hat{\beta}_{4,2}$	*N/A*	*N/A*	*N/A*
$\hat{\beta}_{5,2}$	*N/A*	*N/A*	*N/A*
$\hat{\beta}_{6,2}$	-13634	0.0044	<2 x 10^{-16}
$\hat{\beta}_{0,3}$	0.1363	4.6 x 10^{-4}	<2 x 10^{-16}
$\hat{\beta}_{1,3}$	0.3834	0.0011	<2 x 10^{-16}
$\hat{\beta}_{2,3}$	-1.6136	0.0064	<2 x 10^{-16}
$\hat{\beta}_{3,3}$	1.2485	0.0042	<2 x 10^{-16}
$\hat{\beta}_{4,3}$	*N/A*	*N/A*	*N/A*
$\hat{\beta}_{5,3}$	*N/A*	*N/A*	*N/A*
$\hat{\beta}_{6,3}$	-0.7496	0.0024	<2 x 10^{-16}
SD/Mean preprocessing			
Coefficient	Estimate	Std. Error	p-value
$\hat{\beta}_{0,2}$	1.1298	0.0037	<2 x 10^{-16}
$\hat{\beta}_{1,2}$	0.2608	0.0013	<2 x 10^{-16}
$\hat{\beta}_{2,2}$	-6.6737	0.0261	<2 x 10^{-16}
$\hat{\beta}_{3,2}$	6.0411	0.0217	<2 x 10^{-16}
$\hat{\beta}_{4,2}$	*N/A*	*N/A*	*N/A*
$\hat{\beta}_{5,2}$	*N/A*	*N/A*	*N/A*
$\hat{\beta}_{6,2}$	-6.2131	0.0201	<2 x 10^{-16}
$\hat{\beta}_{0,3}$	0.8646	0.0029	<2 x 10^{-16}
$\hat{\beta}_{1,3}$	0.3761	0.0010	<2 x 10^{-16}
$\hat{\beta}_{2,3}$	-3.6	0.0144	<2 x 10^{-16}
$\hat{\beta}_{3,3}$	3.3201	0.0113	<2 x 10^{-16}
$\hat{\beta}_{4,3}$	*N/A*	*N/A*	*N/A*
$\hat{\beta}_{5,3}$	*N/A*	*N/A*	*N/A*
$\hat{\beta}_{6,3}$	-4.7565	0.0157	<2 x 10^{-16}

Table 4: Alias test for checking multicollinearity of non-determined coefficients in MLR

	$\hat{\beta}_{0,2}$	$\hat{\beta}_{1,2}$	$\hat{\beta}_{2,2}$	$\hat{\beta}_{3,2}$	$\hat{\beta}_{6,2}$
$\hat{\beta}_{4,2}$	-3.2352	0	58.8235	-11.7647	1.5294
$\hat{\beta}_{5,2}$	-9.3529	0	388.2353	-117.6471	8.2941
	$\hat{\beta}_{0,3}$	$\hat{\beta}_{1,3}$	$\hat{\beta}_{2,3}$	$\hat{\beta}_{3,3}$	$\hat{\beta}_{6,3}$
$\hat{\beta}_{4,3}$	-3.2352	0	47.0588	-5.8823	1.5294
$\hat{\beta}_{5,3}$	-9.3529	0	270.5882	-58.8235	8.2941

Once MLR models are fitted, an Analysis of Variance must be performed for the raw data and the two preprocessing strategies in the equations (8) and (9), to check if all the variables are significant. This analysis is carried out by using the R function *ANOVA,* obtaining the results in Table 5, showing that hypothesis tests performed to validate that all the model variables are significant, excepting the *soil type* for crops, pasture type 1, and pasture type 2.

Table 5: ANOVA for MLR in (a) estimation for M_2 and (b) estimation for M_3

(a)

Raw data					
Variable	Df	Sum squares	Mean square	F value	p-value
M_1	1	244.84	244.8495	125163	<2 x10^{-16}
D_1	1	54.990	54.9909	281105	<2 x10^{-16}
D_2	1	9.8928	9.8928	50570	<2 x10^{-16}
soil type	1	18.6491	18.6491	95331	<2 x10^{-16}
Residuals	54488	10.6591	0.000195		
Max-Min preprocessing					
Variable	Df	Sum squares	Mean square	F value	p-value
M_1	1	1881.34	1881.34	125387	<2 x10^{-16}
D_1	1	422.75	422.75	281755	<2 x10^{-16}
D_2	1	76.29	76.29	50829	<2 x10^{-16}
soil type	1	144.07	144.07	96022	<2 x10^{-16}
Residuals	54488	81.75	0		
SD / Mean preprocessing					
Variable	Df	Sum squares	Mean square	F value	p-value
M_1	1	39213	39213	124894	<2 x10^{-16}
D_1	1	8779	8779	279632	<2 x10^{-16}
D_2	1	1594	1594	50777	<2 x10^{-16}
soil type	1	3005	3005	95714	<2 x10^{-16}
Residuals	54488	1711	0		

(b)

Raw data					
Variable	Df	Sum squares	Mean square	F value	p-value

M_1	1	240.583	240.5830	230668	<2 x10^{-16}
D_1	1	37.9458	37.9458	363820	<2 x10^{-16}
D_2	1	0.3932	0.3932	70.236	<2 x10^{-16}
soil type	1	9.4532	9.4532	90636.	<2 x10^{-16}
Residuals	54488	5.6830	0.0001043		

Max Min preprocessing

Variable	Df	Sum squares	Mean square	F value	p-value
M_1	1	1115.72	1115.72	234484	<2 x10^{-16}
D_1	1	174.04	174.04	365759	<2 x10^{-16}
D_2	1	1.78	1.78	3750	<2 x10^{-16}
soil type	1	43.53	43.53	91487	<2 x10^{-16}
Residuals	54488	25.93	0		

SD / Mean preprocessing

Variable	Df	Sum squares	Mean square	F value	p-value
M_1	1	39213	39213	124894	<2 x10^{-16}
D_1	1	8779	8779	239632	<2 x10^{-16}
D_2	1	1594	1594	50777	<2 x10^{-16}
soil type	1	3005	3005	95714	<2 x10^{-16}
Residuals	54488	1711	0		

To assess the best preprocessing strategy, we calculate the R^2, the RMSE and the Index of Agreement for each model with the training data, for both M_2 and M_3 (Table 6). Because the preprocessing strategies change the scale of the measurements, it is evident that the R^2 and the Index of Agreement are similar for all the subsets, however, the RMSE is less for the raw data, since all the moistures are in the range 0.2 up to 0.6 m^3/m^3, and because the standard deviations are from 0.029 to 0.113 (Table 2), the span of the transformed sets is higher than the raw data, according to equations (8) and (9).

