Assessment of Non-parametric Methods for Soil Moisture Retrieval from Active Microwave Data

Tarendra Lakhankar¹, Hosni Ghedira¹, Marouane Temimi¹, Manajit Sengupta², Reza Khanbilvardi¹, and Reginald Blake¹

¹ NOAA - Cooperative Remote Sensing Science & Technology Center (NOAA-CREST), City University of New York, NY USA.
² Cooperative Institute for Research in the Atmosphere (CIRA), Colorado State University, Fort Collins, USA.

Abstract:— Active microwave remote sensing observations hold the potential for efficient and reliable mapping of spatial soil moisture distributions. However, soil moisture retrievals from microwave remote sensing techniques are typically complex because of the inherent difficulty in characterizing the interactions among land surface parameters that contribute to the retrieval process. Therefore adequate physical mathematical descriptions of the interaction of microwave radiation with parameters such as land cover, vegetation density, and soil characteristics are not readily available. On the other hand it may possible to use non-parametric classifiers like neural networks, fuzzy logic and multiple regression models to retrieve soil moisture distributions. In this study we make use such classifiers after using soil moisture data derived using ESTAR for training the non-parametric models due to limited availability of in-situ soil moisture measurements. The fuzzy logic and neural network models performed better when compared to multiple regression models. It was also seen that the inclusion of the vegetation and soil characteristics, derived from infrared and visible measurements, in these models have significant positive impact on soil moisture retrievals with RMSE being reduced by around 30% in the retrievals. Finally the soil moisture derived from these models was compared with ESTAR soil moisture (RMSE ~4.0%) and field soil moisture measurements (RMSE ~6.5%). Additionally, the study showed that soil moisture retrievals from highly vegetated areas are less accurate than that from bare soil areas.

Key Terms:— Soil Moisture, Active Microwave, Neural Network, Fuzzy Logic, Vegetation.

1 Introduction

The temporal and spatial variations of soil moisture are two key parameters needed in many hydrological modeling processes. Conventional field measurement techniques have serious limitations in their ability to appropriately estimate the spatial distribution of soil moisture, particularly over large areas that are characterized by soil surface heterogeneity. In the recent past, most hydrological models that required soil moisture information used point measurements or spatial distributions of soil moisture derived from physically-based models. Presently, however, spatial distributions of high resolution soil moisture estimates are being increasingly used as input to hydrological models to predict the runoff [1].

A number of spaceborne active microwave missions such as ERS-1, ERS-2, JERS-1, SIR-C/X-SAR and RADARSAT-1 demonstrated that soil moisture of the upper ~5 cm of the surface can be measured from space. In the future, the launching of active microwave sensors such as the Advanced Scatterometer (ASCAT) on EUMETSAT’s Polar System METOP, Canadian RADARSAT-2, European SMOS, Indian Radar Imaging Satellite (RISAT), and the newly programmed NASA SMAP mission are expected to enhance the capability to remotely sensed soil moisture over the next decades. The ASCAT scatterometer will be a continuation of the ERS scatterometer mission, and the METOP and SMOS will be the first operational satellite system dedicated to the retrieval of soil moisture.

The accuracy of satellite-derived soil moisture is usually affected by land surface characteristics such as vegetation. The vegetation canopy may affect the backscattered energy by contributing to the volume backscatter of the observed scene and by attenuating the soil component of the total backscatter [2]. Many researchers used a linear regression model to simplify the complex relationship between radar backscatter and soil moisture from a limited number of sample points. It is imperative to consider vegetation characteristic such as NDVI; which is used as a substitute for the directly measured LAI and VWC, due to its strong correlation with those variables [3, 4].

Due to the complex relationship between SAR backscatter and soil moisture, the non-parametric tools like neural
networks and fuzzy logic were used to empirically ascertain the statistical relationship between soil moisture and radar backscatter. These non-parametric tools are considered alternatives to the classical modeling techniques for hydrological and meteorological applications; these non-parametric tools exploit the statistical relationships between hydrologic inputs and outputs without explicitly considering the physical process relationships that exist between them [5].

This paper applies the techniques of Multiple Regression Analysis, Neural Network, and Fuzzy Logic to spatial soil moisture estimation using active microwave data along with soil characteristics and vegetation data. An attempt has been made to: (1) optimize the neural network internal parameters; (2) select the best input parameters to have significant improvement in soil moisture retrieval, (3) use the fuzzy logic technique to retrieve soil moisture via similar variables used in the neural network method, and (4) evaluate the effect of NDVI on the retrieval accuracy of soil moisture.

