

# SOIL WATER CONTENT FORECASTING BY SUPPORT VECTOR MACHINE IN PURPLE HILLY REGION

Wei Wu<sup>1,2</sup>, Xuan Wang<sup>1,2</sup>, Deti Xie<sup>1,3</sup>, Hongbin Liu<sup>1,3,\*</sup>

<sup>1</sup> Chongqing Key Lab of Digital Agriculture, Southwest University, Chongqing, China, 400715

<sup>2</sup> College of Computer and Information, Southwest University, Chongqing, China, 400715

<sup>3</sup> College of Resources and Environment, Southwest University, Chongqing, China, 400715

\* Corresponding author, Address: College of Resources and Environment, Southwest University, 216 Tiansheng Road, Beibei, Chongqing, 400715, P. R.China, Tel:+86-23-68251069, Fax:+86-23-68250444, Email:wuwei\_star@163.com

**Abstract:** Soil water distribution and variation are helpful in predicting and understanding various hydrologic processes, including weather changes, rainfall/runoff generation and irrigation scheduling. Soil water content prediction is essential to the development of advanced agriculture information systems. In this paper, we apply support vector machines to soil water content predictions and compare the results to other time series prediction methods in purple hilly area. Since support vector machines have greater generalization ability and guarantee global minima for given training data, it is believed that support vector machine will perform well for time series analysis. Predictions exhibit good agreement with actual soil water content measurements. Compared with other predictors, our results show that the SVMs predictors perform better for soil water forecasting than ANN models. We demonstrate the feasibility of applying SVMs to soil water content forecasting and prove that SVMs are applicable and perform well for soil water content data analysis.

**Keywords:** support vector machines, soil water content, statistical learning, prediction, forecasting, time series

## 1. INTRODUCTION

Soil water, though very small in volume, provides valuable information for water resources planning and management. Soil water transferring model and prediction are important in agriculture, hydrology, and meteorology. In agriculture, accurate forecasts of future soil water conditions can be helpful in water quality monitoring, irrigation scheduling, and yield forecasting. In hydrology, information about soil water is required for understanding rainfall/runoff generation processes and managing water resources. Similarly, in meteorology, soil water measurements can be helpful for modeling surface/atmospheric interactions. Over the past several years, various attempts have been made to produce soil water content estimates by using different statistical models, such as Artificial Neural Networks (ANNs) (Liou et al., 2001; Baghdadi et al., 2002; Liu et al., 2004; Liu et al., 2003) and Auto Regression (AR) (Liu et al., 2003).

Recently, support vector machines (SVMs), developed by Vapnik and his co-workers (Vapnik et al., 1996), have become a very active research area with machine learning. Motivated by statistical learning theory, SVMs have been successfully applied to solve various problems, among others in data mining, classification, regression, density estimation and times series prediction (Cao et al., 2001; Flake et al., 2002; Mukherjee et al., 1997; Zhang 2003; Vapnik, 1995). SVMs implement the structural risk minimization principle and Vapnik-Chervonenkis (VC) dimension. Based on this principle, SVMs achieve an optimum structure by striking a right balance between the empirical error and the VC-confidence interval. Eventually, this results in better generalization performance than other models (Vapnik et al., 1996). Furthermore, the SVMs deal with non-linear tasks by mapping the input space into high dimensional feature spaces, and then use a kernel function instead of high dimensional inner product. This means that the solution of SVMs is unique, optimal and absent from local minima (Vapnik, 1995; Xian et al., 2005; Yan et al., 2000).

In this paper, we use support vector machines to predict the soil water content and show that SVMs are applicable to soil water prediction and outperforms ANN in purple hilly area. The SVMs are based on statistical learning theory and can be used to predict a quantity forward in time based on the results of "training" that uses past data. This paper is organized in six sections. Section 2 presents the study area and data preparation. Machine learning scheme and kernel function are described in Section 3. Section 4 and Section 5 include the prediction methodology, error measurements and experiment results. Concluding remarks are presented in Section 6.