Table 6: Performance metrics for the sets of raw and preprocessed data for (a) M_2 and (b) M_3

(a)

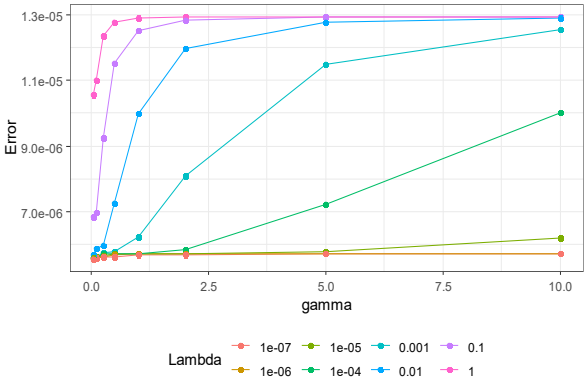
	R^2	RMSE	Index of Agreement
Raw data	0.9685	0.14	0.9217
Max / Min	0.9685	0.0387	0.9217
SD / Mean	0.9683	0.177	0.9217

(b)

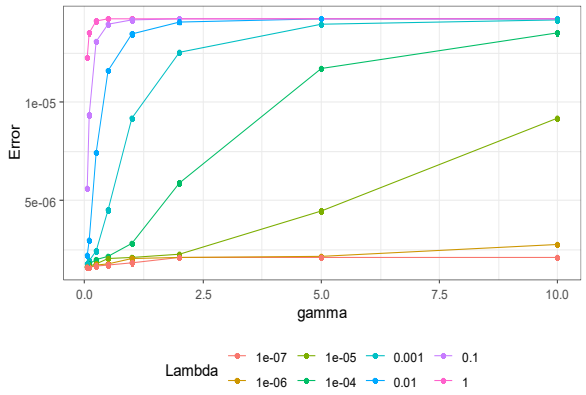
	R^2	RMSE	Index of Agreement
Raw data	0.9808	0.0101	0.9401
Max / Min	0.9808	0.0218	0.9402
SD / Mean	0.9807	0.1387	0.94

4.2. SVR

An SVR is trained to perform the regression of M_2 and M_3, based on M_1, D_1, D_2 or D_3, F, P_1, and P_2. Three types of kernels are tested: linear, polynomial, and Radial Basis Function (RBF) [64]. A cross-validation process is carried out to determine which of the kernels exhibits better performance, by using the function *tune* in R Studio, considering the inverse of the regularization parameter cost, known as lambda (λ) and kernel width (γ) as hyperparameters. Since predictions with linear and polynomial kernels did not exhibit acceptable errors, we selected the RBF kernel to show the regressor tuning process. The steps developed to tune the SVR are as follows (for each preprocessing scheme): *i)* split raw data and preprocessed data into training and test sets, *ii)* configure a 2D grid with the values of the hyperparameters λ = *(0.0001, 0.001, 0.01, 0.1, 1, 10, 100)* and γ = *(0.001, 0.01, 0.1, 1, 10, 100), iii)* determine an SVR model for each subset of hyperparameters, *iv)* evaluate the regression errors of each model, and *v)* choose the model with the least regression error. After tuning all the SVRs, we found the best hyperparameters configuration. Regarding the training sets, Figure 7(a) shows the regression errors for the raw data of M_2 (which exhibited the least regression error = 4.36 x 10^{-6}), considering the hyperparameters λ =1 x 10^{-7} and γ = 0.001, and Figure 7(b) shows the regression errors for the raw data of M_3 (which exhibited the least regression error = 1.59 x 10^{-6}), considering the hyperparameters λ =1 x 10^{-5} and γ = 0.1. Besides, the regression errors for each preprocessing configuration (for the training sets) are shown in Table 7. Finally, for M_2, we obtained R^2= 0.985, RMSE = 0.0115, and Index of Agreement = 0.9422. Similarly, for M_3, we obtained R^2= 0.9885, RMSE = 0.0078, and Index of Agreement = 0.9542.



(a) Raw data for M_2



(b) Raw data for M_3

Figure 7: Hyperparameters selection for M_2 and M_3 for the training subset

Table 7: Regressor errors for each preprocessing scheme for the training subset

M_2	
Preprocessing	Regressor error
Raw	1.43 x 10^{-5}
Max/min	4.36 x 10^{-6}
SD/Mean	5.0387 x 10^{-6}
M_3	
Preprocessing	Regressor error
Raw	1.59 x 10^{-6}
Max/min	7.47 x 10^{-6}
SD/Mean	0.00032046

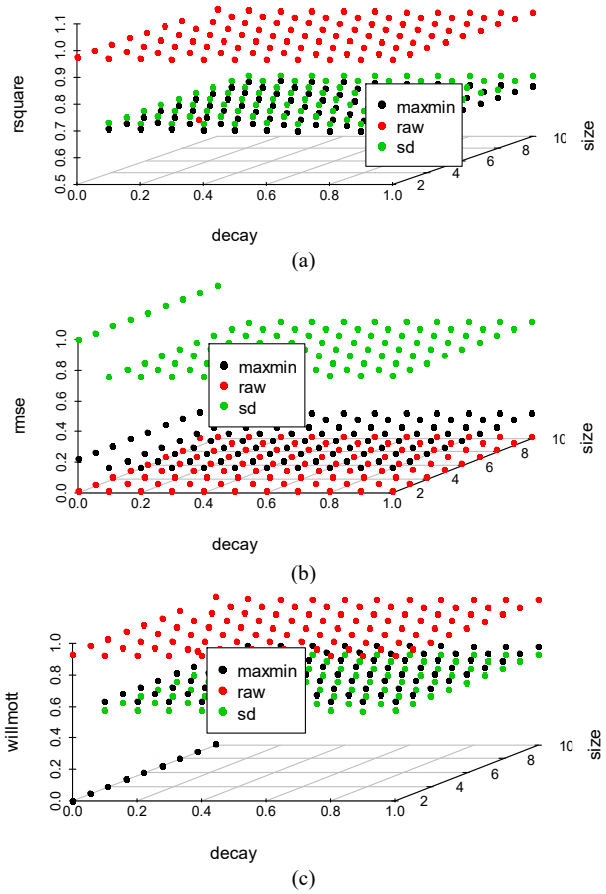


Figure 8: ANN regression performance for each preprocessing scheme for (a) R^2, (b) RMSE, and (c) Index of Agreement (Willmott) for M_2

