# *Comparison of Data Driven Models (DDM) for Soil Moisture Retrieval using Microwave Remote Sensing Data*

*Liauw Hephi*
of *Faculty of Computer Science and Information Technology*
University Malaysia Sarawak
Kuching, Sarawak, Malaysia
lhephi@gmail.com

*Chai Soo See*
of *Faculty of Computer Science and Information Technology*
University Malaysia Sarawak
Kuching, Sarawak, Malaysia
sschai@fit.unimas.my

*Abstract*—**This paper aims to explore the use of various DDM methods for soil moisture retrieval, identifying the advantages and disadvantages of each, compare and evaluate the results for further study. The study looks into the advantages and disadvantages of each DDM method, summarizing the Root-Mean-Square-Error (RMSE) to identify soil moisture condition. In this study, Neural Network Model, Fuzzy-Rule Model, Bayesian Model, Multiple Regression Model and Support Vector Machines (SVM) were reviewed.**

**The Neural Network model performed better compared with other models, proven with the lowest number of RMSE. The SVM model also showed high potential, whereas the Bayesian, Multiple Regression and Fuzzy-Rule Based models showed higher RMSE values, which indicate higher difference in accuracy.**

***Keywords— Data Driven Modelling (DDM,) Neural Network Model, Fuzzy-Rule Model, Bayesian Model, Multiple Regression Model and Support Vector Machines (SVM), Root-Mean-Square-Error (RMSE)***

## I. INTRODUCTION

Soil moisture values derived from remote sensing platforms only accounts for the top 5cm near surface soil layers. The measuring of soil moisture using remote sensing is dependent on a relationship between the remote sensing parameter and soil moisture. Microwave remote sensing measurements (both active (radar) and passive (radiometry)) of bare soil surface are very sensitive to the water content in the surface layer due to the pronounced increase in the soil dielectric constant with increasing water content. Various theoretical and empirical models have been devised to retrieve soil moisture from data. Recent advancements entitled Data Driven Modeling (DDM) which encompasses computational intelligence has emerged. DDM is based on analyzing the data about the system, in particular finding connections between the system variables (input, internal and output variables) without explicit knowledge of the physical behavior of the system.

This paper aims to identify and explore the use of various DDM techniques for soil moisture retrieval using active and passive microwave data, evaluates the strengths and weaknesses of different DDM techniques for soil moisture retrieval, compare and evaluate the result for further study. The common DDM techniques evaluated in this study cover Neural Network Model, Fuzzy-Rule Model, Bayesian Model, Multiple Regression Model, and Support Vector Machines (SVM).

## II. THEORETICAL BACKGROUND

Soil moisture retrievals from microwave remote sensing techniques have the potential for efficient and reliable mapping of spatial soil moisture distributions. However, the techniques used are typically complex due to inherent difficulty in characterizing the interactions among land surface parameters that contribute to the retrieval process. Therefore, adequate physical mathematical descriptions of microwave radiation interaction 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 methods incorporating DDM techniques to retrieve soil moisture distributions.

In this study, common DDM techniques for soil moisture retrieval using microwave remote sensing data are identified, studied and reviewed. Each of these DDM used different input parameters. Neural Network Model, for example, takes into account the land cover, soil texture, soil temperature, and canopy temperature gathered from ancillary data [24]. The Fuzzy Logic Model looks into active microwave SAR data, soil moisture, NDVI, Vegetation Water Content (VWC), Vegetation Optical Depth (VOD), SAR textural images (homogeneity, contrast, dissimilarity, mean, variance, entropy, angular second moment, and correlation), and soil texture (percent of sand) [1]. Bayesian Model takes input parameters such as SD for height, correlation length and dielectric constant were considered for the training purposes. The Multiple Regression Model looks into backscatter, soil characteristics, and Normalized Difference Vegetation Index (NDVI) [1], [3]. SVM Model trains data set consisting of soil moisture and meteorological data (relative humidity, average solar radiation, soil temperature at 5 cm and 10 cm, air temperature, and wind speed).