## 2. SITE AND DATA DESCRIPTION

The study site is located in Southwest University in Chongqing (long. 106°26'E, lat. 30°26'N). The hilly with an area 0.1hm<sup>2</sup> was selected to test the approach presented herein. Climatic conditions are semi-tropical wet with a mean annual temperature of 18.3°C and a precipitation of about 1150.7 mm/year. The soil is classified as purple.

All the data in this study are taken from the Hydrology Experiment. Monitoring has been done about every 5 days, from Dec, 2001 to Apr, 2004. Samples were collected from a depth of 30 to 40 cm, placed in a can, sealed and transported to a lab, where they were weighed before and after oven drying.

## 3. SUPPORT VECTOR MACHINE

### 3.1 Theory of SVMs in regression approximation

Compared to other neural network regressors, there are three distinct characteristics when SVMs are used to estimate the regression function. First of all, SVMs estimate the regression using a set of linear functions which are defined in a high dimensional space. Secondly, SVMs carry out the regression estimation by risk minimization where the risk is measured using Vapnik's  $\epsilon$ -insensitive loss function. Thirdly, SVMs use a risk function consisting of the empirical error and a regularization term which is derived from the structure risk minimization principle.

Given a set of data points  $G = \{(x_i, d_i)\}_i^n$  ( $x_i$  is the input vector,  $d_i$  is the desired value and  $n$  is the total number of data patterns), SVMs approximate the function using the following:

$$y = f(x) = w\phi(x) + b \quad (1)$$

where  $\phi(x)$  is the high dimensional feature space which is non-linearly mapped from the input space  $X$ . The coefficients  $w$  and  $b$  are estimated by minimizing

$$R_{SVMs}(C) = C \frac{1}{n} \sum_{i=1}^n L(d_i, y_i) + \frac{1}{2} \|w\|^2 \quad (2)$$

$$L\epsilon(d, y) = \begin{cases} |d - y| - \epsilon & |d - y| \geq \epsilon \\ 0 & otherwise \end{cases}$$

In the regularized risk function given by Eq.(2), the first term  $C(1/n)\sum_{i=1}^n L\mathcal{E}(d_i, y_i)$  is empirical error (risk), and measured by function  $L\mathcal{E}$ .

The second term  $0.5\|w\|^2$ , is the regularization term.  $C$  is referred to as the regularized constant and is determines the trade-off between the empirical risk and the regularization term. Increasing the value of  $C$  will result in the relative importance of the empirical risk with respect to the regularization term to grow.  $\mathcal{E}$  is called the tube size and it is equivalent to the approximation accuracy placed on the training data points.

To obtain the estimations of  $w$  and  $b$ , Eq.(2) is transformed to the primal function given by Eq.(3) by introducing the positive slack variables  $\zeta_i$  and  $\zeta_i^*$  as follows:

$$\text{Minimize } R_{SVMs}(w, \zeta^{(*)}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n (\zeta_i + \zeta_i^*) \quad (3)$$

$$\text{Subjected to } \begin{aligned} d_i - w\phi(x_i) - b_i &\leq \mathcal{E} + \zeta_i \\ w\phi(x_i) + b_i - d_i &\leq \mathcal{E} + \zeta_i^*, \zeta_i^* \geq 0 \end{aligned}$$

Finally, by introducing Lagrange multipliers and exploiting the optimality constraints, the decision function given by Eq.(1) has the following explicit form Zhang (2003):

$$f(x, a_i, a_i^*) = \sum_{i=1}^n (a_i - a_i^*)K(x, x_i) + b \quad (4)$$

The detail computation procedure can be found in Vapnik (1996) and Flake et al. (2002).