4.3. ANN

An ANN is proposed to predict soil moisture. Two multilayer perceptrons are trained to perform the regressions of M_2 and M_3, based on M_1, D_1, D_2 or D_3, F, P_1, and P_2. A cross-validation process was carried out aiming to tune the hyper parameters of the ANN, i.e., *i)* best number of hidden layers, taking into account a tradeoff between low estimation error and simplicity, *ii)* weights values for each connection and *iii)* activation function the neurons (binary step, linear, sigmoid, tanh and Rectified Linear Unit - RELU- were considered). We have followed the next procedure to choose the best configuration: *i)* take raw data, or preprocessed data with the equations (8) or (9) and divide them into training (70%) and test (30%) sets, *ii)* train a perceptron with a hidden layer and a number of neurons from 2 up to 10, and a decay from 0 up to 1 with steps of 0.1, *iii)* for each pair of hyperparameters, evaluate R^2, RMSE and the Index of Agreement (Willmott) for the training set, *iv)* inspect and select the best preprocessing scheme, *v)* plot errors for the selected preprocessing scheme according to the set of hyperparameters chosen, and *vi)* select the configuration of the neural network. Figure 8 and Figure 9 show different regression errors for the different preprocessing schemes for M_2 and M_3.

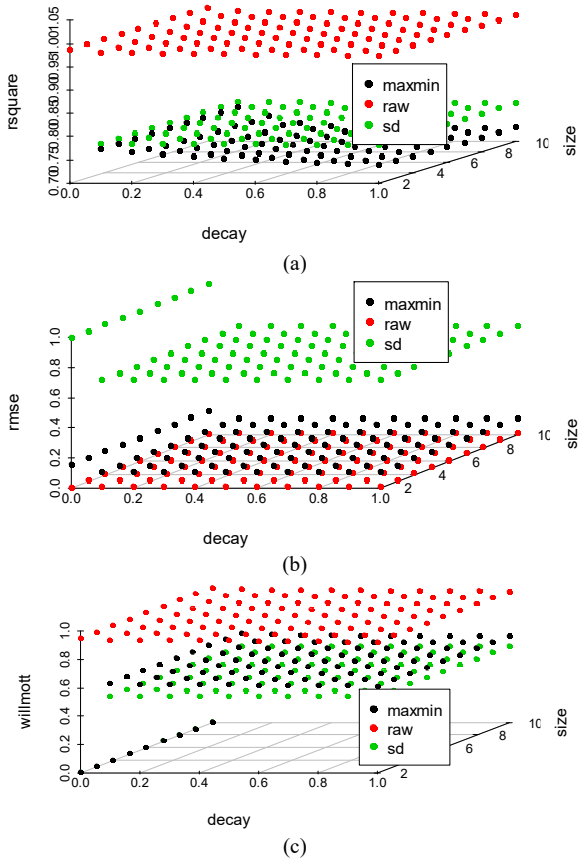
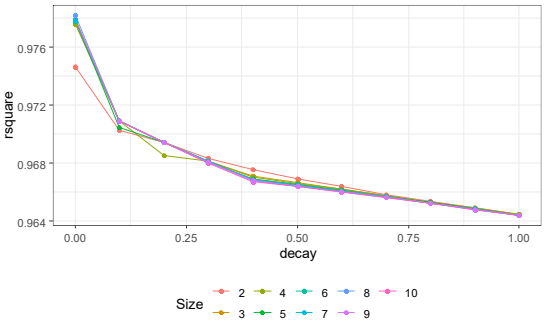
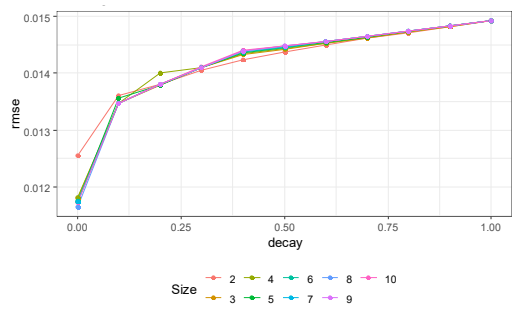


Figure 9: ANN regression performance for each preprocessing scheme for (a) R2, (b) RMSE, and (c) Index of Agreement (Willmott) for M_3

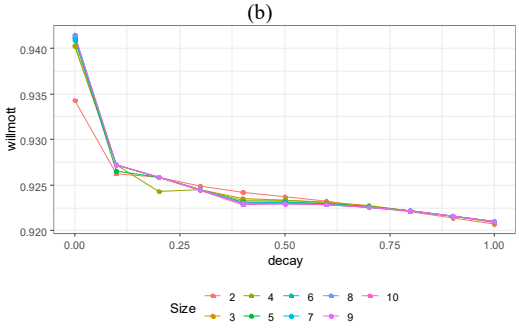
From Figure 8 and Figure 9, it can be seen that all the configurations of ANNs have better performance with raw data, since R^2 and Index of Agreement are close to 1, and RMSE is close to 0. Thus, we selected the raw data to get the best ANN configuration, as shown in Figure 10 and Figure 11.