The expected soil moisture accuracy is indicated by RMSE – a frequently used measure of the differences between values (in this case, data set studied by each DDM model) predicted and the values actually observed (in this case, the ground measurement). RMSE represents the sample standard deviation of the differences between predicted values and observed values. It serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSE is a good measure of accuracy and is scale-dependent. The accuracy of soil moisture condition using microwave remote sensing data is at RMSE value of 4%. Higher values indicate lower level of accuracy, whereas lower values indicate higher level of accuracy in identifying soil moisture condition.

In this paper, the five commonly used DDM techniques, namely Neural Network Model, Fuzzy-Rule Model, Bayesian Model, Multiple Regression Model, and Support Vector Machines (SVM) are identified. The different type of data input, strengths and weaknesses are compared to support future study in this domain.

## III. COMMON DDM USED

### *A. Neural Network Model*

Neural network is a highly interconnected system of simple processing elements (nodes) designed to mimic the highly parallel human biological neurons [1]; usually organized into a sequence of fully connected layers [2]. TWO (2) phases involved: training and validation: training is the process of adapting the connection weights in response to the training data presented at the input layer and the desired response at the output layer; whereas validation is then used to assess the performance of the trained neural network via the input vector and the created response at the output layer [1]. Prediction made by neural network is high in accuracy, validated by the ground measurements [5], [10]. The prediction has an ability to identify subtle and nonlinear patterns [5]. Performance is not affected significantly by the variation of the architecture configuration [1]. Another strong point of the neural network model is that it does not require normally distributed continuous data and may be used to integrate data from different sources with poorly defined or unknown distributions [5].

Neural network model is able to take a specific set of input data and generalize a solution set [10]. Better correlation between soil moisture and SAR backscattering was found in areas with high soil moisture content. Retrieval of soil moisture in highly vegetated areas was less accurate than bare soil areas [4]. Neural network model shows higher potential to estimate soil moisture. The inverse relationship between inputs and outputs was established in the training phase when the inputs were presented [1].

The results of the study, identified from the experiments conducted, showed diverse outcomes in RMSE value with the lowest RMSE at the value of 3.4% [24] as per study conducted by Chai *et al.* [24] with a 2-day experiment conducted in the northern part of the Goulburn River catchment, located in a semiarid area of south-eastern Australia. This catchment extends from 31°46'S to 32°51'S and 149°40'E to 150°36'E with elevations ranging from 106 m in the floodplains to 1,257 m in the northern and southern mountain ranges [24]. The input parameters carried out in the study conducted by Chai *et al.* [24] consist of land cover, soil texture, soil temperature, and canopy temperature gathered from ancillary data [24].

### *B. Fuzzy Logic Model*

Fuzzy logic model makes use of approximate and effective linguistic approach; it has the capability to describe the behavior of systems that are imprecise and vague, and too complex to be analyzed with precise mathematical approaches [1]. Fuzzy systems provide a computational framework in which linguistic knowledge is expressed in the form of fuzzy IF-THEN rules [15], [16]. 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. The fuzzy inference system is based on the concepts of fuzzy set theory, fuzzy if-then rules, and fuzzy reasoning. Defuzzification involves the process of transposing the fuzzy crisp outputs in the form of image data [1]. The model is suitable and able to deal with uncertainty and imprecision in a decision-making process, and offers a new approach for classifying remotely sensed images [17]. Estimation of soil moisture is more balanced [1], [17]. Better correlation between soil moisture and SAR backscattering was found in areas with high soil moisture content. Retrieval of soil moisture in highly vegetated areas was less accurate than bare soil areas [2].

The data used for this study governs active microwave SAR data, soil moisture, NDVI, Vegetation Water Content (VWC), Vegetation Optical Depth (VOD), SAR textural images (homogeneity, contrast, dissimilarity, mean, variance, entropy, angular second moment, and correlation), and soil texture (percent of sand) [1].