2 Study Area and Data Sets

The study area was selected based on the availability of intensive soil moisture measurements collected in 1997 during the Southern Great Plain Mission (SGP97). This mission was a large, interdisciplinary field campaign performed over a one month period (18 June–17 July) with the objective of testing previously established passive microwave - Electronically Scanned Thinned Array Radiometer (ESTAR) - based soil-moisture retrieval algorithms [4]. The soil moisture data, retrieved from L-band ESTAR instrument during the SGP97 experiment, were used in this study. The technical details about the instruments and the methodology used in soil moisture field measurement can be found in [4].

Gravimetric soil moisture measurements of the surface layer have been taken from 50 sites distributed in Central facility (CF), El Reno (ER), and the Little Washita (WS) area Oklahoma USA. For each sampling site an average of 14 samples of soil moisture were measured. Detailed information about the SGP97 mission, the geographical coordinates and the methodology used to measure in situ soil moisture as well as other soil and land cover parameters can be found in Jackson et al. (1999).

Soil classification data were acquired from the state soil geographic database (STATSGO) of the USDA Natural Resources Conservation Service (NRCS), on a 1 km grid; the data were further re-sampled to 800 meters to match ESTAR resolution. The NDVI data were obtained from one Landsat TM image acquired on July 25, 1997. The NDVI values were originally calculated at 30 meter resolution, and then aggregated to 800 meter resolution to match the soil moisture resolution. The active microwave, Synthetic Aperture Radar (SAR) backscatter data from RADARSAT-1 satellite was chosen for this study. Two RADARSAT-1 images taken in SCANSAR Narrow Mode were acquired for July 2nd and July 12th, 1997. Two regions (A and B) were selected within the study area for both the images.

3 Methodology

3.1 Multiple Regression Analysis

The stepwise regression method is used to generate the multivariable model. The data fitting in stepwise regression is tested by correlations and an overall test of significance. The correlations are actually values of R² for the observed values versus predicted values. The test of significance is done by: (1) a standardized regression coefficient (b if all variables are standardized), (2) a t value, and (3) a p value associated with the t value. The standardized coefficient equates the value of r between the variable of interest and the residuals from the regression. The general multivariable model is expressed as:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \ldots + c \]  

where the ß's are the regression coefficients; they express the amount to which the dependent variable y changes with the change in 1 unit of the corresponding independent variable. The \( \beta_0 \) is the constant intercept of the regression line at the y axis. The ratio of the beta coefficients determines the relative predictive power of the associated independent variables [6].

3.2 Neural Network Model

Neural networks have been applied to a wide range of problems in remote sensing and have produced improved accuracy when compared to traditional statistical methods. Neural networks have been increasingly used for the classification of remotely sensed images [5]. A neural network is a highly interconnected system of simple processing elements (called nodes) designed to mimic the highly parallel human biological neurons. These nodes are usually organized into a sequence of fully connected layers. The strength of these connections is given through the connecting weights of the network. Each node calculates a summation of weighted inputs and then outputs its transfer function value to other nodes. There
are three main phases in the operation of a network: training, validation and testing.

3.2 Fuzzy Logic Model
The basic principle of fuzzy set theory was formulated in linguistic form by Dr. Lotfi Zadeh in 1965. This linguistic approach is an approximate and effective means of describing the behavior of systems that are imprecise and vague, and too complex to be analyzed with precise mathematical approaches. A strong advantage of the fuzzy approach over traditional methods is that it does not require a detailed mathematical description of the system. A typical Fuzzy system includes the processes of fuzzification, inference system, and defuzzification. Fuzzification involves the process of transforming crisp values of image data into grades of membership in linguistic terms. Defuzzification involves the process of transposing the fuzzy crisp outputs in the form of image data [7].

4 Results and Discussion

4.1 Input Parameter Selection
The potential inputs that affect soil moisture retrievals can be better understood by techniques like neural network, fuzzy logic, and multiple regression models where more than one input can be automatically weighted and used to improve the model retrieving capabilities. The identification of important and redundant inputs is an important issue that leads directly to reduced size, faster training, and possibly more accurate results from the neural network, fuzzy logic and multiple regression models. Initially, each input is individually used to train the models and then the mean square error of soil moisture retrieval is analyzed. The model trained with SAR backscatter was able to predict soil moisture with lower RMSE (4.857) than the model trained with other individual inputs. The models were then trained in combination with two or more inputs such as SAR textural images, VOD, VWC, NDVI, and soil characteristics. The results show that the soil moisture retrieval process is sensitive to the vegetation and soil characteristics (percent of sand). When NDVI was added to the SAR backscatter in the input layer, the accuracy in the soil moisture estimate improved. However, no significant improvement was observed by adding VOD and VWC to an already existing NDVI in the neural network input layer. This can be explained by the indirect relationship between VOD and VWC with NDVI [4].