(a)

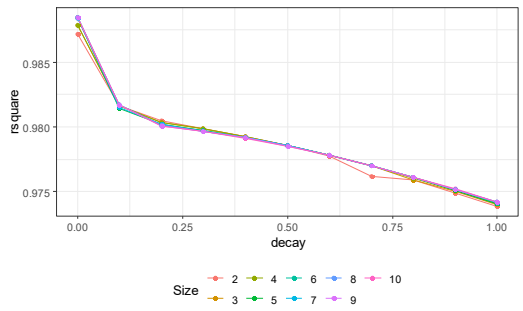


(b)

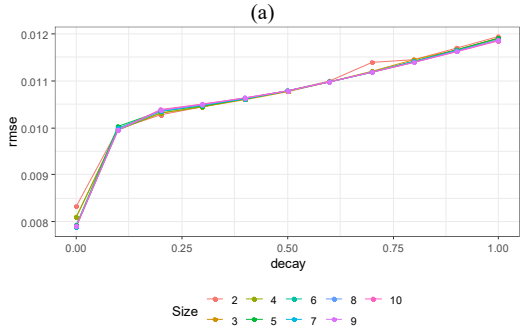


(c)

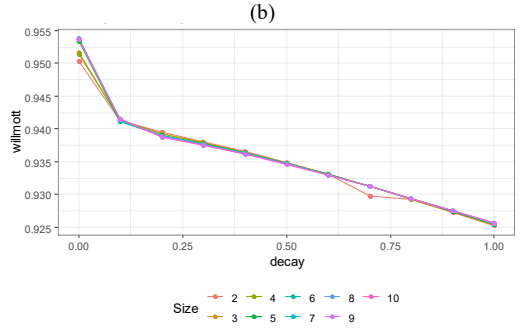
Figure 10: ANN regression performance for raw data for (a) R^2, (b) RMSE, and (c) Index of Agreement (Willmott) for M_2



(a)



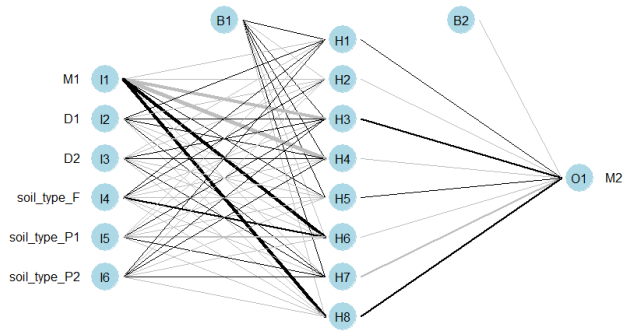
(b)



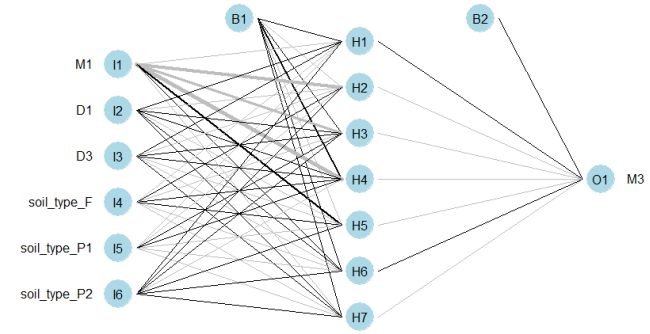
(c)

Figure 11: ANN regression performance for raw data for (a) R^2, (b) RMSE, and (c) Index of Agreement (Willmott) for M_3

From Figure 10, we obtained the best ANN configuration for M_2 as follows: an input layer with six neurons, a hidden layer with eight neurons, an output layer with one layer, and a decay = 0 (R^2=0.9781, RMSE=0.0116, and Index of Agreement=0.9414). Similarly, for M_3, we found the best configuration with a hidden layer of 7 neurons and a decay = 0 (R^2=0.9884, RMSE=0.0079, and Index of Agreement=0.9534). The ANNs' configurations are depicted in Figure 12.



(a)



(b)

Figure 12: Multilayer perceptron ANN for (a)M_2 estimation and (b) M_3 estimation

In Figure 12, the darker and broader the line between two neurons is, the greater the weight is, which represents the relative importance (positive) of such connection. On the other hand, the grayer and broader the line between two neurons is, the lesser the weight is, which represents the relative importance (negative) of such connection. The tuned weights are shown in Table 8.

Table 8: Tuned weights for neurons connections of the ANN. Table (a) shows the weights from the input layer to the hidden layer for M_2 estimation, table (b) shows the weights from the hidden layer to the output for M_2 estimation, table (c) shows the weights from the input layer to the hidden layer for M_3 estimation, table (d) shows the weights from the hidden layer to the output for M_3 estimation.

(a)

	B1	I1	I2	I3	I4	I5	I6
H1	0.54	-2.61	0.15	-0.90	4.58	-8.25	1.71
H2	-5.51	-5.17	-0.63	-2.37	4.15	-6.48	-0.29
H3	11.76	-49.2	2.19	5.04	-9.21	8.75	0.46
H4	13.42	-54.7	1.68	2.96	-7.80	-11.8	0.25
H5	2.61	10.10	-0.60	-0.37	-2.94	-9.21	-5.50
H6	-13.4	52.54	-2.25	-6.89	14.50	-16.1	-0.90
H7	1.84	-2.64	0.86	1.68	-5.57	5.91	3.73
H8	-12.6	49.26	-2.26	-4.89	-5.09	-8.00	-0.66

(b)

	B2	H1	H2	H3	H4	H5	H6	H7	H8
O1	-2.9	7.86	-2.07	17.10	-7.78	2.27	-7.57	-14.44	16.95

(c)

	B1	I1	I2	I3	I4	I5	I6
H1	4.40	-6.37	0.81	1.36	-2.21	-0.84	0.66
H2	-1.44	-36.24	-0.70	-0.80	5.94	-0.48	2.11
H3	5.53	-24.60	0.78	4.76	-1.36	9.97	0.05
H4	15.28	-45.70	1.70	6.19	4.33	3.73	-4.13
H5	-4.00	17.68	-0.98	-2.84	2.12	-5.66	0.38
H6	0.14	11.19	0.51	-2.04	-8.67	-7.35	4.88
H7	2.07	-1.97	0.64	0.63	-9.89	-3.19	0.51

(d)

	B2	H1	H2	H3	H4	H5	H6	H7
O1	1.47	2.42	-11.36	-1.99	-0.28	-3.38	5.10	-6.77

Table 9: Performance metrics for regression models (test set) ANN, SVR and MLR

	R^2	RMSE	Index of Agreement
Test set ANN M_2	0.978	0.0116	0.9411
Test set MLR M_2	0.9686	0.014	0.9217
Test set SVR M_2	0.9786	0.0115	0.9423
Test set ANN M_3	0.988	0.008	0.9551
Test set MLR M_3	0.9885	0.01	0.9411
Test set SVR M_3	0.9808	0.007	0.9542