### 3.2 Kernel function

$K(x_i, x_j)$  is defined as kernel function. The value of the kernel is equal to the inner product of two vectors  $X_i$  and  $X_j$  in the feature space  $\phi(x_i)$  and  $\phi(x_j)$ , that is,  $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ . The elegance of using the kernel function is that one can deal with feature spaces of arbitrary dimensionality without having to compute the map  $\phi(x)$  explicitly. Any function satisfying Mercer's condition (Flake et al. 2002) can be used as kernel function. The typical examples of kernel function are the polynomial kernel  $K(x, y) = (x \cdot y + 1)^d$  and the Gaussian kernel  $K(x, y) = \exp(-1/\delta^2(x-y)^2)$  where  $d$  is the degree of polynomial kernel and  $\delta^2$  is the bandwidth of the Gaussian kernel. The kernel parameter should be carefully chosen as it implicitly defines the structure of the high dimensional feature space  $\phi(x)$  and thus controls the complexity of the final solution.

## 4. METHODOLOGY

### 4.1 Prediction methodology

Suppose the current time is  $t$ , we want to predict  $y(t+l)$  for the future time  $t+l$  with the knowledge of the value  $y(t-n)$ ,  $y(t-n+1), \dots, y(t)$  for past time  $t-n$ ,  $t-n+1, \dots, t$ , respectively. The prediction function is expressed as:

$$y(t+l) = f(t, l, y(t), y(t-1), \dots, y(t-n)) \quad (5)$$

### 4.2 Error Measurements

In addition, we examine the soil water content time series of different prediction methods. Relative Mean Errors (RME), Root Mean Square Error (RMSE) and Coefficient of Variation (CV) are applied as performance indices. The computational methods are described as follows:

$$RME = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (6)$$

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

where  $n$ ,  $y$ ,  $\hat{y}$  and  $\bar{y}$  represents the number of test data, the observation value, the predicted value and the average of the observation samples, respectively.

All the mentioned procedures above are carried out in MATLAB 6.5.

## 5. RESULTS

The experiment samples are classified into two groups. The first one includes the soil water content every 5 days, and the monthly average soil water measurements consist in the second group. For the first group, we use data from the first 160 samples as the training set and use the last 21 samples as our testing set. Meanwhile, for the second group, the first 21 samples are selected as the training set and the last 8 samples as the testing set.

SVMs combined with three kinds of kernel functions, that are linear (SVM\_Linear), polynomial (SVM\_Poly) and radius base function (SVM\_Rbf), are applied to soil water content forecast. In addition, the application presented in this paper is also compared to a very well known

machine learning tool used in hydrology, Artificial Neural Networks (ANNs). The ANN model was set up in such a way as to have one output node, one hidden layer, and one input layer. Moreover, a tan sigmoid function was employed in the hidden layer neurons and a linear transfer function was used at the output node. ANNs use a least squares loss function, unlike SVMs, which use an  $\epsilon$ -insensitive loss function as a fitness measure. Moreover, the ANN predictions are not stable and depend on the averages from various network initializations, which may give a different result each time a model is trained. On the other hand, the SVM results are stable and unique.

The experimental results are summarized in Table 1 and Table 2. It is observed that the SVMs regression outperform the ANN predictors, especially for the monthly average soil water content forecasting. The results in Table 2 show that the radius base function SVM predictor reduces relative mean errors, the root mean squared errors and the coefficient of variation than those achieved by the linear, polynomial SVMs and the ANN predictor.

Furthermore, this experiment examines the errors greater than 5% which are produced by SVMs and ANN prediction methods for the different soil water content. For the first group, the results in Table 1 shows that 30.17% portion of total errors produced by linear and polynomial function SVM predictor are greater than 5% whereas radius base function SVM and ANN predictor produce the number of 78.83% and 70.76% to total errors which are over the 5% RME threshold. Moreover, as far as the monthly average soil water content concerned, Table 2 shows that the bad parts (the portion of errors exceed 5%) of the linear, polynomial function SVMs and ANN prediction errors occupy 10.90%, 11.12% and 9.87% of total errors, respectively. However, for the radius base function SVM predictor, there are only 1.33% of the errors belongs to the bad portion.