Based on thorough analysis of several runs of the neural network, fuzzy logic and multiple regression models, SAR backscatter, NDVI and soil characteristics, were determined to be the three critical inputs that affected the retrieval process the most. The total available in-situ soil moisture measurements corresponding to the active microwave data taken on 2nd and 12th July (76 data points) were used to train the models. However, we found that these data points were not enough to train the neural network and fuzzy logic models. Therefore, we used 500 data points were used from soil moisture data derived from using ESTAR instead of the in-situ soil moisture measurements.

4.2 Finding Model using Multiple Regression Analysis
The basic procedures in stepwise regression analysis involve (A) identifying an initial model, (B) iteratively altering the initial model by adding or dropping an independent variable in agreement with the "significant test criteria", and (C) terminating the search when stepping is no longer possible given the significant test criteria, or when a specified maximum number of steps have been reached [6]. The significance test is performed using the $p$-values to compare the effect of different variables on regression analysis. It is necessary to run a stepwise procedure a number of times using random selection of data to find meaningful patterns. The model coefficients were estimated based on numerous runs (100) of the stepwise regression analysis with different datasets. Thus equation below was derived, where volumetric soil moisture ($M_v$) is:

$$M_v (%) = 0.313 \ (\sigma_0) + (4.471 \ * \ \text{NDVI}) - 8.50 \ * \ \text{PS}$$

Where, $\sigma_0$ is SAR backscattering in digital numbers and PS is the percent of sand.

4.3 Neural Network Training
The training stage consists of adjusting the connection weights (randomly initialized) in order to decrease the difference between the network output and the desired output. The training data were presented to the input layer and propagated through the hidden layers to the output layer. The differences between the neural network outputs and the desired outputs were computed and fed backwards to adjust the network connections. This iterative process continued until the mean square error reached a preset goal or when the validation criteria were reached. When one of the two criteria is met, the training is stopped and the weight values saved. The trained network is then used to simulate new data.
Often, the increase in the number of training pixels increases the training time, so it is necessary to find out the optimum size of the training set. Furthermore, the training data must represent the entire range of values associated with a particular class [8]. After several successive runs of the same network, the analysis showed that by increasing the number of training pixels, However, once the size of training data reaches maturity, no significant increase in accuracy was observed. When a single hidden layer is used, the number of nodes should be greater than the number of input data layers to get reliable results. Based on this optimization process for these datasets, it was determined that there was no apparent advantage for using multi-hidden-layer networks instead of single-hidden-layer networks for our data. Therefore, a single hidden layer network structure was used to predict the soil moisture. The network architecture 3:10:1 (Input layer: Hidden nodes: Output layer) was used to train the network with normalized values of soil moisture.

4.4 Fuzzy Logic Training
A subtractive clustering-based fuzzy identification method and a Sugeno type fuzzy inference system were used to retrieve of soil moisture from SAR backscatter data. The fuzzy logic model has been applied to various potential input parameters such as SAR backscatter data, NDVI, and soil characteristics. Based on several runs of the fuzzy logic model, the retrieval process was found to be very sensitive to SAR backscatter, NDVI and soil characteristics. The RMSE errors in soil moisture retrievals were minimized through an exhaustive search of clustering parameters. The accuracy of the fuzzy model depends on the number of the rules that are applied as well as to the number clustering parameters.

4.5 Comparison of Results
The neural network, fuzzy logic and multiple regression models were trained by using combinations of inputs; SAR backscatter, soil characteristics, NDVI data. The ESTER derived soil moisture data (500 pixels) from study area A of the July 12 image were used in output layer to train the model. The models were tested for areas A and B for SAR images taken on July 02 and July 12. Independent pixels that were not used in training the model were used as tests for Area A of 12th July. Same data were also used to determine multivariable model by using multiple regression analysis as previously outlined. The soil moisture data that were predicted by the neural network, fuzzy logic and multiple regression models were compared with two independent data sets: ESTAR derived soil moisture and field soil moisture measurements (FIGURE 1).

The RMSE between ESTAR derived soil moisture and the neural network, fuzzy logic and multiple regression model outputs for independent soil moisture pixels are given in TABLE 1. Lower RMSE of soil moisture for all the area has been observed for fuzzy logic prediction compared to the neural network model. The RMSE for area A of 12th July is found to be smaller than that for Area B and 2nd July data. The correlation coefficient between the neural network and the fuzzy logic predicted soil moisture and the ESTAR derived soil moisture varies between 0.62% and 0.77%.

Soil moisture values predicted by the neural network and the fuzzy logic model compared to ESTAR soil moisture and in-situ soil moisture measurements at field sites are shown in FIGURE 1. The RMSE of the predicted and the in-situ soil moisture measured at 38 field sites locations, in terms of soil moisture percentage, using neural network and fuzzy logic are 6.44 and 6.97 respectively. Lower RMSE was observed for Washita area (LW) compared to El Reno and Central Facility of study area.