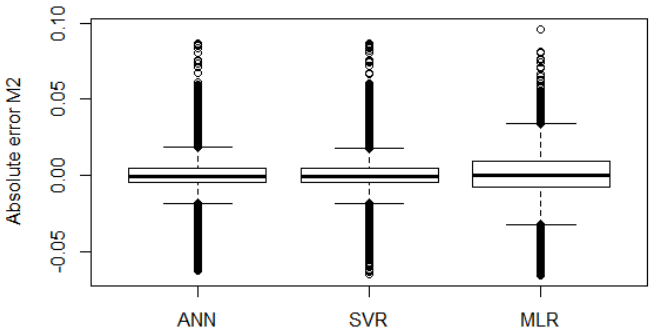


Figure 13: Absolute errors for the test set of M_2

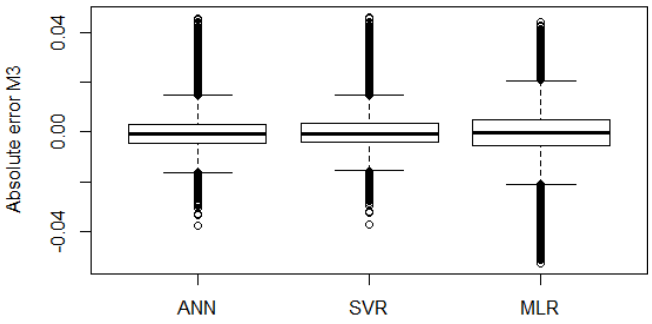


Figure 14: Absolute errors for the test set of M_3

5. Results

After fitting the parametric and non-parametric methods, some performance metrics will be applied to test what method could be used with better results to represent the soil moisture variability toward the soil profile for different type of soil and land cover configurations. It will indicate what model could be applied to extrapolate the soil moisture profiles toward areas with scarce information. To evaluate the models' performances, R^2 , RMSE, and the Index of Agreement are proposed, as shown in Table 9. It can be noticed that, at first glance, the general performance of the methods is almost the same with slightly worse results for the MLR method for the sites presented. A deeper analysis can be carried out from the error dispersions of each method, as depicted in Figure 13 and Figure 14, where it can be noticed that error dispersion is higher for MLR, and the absolute error performances of ANNs and SVR are similar. Furthermore, when plotting the time-series (Figure 15 and Figure 16), it can be noticed that a deeper analysis should be carried out regarding the soil types.

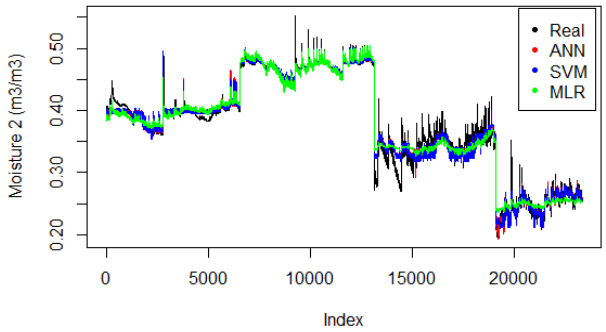


Figure 15: Regression results for the test set of M_2. Index interval from 0 up to 6.500 corresponds to crops, from 6.501 to 13.100 corresponds to pasture 1, from 13.101 up to 19.100 corresponds up to pasture 2, and from 19.101 up to 23.355 corresponds to the forest

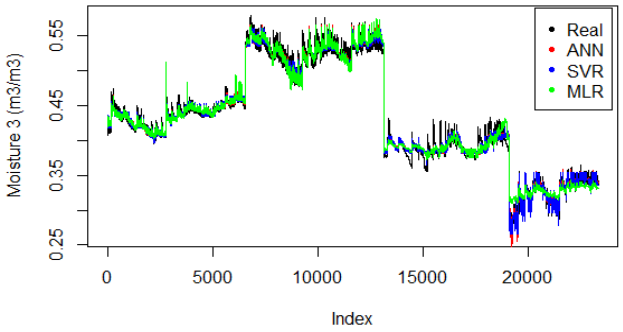


Figure 16: Regression results for the test set of M_3 . Index interval from 0 up to 6.500 corresponds to crops, from 6.501 to 13.100 corresponds to pasture 1, from 13.101 up to 19.100 corresponds up to pasture 2, and from 19.101 up to 23.355 corresponds to the forest.

It can be noticed that the methods (ANN, MLR, and SVR) exhibit similar results for the crop, pasture 1, and pasture 2 in terms of average variability of the soil moisture (Figure 15 and Figure 16). Nonetheless, SVR outperforms the other two methods in terms of high variability when comparing versus the measured data. In the case of the forest, it is more evident that SVR can better predict the real values of both M_2 and M_3. On the other hand, none of the methods are accurate enough to predict behaviours in pasture 2. Due to this visual inspection, RMSE, R^2, and Index of Agreement are recomputed but segmented by each *soil type,* as shown in Table 10.

Table 10: Performance metrics for regression models ANN, MLR, and SVR differenced by *soil type*

Soil type – Moisture	Performance metric	R^2	RMSE	Index of Agreement
Crop M_2	Test ANN	0.8095	0.0425	0.7922
	Test MLR	0.8129	0.0424	0.8027
	Test SVR	0.6542	0.1168	0.6031

Crop M_3	Test ANN	0.9307	0.0044	0.8725
	Test MLR	0.9325	0.0043	0.8782
	Test SVR	0.9344	0.0044	0.8993
Pasture 1 M_2	Test ANN	0.8176	0.0583	0.7722
	Test MLR	0.8174	0.0588	0.7695
	Test SVR	0.8246	0.0049	0.7715
Pasture 1 M_3	Test ANN	0.8439	0.0077	0.7824
	Test MLR	0.8444	0.0081	0.7902
	Test SVR	0.8585	0.0074	0.8136
Pasture 2 M_2	Test ANN	0.2836	0.0551	0.2526
	Test MLR	0.2770	0.0551	0.2468
	Test SVR	0.3386	0.0196	0.3121
Pasture 2 M_3	Test ANN	0.3749	0.0125	0.3185
	Test MLR	0.3560	0.0128	0.3054
	Test SVR	0.4419	0.0119	0.4112
Forest M_2	Test ANN	0.8247	0.0785	0.8063
	Test MLR	0.8245	0.0795	0.7985
	Test SVR	0.8264	0.0079	0.8111
Forest M_3	Test ANN	0.8247	0.0185	0.7954
	Test MLR	0.3245	0.0796	0.2212
	Test SVR	0.9207	0.0059	0.8924

It can be noted, according to Table 10, that SVR outputs better results in most of the cases, excepting for M_2 in the crop, where MLR fits better data with the test set. In the rest of the cases, ANN outperforms MLR. Since the R^2 value depends on the individual standard deviations, it is possible that the segmentations conducted for analyzing the individual performance regarding the soil type present different fitting values than the case of the whole dataset.