Prediction made by fuzzy logic is more stable in nature [1], [2]. The lowest RMSE and correlation coefficient (R) of the model was at the value of 4.82% and 0.55 respectively, as identified by Lakhankar *et al.* [1] from a study area located at Oklahoma, USA (97°35'W, 36°15'N). This area was selected based on the availability of soil moisture measurements collected in 1997 during the Southern Great Plain Mission (SGP97) [1] [2].

### *C. Bayesian Model*

Bayesian model is a statistically based inversion method [8]. Results were based on probabilities that a given set of measurements comes from certain surface parameter values. The probability density functions (pdfs) are estimated by training, where samples of sensor and surface measurements are presented in the algorithm [5]. The result turns probabilities that can be estimated from a training set into those that are required for the estimation of the unknown surface parameters [13]. The model relies on limited surface parameter conditions. It is optimal in minimizing expected errors, although large amount of experimental data is needed [5], [8], [13].

For the training, input parameters such as SD for height, correlation length and dielectric constant were considered.

Errors for the Bayesian method increase as the number of input increase [5], [8]. Results of studies indicate lowest values of soil moisture with RMSE of 4.2% based on the study performed by Notarnicola *et al.* [5]. Notarnicola *et al.* [5] took experimental data sets acquired by the University of Bern's truck-mounted radiometer–scatterometer that operates with the following frequencies: 2.5, 3.1, 4.6, 7.2, 10.2, and 11 GHz over the incidence angle range of 10◦–70◦ [22].

### D. Multiple Regression Model

The multiple regression model is used to generate the multivariable model [9]. Basic procedures in multiple 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 [1]. Combinations of input in the study conducted by Lakhankar *et al.* [1] and Njoku *et al.* [3] consist of backscatter, soil characteristics, and Normalized Difference Vegetation Index (NDVI) [1], [3].

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 [1], [3]. It may not be suitable for the feature estimation applies to the data acquired on other sites [11], [12]. Results indicate values of soil moisture with lowest RMSE of 4.2% based on the study conducted by Lakhankar *et al.* [1] with the study area located at Oklahoma, USA (97°35'W, 36°15'N). This area was selected based on the availability of soil moisture measurements collected in 1997 during the Southern Great Plain Mission (SGP97) [1] [2].

### E. Support Vector Machine (SVM) Model

SVM represents a learning paradigm where prediction error and model complexity are simultaneously minimized. The structure of the SVM model is not fixed in advance with a specific number of adjustable parameters, but can adapt with the data. SVMs have 3 distinct characteristics when applied to estimate the regression function: (1) SVMs estimate the regression using a set of linear functions that are defined in a high dimensional space, (2) SVMs carry out the regression estimation by risk minimization, (3) SVMs use a risk function consisting of the empirical error and a regularization term which is derived from the structure risk minimization principle [10]. SVM algorithm has a remarkable prediction capacity and it performed better than polynomial and rational approximations, local polynomial techniques, radial basis functions, and feed-forward artificial neural networks when applied on a database of chaotic time series [19].

The study for this model showed diverse result in RMSE values, ranging from 3.65% [20] to more than 5% [10]. Some papers reviewed showed error greater than 5% [10]; however the lowest RMSE value was achieved at the value of 3.65%, as studied by Khalil *et al.* [20]. The plausibility of this model is evaluated using the data from Soil Climate Analysis Network (SCAN). The Soil Climate Analysis Network (SCAN) was created by the US Department of Agriculture with a purpose to provide soil and climate data bank. There are more than 90 SCAN stations all over the US, at which daily and hourly measurements for meteorological and soil moisture data are measured using various sensors and instruments. The data in this study is taken from the SCAN site located at the Little Washita River Experimental Watershed (LWREW) in Southwestern Oklahoma in the Southern Great Plains region of the United States [1]. The SVM model in this study are trained to predict soil moisture at time step 't+5' (t is in days), taking into account parameters such as soil moisture and meteorological data (relative humidity, average solar radiation, soil temperature at 5 cm and 10 cm, air temperature, and wind speed) at time steps 't-l' and 't'. The output is the soil moisture value at 't+5'.