4.6 Effect of Vegetation and Soil Characteristics
The impact of addition of vegetation information (in terms of NDVI) to neural network, fuzzy logic and multiple regression models in retrieval of soil moisture were given in TABLE 2. The addition of NDVI along with SAR backscatter to the neural network models reduced the RMSE error by 18% and 9.5% using fuzzy logic model. The largest impact was observed using multiple regression models reduced RMSE by 27%. However, RMSE of predicted soil moisture were lower using neural network and fuzzy logic models.

Soil texture is the relative composition of the three major soil classes: sand, silt and clay. Previous studies showed a strong linear correlation between the backscatter and soil moisture at a particular soil texture. The SAR backscattering increases when the clay content of the soil decreases at any given value of soil moisture [9]. In this study, the addition soil characteristic (in terms of percent of sand) were improved the soil moisture prediction. The addition of soil characteristic along with SAR backscatter to the neural network models reduced the RMSE error by 10% and 12% using fuzzy logic model. The addition of vegetation and soil characteristic along with SAR backscatter to these models reduced the RMSE by 30%, 23% and 39% using neural network, fuzzy logic and multiple regression models respectively.

Based on the comparison of the soil moisture output, the analysis showed that areas with low NDVI values have
lower RMSE than areas with higher NDVI. This can be explained by the decrease of vegetation contribution to the backscatter. The RMSE error observed with field point measured soil moisture was higher than that observed with ESTAR based soil moisture data. This is due to variability of class covers within a pixel, which generate additional errors when soil moisture point measurements are converted to spatial maps. With the magnitude of heterogeneity typically observed in surface soil moisture fields, and with the uncertainty associated with gridded point-scale observations mapped to space-borne sensor footprint scales, it is difficult to correlate the soil moisture values [10-12]. Additionally, soil moisture retrieval is highly influenced by heterogeneity of surface land cover. The addition of vegetation data to the neural network and to the fuzzy logic models is important to improving the accuracy of dry soil moisture pixels estimates, where the dominance of VWC can be greater than soil water content.

Acknowledgments
This study was supported and monitored by National Oceanic and Atmospheric Administration (NOAA) under Grant NA06OAR4810162. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official National Oceanic and Atmospheric Administration or U.S. Government position, policy, or decision.

References

<table>
<thead>
<tr>
<th>Date</th>
<th>Area</th>
<th>Neural Network Model</th>
<th>Fuzzy Model</th>
<th>Logic Model</th>
<th>Multiple Regression Model</th>
<th>Regression Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>R</td>
<td>RMSE</td>
<td>R</td>
<td>RMSE</td>
</tr>
<tr>
<td>2nd July</td>
<td>A</td>
<td>7.967</td>
<td>0.414</td>
<td>5.763</td>
<td>0.493</td>
<td>6.696</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>8.294</td>
<td>0.397</td>
<td>6.195</td>
<td>0.448</td>
<td>5.834</td>
</tr>
<tr>
<td>12th July</td>
<td>A*</td>
<td>3.396</td>
<td>0.767</td>
<td>3.454</td>
<td>0.759</td>
<td>4.063</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>4.493</td>
<td>0.458</td>
<td>3.853</td>
<td>0.504</td>
<td>4.822</td>
</tr>
</tbody>
</table>
TABLE 2: Effect of data input configuration used in neural network, fuzzy logic and multiple regression model in terms of Root means square error (RMSE) and correlation coefficient (R) values of predicted soil moisture and ESTAR soil moisture for independent 300 pixels from Area A on 12th July data.

<table>
<thead>
<tr>
<th>Data Input</th>
<th>Neural Network Model</th>
<th>Fuzzy Logic Model</th>
<th>Multiple Regression Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>R</td>
<td>RMSE</td>
</tr>
<tr>
<td>SAR</td>
<td>4.847</td>
<td>0.620</td>
<td>4.506</td>
</tr>
<tr>
<td>SAR+NDVI</td>
<td>3.940</td>
<td>0.716</td>
<td>4.075</td>
</tr>
<tr>
<td>SAR+PS</td>
<td>4.344</td>
<td>0.660</td>
<td>3.955</td>
</tr>
<tr>
<td>SAR+NDVI+PS</td>
<td>3.396</td>
<td>0.767</td>
<td>3.454</td>
</tr>
</tbody>
</table>

FIGURE 1. Comparison of predicted soil moisture from fuzzy logic model with field and ESTAR soil moisture with field soil moisture measuring area: Central Facility (CF), El Reno (ER), and Little Washita (LW) for July 02nd (a), July 12th (b) data.