For the sake of statistical validity, some goodness-of-fit tests must be carried out for the parametric method (MLR): normality, autocorrelation, and homoskedasticity. All the tests are calculated using residual errors. For normality, the Kolmogorov Smirnov test [38] is applied, obtaining a p-value = 2×10^{-16}, so residuals are not normal. For autocorrelation, the Durbin test is applied, obtaining a DW = 0.02. According to [37], if the value is too far from 2, there is a autocorrelation. Finally, for homoskedasticity, the Breusch Pagan test [39] is carried out, obtaining a p-value=2×10^{-16}, which means that residuals are heteroskedastic. Because there is no normality, correlation, nor homoskedasticity, the linear regression model could not be applied.

6. Conclusions

In this paper, a comparison between parametric and non-parametric methods for fitting soil moistures in Las Palmas Andean Basin. This research was motivated due to scarce information from satellite images and in-situ measurements toward the tropical areas where the soil moisture plays an important role in climate variability, so more in-depth soil moisture research is not possible.

The proposed methodology based on parametric and non-parametric methods employing a database collected from three typical land covers: crop, pasture, and forest, and typical soil types demonstrated that there are high correlations between superficial and intermediate soil moistures, as well as superficial and deepest soil moistures, so it is possible to propose prediction models to avoid the need to install many instruments per point or even the high ability of the proposed method to use surface information from satellite data to obtain soil moisture profiles.

A model fitting using MLR, SVR, and ANN was carried out, aiming to predict the soil moisture, by using as predictor variables the superficial soil moisture, its depth, the depth of the soil moisture that wants to be predicted, and the *soil type*. All models offer similar performance metrics, except for forest, where the SVR outperforms the other methods. However, goodness-of-fit tests are not met by residual errors of the MLR model, so it cannot be used with statistical validity. In that way, SVR is the best model to fit moistures based on superficial soil moisture.

Finally, it is important to state that the traditional hydrological models had been calibrated employing few information of the full complex and non-linear hydrological rainfall runoff transformation process (normally flow in specific points), avoiding the understanding of the spatial variability along the full catchment and imposing calibration parameters representing just a small part of the physical phenomenon. The inclusion of distributed spatial parameters in the calibration process such as infiltration parameters or soil moisture profiles can increase strongly the quality of the results obtained along the full catchment or even increase the understanding of the physics involved. In this direction the proposed methodology represents an important contribution to several engineering and hydrological applications.

7. Future work

We are planning to install two additional soil moisture stations in another forest, and residential soil uses, to better tune the models.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Special thanks to the University the Medellín for financing the research project “Interacciones entre la cobertura vegetal, la variabilidad climática y el contenido de humedad en el suelo, en cuencas agrícolas neotropicales”, Project code 1070.

References

[1] M. Gonzalez-Palacio, L. Sepulveda-Cano, J.D. Valencia-Calvo, J. Quiza-Montealegre, “System dynamics baseline model for determining a multivariable objetive function in Wireless Sensor Networks,” in CISTI 2020, 2020.

[2] V.R. Pauwels, R. Hoeben, N.E. Verhoest, F.P. De Troch, “The importance of the spatial patterns of remotely sensed soil moisture in the improvement of discharge predictions for small-scale basins through data assimilation,” Journal of Hydrology, **251**, 88–102, 2001.

[3] H. Sharma, M.K. Shukla, P.W. Bosland, R. Steiner, “Soil moisture sensor

calibration, actual evapotranspiration, and crop coefficients for drip irrigated greenhouse chile peppers," Agricultural Water Management, **179**, 81–91, 2017, doi:10.1016/j.agwat.2016.07.001.

[4] S. Walther, G. Duveiller, M. Jung, L. Guanter, A. Cescatti, G. Camps-Valls, "Satellite Observations of the Contrasting Response of Trees and Grasses to Variations in Water Availability," Geophysical Research Letters, **46**(3), 1429–1440, 2019, doi:10.1029/2018GL080535.

[5] H. Janssen, G.A. Scheffler, R. Plagge, "Experimental study of dynamic effects in moisture transfer in building materials," International Journal of Heat and Mass Transfer, **98**, 141–149, 2016, doi:10.1016/j.ijheatmasstransfer.2016.03.031.

[6] L. Zhuo, Q. Dai, D. Han, N. Chen, B. Zhao, M. Berti, "Evaluation of Remotely Sensed Soil Moisture for Landslide Hazard Assessment," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, **12**(1), 162–173, 2019, doi:10.1109/JSTARS.2018.2883361.

[7] L. Brocca, L. Ciabatta, C. Massari, S. Camici, A. Tarpanelli, "Soil moisture for hydrological applications: Open questions and new opportunities," Water (Switzerland), **9**(2), 2017, doi:10.3390/w9020140.

[8] X. Huang, Z.H. Shi, H.D. Zhu, H.Y. Zhang, L. Ai, W. Yin, "Soil moisture dynamics within soil profiles and associated environmental controls," Catena, **136**, 189–196, 2016, doi:10.1016/j.catena.2015.01.014.

[9] I. V. Florinsky, Digital Terrain Analysis in Soil Science and Geology: Second Edition, Elsevier Inc., 2016.

[10] J. Liu, B.A. Engel, Y. Wang, Y. Wu, Z. Zhang, M. Zhang, "Runoff Response to Soil Moisture and Micro-topographic Structure on the Plot Scale," Scientific Reports, **9**(1), 2019, doi:10.1038/s41598-019-39409-6.

[11] M. Pan, E.F. Wood, "Impact of Accuracy, Spatial Availability, and Revisit Time of Satellite-Derived Surface Soil Moisture in a Multiscale Ensemble Data Assimilation System," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, **3**(1), 49–56, 2010, doi:10.1109/JSTARS.2010.2040585.