## IV. RESULT OF STUDY

The expected soil moisture accuracy is indicated by RMSE at the value of 4%. Higher values indicate lower level of accuracy, whereas lower values indicate higher level of accuracy in identifying soil moisture condition.

| DDM | Advantages | Disadvantages | RMSE |
|---|---|---|---|
| Neural networks | Able to identify subtle and nonlinear patterns. Performance is not affected significantly by the variation of the architecture configuration. Able to take a specific set of input data and generalize a solution set. | Retrieval of soil moisture in highly vegetated areas was less accurate than bare soil areas. More inputs result in high variation in accuracy for successive runs due to more variations. | 3.4% |
| Fuzzy-rule based system | Estimation of soil moisture is more balanced (low variation in accuracy). Prediction is stable in nature. | Performance depends significantly on the cluster radius selection. Retrieval of soil moisture in highly vegetated areas was less accurate than bare soil areas | 4.82% |
| Bayesian model | Optimal in minimizing expected errors, although large amount of experimental data is needed. | Errors increase as the number of inputs increase. | 4.2% |
| Multiple regression model | Prediction is more accurate in areas with low NDVI values than areas with higher NDVI. | Performance depends significantly on NDVI. | 4.2% |
| Support vector machine | Represents a learning paradigm where prediction error and model complexity are simultaneously | Diverse results gathered from studies. | 3.65% - 5% |

| | | | |
|---|---|---|---|
| | minimized. The model structure is not fixed in advance with a specific number of adjustable parameters, but can adapt with the data. | | |

Table 1. Result of study

In identifying soil moisture retrieval using microwave remote sensing data by taking account the RMSE, the result of the study identified the lowest level of RMSE achieved by the Neural Network model. Although results vary as more input were gathered and examined, the model was capable of showing the lowest level of RMSE as compared with other models. The expected soil moisture accuracy of 4% was achieved with the result of study indicating lowest RMSE at the value of 3.4%. However, study also showed higher RMSE values when different input parameters were used and analyzed. Inconsistency might depend on the input parameters, when the same model is concerned.

The Support Vector Machine (SVM) showed high potential in identifying the soil moisture values. The model was capable of indicating soil moisture accuracy of 4% at some points. Although as more input values were gathered, processed and examined; the more varied the output values turned out to be. The RMSE value ensued by the model range in between 3.65% to 5%.

The Bayesian and Multiple Regression models showed average performance at 4.2% RMSE value, whereas the Fuzzy-Rule Based model showed the highest RMSE value of 4.82%, which indicates highest difference in accuracy.

Additional note on the study showed that soil moisture retrievals from highly vegetated areas are less accurate than that from bare soil areas.

All of the DDM models tested and analyzed in this study used different data set. The use of the same data set over different range of DDM models has yet to be studied. There has not been any research studying the optimal soil moisture retrieval model. The effects of active and passive or assimilation of these two types of microwave data on different DDM models were also yet to be explored. Therefore, a proceeding study would look into a repository of data widely used in soil moisture retrieval and processing it using different DDM models. At the end of the proceeding study, an optimal DDM model will be developed for soil moisture retrieval using microwave data, focusing on either one or assimilation of both of the types of microwave data using the same data set.

## V. CONCLUSION

Data Driven Modeling (DDM) which encompasses computational intelligence in measuring soil moisture using microwave remote sensing has been accurate in identifying soil moisture values. DDM is based on analyzing the data about the system, in particular finding connections between the system variables (input, internal and output variables) without explicit knowledge of the physical behavior of the system. In this paper, various DDM methods for soil moisture retrieval were identified, explored, compared and evaluated.

The study in this paper evaluates Neural Network Model, Fuzzy-Rule Model, Bayesian Model, Multiple Regression Model, and Support Vector Machines (SVM). The result of the study identifies the strengths and weaknesses of different DDM techniques for soil moisture retrieval based on the RMSE value obtained from the models examined.

Based on the result of study shown in Table 1, in identifying the soil moisture condition using microwave remote sensing data by taking account the RMSE, the Neural Network model shows the best result in identifying soil moisture accuracy. The Support Vector Machine (SVM) model shows high potential, although results vary as more input values were gathered. The Bayesian and Multiple Regression models showed average performance, whereas the Fuzzy-Rule Based model showed the highest RMSE value, which indicates highest difference in accuracy. Additionally, the study showed that soil moisture retrievals from highly vegetated areas are less accurate than that from bare soil areas.

## *Acknowledgment*

This project is registered under Exploratory Research Grant Scheme (ERGS) of Ministry of Higher Education (MOHE) Malaysia with the project no. ERGS/STWN06(01)/1016/2013(13).

## *References*

[1] Tarendra Lakhankar, Hosni Ghedira, Marouane Temimi, Manajit Sengupta, Reza Hanbilvardi, and Reginald Blake, “Non-parametric Methods for Soil Moisture Retrieval from Satellite Remote Sensing Data”, International Journal of Systems Applications, engineering and Development Issus 4, Volume 2, 2008.

[2] Tarendra Lakhankar, Hosni Ghedira, and Reza Khanbilvardi, “Soil Moisture Retrieval from Radarsat Data: A Neuro-Fuzzy Approach”, NOAA-CREST, City University of New York, New York, NY 10031, 2006.

[3] Eni G. Njoku, Fellow, IEEE, Thomas J, Jackson, Fellow, IEEE, Venkataraman Lakshmi, Senior Member, IEEE, Tsz K. Chan., Member, IEEE, and Son V. Nghiem, Senior Member, IEEE, “Soil Moisture Retrieval From AMSR-E”, IEEE Transactions on GeoScience and Remote Sensing, Vol. 41, No. 2, February 2003.

[4] Claire Gruhier, Patricia de Rosnay, Stefan Hasenauer, Thomas Holmes, Richard de Jeu, Yann Kerr, Eric Mougin, Eni Njoku, Frank Timouk, Wolfgang Wagner, and Mehrez Zribi, “Soil moisture active and passive microwave products: intercomparison and evaluation over a Sahelian site”, Hydrol. Earth Syst. Sci., 18 December 2009.

[5] Claudia Notarnicola, Mariella Angiulli, and Franzesco Posa, Member, IEEE, “Soil Moisture Retrieval From Remotely Sensed Data: Neural Network Approach Versus Bayesian Method”, IEEE Transactions on GeoScience and Remote Sensing, Vol. 46, No. 2, February 2008.

[6] Dara Entekhabi, Eni G. Njoku, Peggy E. O’Neill, Kent H. Kellogg, Wade T. Crow, Wendy N. Edelstein, Jared K. Entin, Shawn D. Goodman, Thomas J. Jackson, Joel Johnson, John Kimball, Jeffrey R. Piepmeier, Randal D. Koster, Neil Martin, Kyle C. McDonald, Mahta Moghaddam, Susan Moran, Rolf Reichle, J. C. Shi, Michael W. Spencer, Samuel W. Thurman, Leung Tsang, and Jakob Van Zyl, “Soil Moisture Active Passive (SMAP) Mission”, Proceedings of the IEEE, Vol. 98, No. 5, May 2010.

[7] A. Al-Yaari, J. -P. Wigneron, A. Ducharne, Y.H. Kerr, W. Wagner, G. De Lannoy, R. Reichle, A. Al Bitar, W. Dorigo, P. Richaume, A. Mialon, “Global-scale comparison of passive (SMOS) and active (ASCAT) satellite based microwave soil moisture retriavals with soil

moisture simulations (MERRA-Land)", Remote Sensing of Environment 152 (2014) 614-626.

[8] J.-P. Wignerona, J.-C. Calvetb, T. Pellarinb, A.A. Van de Griendc, M. Bergerd, P. Ferrazzolie, "Retrieving near-surface soil moisture from microwave radiometric observations: current status and future plans", Remote Sensing of Environment 85 (2003) 489–506.