[12] B. Kuang, Y. Tekin, A.M. Mouazen, "Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content," Soil and Tillage Research, **146**(PB), 243–252, 2015, doi:10.1016/j.still.2014.11.002.

[13] K. Were, D.T. Bui, Ø.B. Dick, B.R. Singh, "A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape," Ecological Indicators, **52**, 394–403, 2015, doi:10.1016/j.ecolind.2014.12.028.

[14] S. Maroufpoor, E. Maroufpoor, O. Bozorg-Haddad, J. Shiri, Z. Mundher Yaseen, "Soil moisture simulation using hybrid artificial intelligent model: Hybridization of adaptive neuro fuzzy inference system with grey wolf optimizer algorithm," Journal of Hydrology, **575**, 544–556, 2019, doi:10.1016/j.jhydrol.2019.05.045.

[15] M.K. Gill, T. Asefa, M.W. Kemblowski, M. McKee, "Soil moisture prediction using support vector machines," Journal of the American Water Resources Association, **42**(4), 1033–1046, 2006, doi:10.1111/j.1752-1688.2006.tb04512.x.

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

[17] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang, J. Gao, L. Zhang, "Deep learning in environmental remote sensing: Achievements and challenges," Remote Sensing of Environment, **241**, 2020, doi:10.1016/j.rse.2020.111716.

[18] G. Dumedah, J.P. Walker, L. Chik, "Assessing artificial neural networks and statistical methods for infilling missing soil moisture records," Journal of Hydrology, **515**, 330–344, 2014, doi:10.1016/j.jhydrol.2014.04.068.

[19] M. Khalil, U.S. Panu, W.C. Lennox, "Groups and neural networks based streamflow data infilling procedures," Journal of Hydrology, **241**(3–4), 153–176, 2001, doi:10.1016/S0022-1694(00)00332-2.

[20] F.D. Mwale, A.J. Adeloye, R. Rustum, "Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi - A self organizing map approach," Physics and Chemistry of the Earth, **50**–**52**, 34–43, 2012, doi:10.1016/j.pce.2012.09.006.

[21] T.R. Nkuna, J.O. Odiyo, "Filling of missing rainfall data in Luvuvhu River Catchment using artificial neural networks," Physics and Chemistry of the Earth, **36**(14–15), 830–835, 2011, doi:10.1016/j.pce.2011.07.041.

[22] P. Coulibaly, N.D. Evora, "Comparison of neural network methods for infilling missing daily weather records," Journal of Hydrology, **341**(1–2), 27–41, 2007, doi:10.1016/j.jhydrol.2007.04.020.

[23] M. Pal, R. Maity, "Development of a spatially-varying Statistical Soil Moisture Profile model by coupling memory and forcing using hydrologic soil groups," Journal of Hydrology, **570**, 141–155, 2019, doi:10.1016/j.jhydrol.2018.12.042.

[24] M. Aboutalebi, N. Allen, A.F. Torres-Rua, M. McKee, C. Coopmans, "Estimation of soil moisture at different soil levels using machine learning techniques and unmanned aerial vehicle (UAV) multispectral imagery," SPIE-Intl Soc Optical Eng: 26, 2019, doi:10.1117/12.2519743.

[25] R. Girden, "ANOVA: Repeated measures," Computer Science, 1991.

[26] N. Rodríguez-Fernández, P. de Rosnay, C. Albergel, P. Richaume, F. Aires, C. Prigent, Y. Kerr, "SMOS neural network soil moisture data assimilation in a land surface model and atmospheric impact," Remote Sensing, **11**(11), 2019, doi:10.3390/rs11111334.

[27] X. Dai, Z. Huo, H. Wang, "Simulation for response of crop yield to soil moisture and salinity with artificial neural network," Field Crops Research, **121**(3), 441–449, 2011, doi:10.1016/j.fcr.2011.01.016.

[28] W.E.H. Blum, Functions of soil for society and the environment, Reviews in Environmental Science and Biotechnology, **4**(3), 75–79, 2005, doi:10.1007/s11157-005-2236-x.

[29] K. Liao, X. Lai, Z. Zhou, Q. Zhu, "Applying fractal analysis to detect spatio-temporal variability of soil moisture content on two contrasting land use hillslopes," Catena, **157**, 163–172, 2017, doi:10.1016/j.catena.2017.05.022.

[30] J. Geris, D. Tetzlaff, J.J. McDonnell, C. Soulsby, "Spatial and temporal patterns of soil water storage and vegetation water use in humid northern catchments," Science of the Total Environment, **595**, 486–493, 2017, doi:10.1016/j.scitotenv.2017.03.275.

[31] L. Brocca, T. Moramarco, F. Melone, W. Wagner, "A new method for rainfall estimation through soil moisture observations," Geophysical Research Letters, **40**(5), 853–858, 2013, doi:10.1002/grl.50173.

[32] W. Dorigo, W. Wagner, C. Albergel, F. Albrecht, G. Balsamo, L. Brocca, D. Chung, M. Ertl, M. Forkel, A. Gruber, E. Haas, P.D. Hamer, M. Hirschi, J. Ikonen, R. de Jeu, R. Kidd, W. Lahoz, Y.Y. Liu, D. Miralles, T. Mistelbauer, N. Nicolai-Shaw, R. Parinussa, C. Pratola, C. Reimer, R. van der Schalie, S.I. Seneviratne, T. Smolander, P. Lecomte, "ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions," Remote Sensing of Environment, **203**, 185–215, 2017, doi:10.1016/j.rse.2017.07.001.

[33] H. Lin, H.J. Vogel, J. Phillips, B.D. Fath, Complexity of soils and hydrology in ecosystems, Ecological Modelling, **298**, 1–3, 2015, doi:10.1016/j.ecolmodel.2014.11.016.

[34] X. Jia, M. Shao, Y. Zhu, Y. Luo, "Soil moisture decline due to afforestation across the Loess Plateau, China," Journal of Hydrology, **546**, 113–122, 2017, doi:10.1016/j.jhydrol.2017.01.011.

[35] S. Zhang, W. Fan, Y. Li, Y. Yi, "The influence of changes in land use and landscape patterns on soil erosion in a watershed تاثیر تغییرات کاربری اراضی و الگوی شکل شناسی زمین برروی فرسایش خاک در حوزه," Science of The Total Environment, **574**, 34–45, 2017.