[9] Thomas R. H. Holmes, Wade T. Crow, M. Tugrul Yilmaz, Thomas J. Jackson, and Jeffrey B. Basara, "Enhancing model-based land surface temperature estimates using multiplatform microwave observations", Journal if Geophysical Research: Atmospheres, Vol. 118, 577- 591, doi:10.1002/jgrd.50113, 2013.

[10] Hongbin Liu & Deti Xie & Wei Wu, "Soil water content forecasting by ANN and SVM hybrid architecture", Springer Science + Business Media B.V. 2007.

[11] L. Prevot, M. Dechambre, et al., "Estimating the characteristics of vegetation canopies with airborne radar measurements," Int. J. Remote Sens., pp. 2803-2818, vol 14, no. 15, 1993.

[12] D. S. Lin, E. F. Wood, K. Bevan, and S. Saatchi, "Soil moisture estimation over grass-covered areas using AIRSAR," Int. J. Remote Sens., pp. 2323-2343, vol 15, no. 11, 1994.

[13] J. A. Marchant and C. M. Onyango, "Comparison of a Bayesian classifier with a multilayer feed-forward neural network using the example of plant/weed/soil discrimination," Comput. Electron. Agric., vol. 39, no. 1, pp. 3-22, Apr. 2003.

[14] F. T. Ulaby, R. K. Moore, and A. K. Fung, "Microwave Remote Sensing: Active and Passive", vol. II., Norwood, MA: Artech House, 1986.

[15] S. L. Chiu., "An Effective Method for Extracting Fuzzy Classification Rules from High Dimensional Data", Hournal of Advanced Computational Intelligence, vol. 1, pp. 1-7, 1997.

[16] M. T. Hayajneh and S. M. Radaideh, "Modeling Surface Finish in End Milling Using Fuzzy Subtractive Clustering-Based System Identification Method", Materials and Manufacturing Processes, vol. 18, pp. 653-665, 2003.

[17] I. Nedeljkovic, "Image classification based on fuzzy logic", in The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences", 2004, pp. 74-79.

[18] Salman Nazari-Shirkouhi, Abbas Keramati, Kamran Rezaie, "Improvement of customers' satisfaction with new product design using an adaptive neuro-fuzzy inference systems approach", Neural Comput & Applic, 2013.

[19] Mukherjee, S., E. Osuma, and F. Girosi, "Nonlinear prediction of chaotic time series using support vector machines", Proceedings of IEEE Workshops on Neural Network for Signal Processing, Amelia Island, FL, 1997.

[20] Abedalrazq Khalil, M. Kashif Gill, Mac McKee, "New applications for Information Fusion and Soil Moisture Forecasting", 7th International Conference on Information Fusion (FUSION), 2005.

[21] F. Ulaby, R. Moore, and, A. Fung, "Microwave Remote Sensing, Active and Passive, vol. 3. From Theory to Applications", Artech House, Norwood, MA, 1986.

[22] U. Wegmueller, C. Maeztler, R. Hueppi, and E. Schanda, "Active and passive microwave signature catalog on bare soil (2–12 GHz)," IEEE Trans. Geosci. Remote Sens., vol. 32, no. 3, pp. 698–702, May 1994.

[23] J. Cihlar, M. C. Dobson, T. Schmugge, P. Hoogeboom, A. R. P. Janse, F. Baret, G. Guyot, T. L. Toan, and P. Pampaloni, "Procedures for the description of agricultural crops and soils in optical and microwave remote sensing studies," Int. J. Remote Sens., vol. 8, no. 3, pp. 427–439, 1987.

[24] Soo-See Chai, Jeffrey P. Walker, Oleg Makarynskyy, Michael Kuhn, Bert Veenendaal and Geoff West, "Use of Soil Moisture Variability in Artificial Neural Network Retrieval of Soil Moisture", Remote Sensing 2010, 2, 166-190; doi:10.3390/rs2010166.