[36] L. Gao, Y. Lv, D. Wang, T. Muhammad, A. Biswas, X. Peng, "Soil water storage prediction at high space-time resolution along an agricultural hillslope," Agricultural Water Management, **165**, 122–130, 2016, doi:10.1016/j.agwat.2015.11.012.

[37] K.J. White, "The Durbin-Watson Test for Autocorrelation in Nonlinear Models," The Review of Economics and Statistics, **74**(2), 370, 1992, doi:10.2307/2109675.

[38] M. F.J.Jr., "The Kolmogorov-Smirnov test for goodness of fit," Journal of the American Statistical Association, **56**(1951), 68–78, 1951.

[39] D.M. Waldman, "A note on algebraic equivalence of White's test and a variation of the Godfrey/Breusch-Pagan test for heteroscedasticity," Economics Letters, **13**(2–3), 197–200, 1983, doi:10.1016/0165-1765(83)90085-X.

[40] W.S. Noble, What is a support vector machine?, Nature Biotechnology, **24**(12), 1565–1567, 2006, doi:10.1038/nbt1206-1565.

[41] N. Deng, Y. Tian, C. Zhang, Support vector machines: Optimization based theory, algorithms, and extensions, CRC Press, 2012, doi:10.1201/b14297.

[42] T. Kavzoglu, I. Colkesen, "A kernel functions analysis for support vector machines for land cover classification," International Journal of Applied Earth Observation and Geoinformation, **11**(5), 352–359, 2009, doi:10.1016/j.jag.2009.06.002.

[43] W. Zhou, L. Zhang, L. Jiao, J. Pan, "Support vector regression based on unconstrained convex quadratic programming," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag: 167–174, 2006, doi:10.1007/11881070_27.

[44] M. Awad, R. Khanna, Support vector regression, Apress, Berkeley: 67–80, 2015, doi:doi.org/10.1007/978-1-4302-5990-9_4.

[45] L. Tian, X. ZHANG, A Convergent Nonlinear Smooth Support Vector Regression Model, 205–207, 2015, doi:10.2991/978-94-6239-102-4_43.

[46] Y.O. Ouma, C.O. Okuku, E.N. Njau, "Use of Artificial Neural Networks and Multiple Linear Regression Model for the Prediction of Dissolved Oxygen in Rivers: Case Study of Hydrographic Basin of River Nyando, Kenya," Complexity, **2020**, 2020, doi:10.1155/2020/9570789.
[47] A. Landi, P. Piaggi, M. Laurino, D. Menicucci, "Artificial neural networks for nonlinear regression and classification," in Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications, ISDA'10, 115–120, 2010, doi:10.1109/ISDA.2010.5687280.
[48] A. Biglarian, E. Bakhshi, A.R. Baghestani, M.R. Gohari, M. Rahgozar, M. Karimloo, "Nonlinear survival regression using artificial neural network," Journal of Probability and Statistics, 2013, doi:10.1155/2013/753930.
[49] P.L. Fernández-Cabán, F.J. Masters, B.M. Phillips, "Predicting roof pressures on a low-rise structure from freestream turbulence using artificial neural networks," Frontiers in Built Environment, **4**, 2018, doi:10.3389/fbuil.2018.00068.
[50] B. Liu, M. Shao, "Modeling soil-water dynamics and soil-water carrying capacity for vegetation on the Loess Plateau, China," Agricultural Water Management, **159**, 176–184, 2015, doi:10.1016/j.agwat.2015.06.019.
[51] H. YiLong, C. LiDing, F. BoJie, H. ZhiLin, G. Jie, L. XiXi, "Effect of land use and topography on spatial variability of soil moisture in a gully catchment of the Loess Plateau, China.," Ecohydrology, **5**(6), 826–833, 2012.
[52] X. Fang, W. Zhao, L. Wang, Q. Feng, J. Ding, Y. Liu, X. Zhang, "Variations of deep soil moisture under different vegetation types and influencing factors in a watershed of the Loess Plateau, China," Hydrology and Earth System Sciences, **20**(8), 3309–3323, 2016, doi:10.5194/hess-20-3309-2016.
[53] C. Zhu, Y. Li, "Long-Term Hydrological Impacts of Land Use/Land Cover Change From 1984 to 2010 in the Little River Watershed, Tennessee," International Soil and Water Conservation Research, **2**(2), 11–21, 2014, doi:10.1016/S2095-6339(15)30002-2.
[54] L. Gao, M. Shao, "Temporal stability of soil water storage in diverse soil layers," Catena, **95**, 24–32, 2012, doi:10.1016/j.catena.2012.02.020.
[55] X. Mei, Q. Zhu, L. Ma, D. Zhang, H. Liu, M. Xue, "The spatial variability of soil water storage and its controlling factors during dry and wet periods on loess hillslopes," Catena, **162**, 333–344, 2018, doi:10.1016/j.catena.2017.10.029.
[56] B. Yang, X. Wen, X. Sun, "Seasonal variations in depth of water uptake for a subtropical coniferous plantation subjected to drought in an East Asian monsoon region," Agricultural and Forest Meteorology, **201**, 218–228, 2015, doi:10.1016/j.agrformet.2014.11.020.
[57] Soil Survey Staff, Keys to soil taxonomy, 2014.
[58] IGAC, General Study of Soils and Land Zoning: Department of Antioquia (In Spanish), Instituto Geografico Agustin Codazzi, Bogota, Colombia, 2007.
[59] IDEA, Semi-detailed Study of Soil in Zone 13 of the Municipality of Envigado for Potential Use Purposes (In Spanish), 2014.
[60] Y. Zhang, L. Qiao, C. Chen, L. Tian, X. Zheng, "Effects of organic ground covers on soil moisture content of urban green spaces in semi-humid areas of China," Alexandria Engineering Journal, 2020, doi:10.1016/j.aej.2020.08.001.
[61] Unidots, Ubidots IoT Platform, 2020.
[62] R.L. Burden, F. J.D., Numerical Analysis, Brooks/Cole, Cengage Learning, 2011.
[63] J.M. Chambers, A.E. Freeny, R.M. Heiberger, Analysis of variance; designed experiments, CRC Press: 145–193, 2017, doi:10.1201/9780203738535.
[64] S. Fine, K. Scheinberg, "Efficient svm training using low-rank kernel representations," Journal of Machine Learning Research, **2**(Dec), 243–264, 2002.