Table 1. The forecast performance of different models for the soil water content every 5days

<i>Model</i>	<i>RME</i>	<i>RMSE</i>	<i>CV</i>	<i>Proportion</i>
SVM_Linear	1.0573	0.2200	-2.0938	30.17%
SVM_Poly	1.0573	0.2200	-2.0938	30.16%
SVM_Rbf	58.9678	3.2695	-31.1202	78.83%
ANN	52.4897	2.9878	-28.4386	70.76%

Table 2. The forecast performance of different models for the monthly average soil water content

<i>Model</i>	<i>RME</i>	<i>RMSE</i>	<i>CV</i>	<i>Proportion</i>
SVM_Linear	3.7727	0.3023	-3.4011	10.90%
SVM_Poly	3.7790	0.3024	-3.4017	11.12%
SVM_Rbf	1.2057	0.2440	-2.7448	1.33%
ANN	3.2326	0.2891	-3.2523	9.87%

Moreover, the performance indices of the radius base function SVM for the monthly average soil water content forecasting improve remarkably than that of the soil water content every 5 days. It is can be conferred that to some extent the noise in monthly average samples reduce to less than those of every 5 days.

## **6. CONCLUSION**

Support vector machine and support vector regression have demonstrated their success in time-series analysis and statistical learning. However, little work has been done for soil water content analysis. Prior knowledge of soil water content behavior can not only help in better management and understanding of hydrological systems but also result in improved forecasting, especially for agricultural basins. In this paper we examine the feasibility of applying support vector regression to soil water content time series prediction in purple hilly area. The application presented here uses measured soil water data to predict future soil water. After numerous experiments, we propose a set of SVR parameters that can predict soil water content time series very well. The results show that the SVM predictor significantly outperforms the other baseline predictors. This evidences the applicability of support vector regression in soil water content analysis.

## **ACKNOWLEDGEMENTS**

Authors are thankful to Chongqing key lab of digital agriculture, Southwest University for financial help. We are also thankful to the anonymous referees for their helpful comments in improving the manuscript.

## **REFERENCES**

- Baghdadi N, Gaultier S, King C. Retrieving Surface Roughness and Soil Moisture From Synthetic Aperture Radar (SAR) Data Using Neural Networks. *Canadian Journal of Remote Sensing*, 2002, 5: 701-711
- Cao L J, Tay F E H. Financial forecasting using support vector machines. *Neural Comput.* 2001, Appl(10): 184-182
- Flake G, Lawrence S. Efficient SVM regression training with SMO. *Machine Learning*. 2002, 1/3: 271-290
- Liou Y A, Liu S F, Wang W J. Retrieving Soil Moisture From Simulated Brightness Temperature by a Neural Network, *IEEE Transactions on Geoscience Remote Sensing*, 2001, 8: 1662-1672

- Liu H B, Wu W, Wei C F, Xie D T. Comparison of autoregression and neural network models for soil water content forecasting, *Transactions of the CSAE*, Vol.19, No.4, 2003, pp. 33-36 (in Chinese)
- Liu H B, Wu W, Wei C F, Xie D T. Soil water dynamics simulation by autoregression models, *Journal of Mountain Science*, 2004, 1: 121-125 (in Chinese)
- Liu H B, Wu W, Wei C F. Study of soil water forecast with neural network, *Journal of Soil and Water Conservation*, 2003, 5: 59-62 (in Chinese)
- Mukherjee S, Osuna E, Girosi F. Nonlinear prediction of chaotic time series using support vector machine. In: *NNSP'97: Neural Networks for Signal Processing VII: Proceedings of the IEEE Signal Processing Society Workshop*, Amelia Island, FL, USA. 1997, 243-254
- Thissen U, van Brackel R. de Weijer A P, Melssen W J, Buydens L M C. Using support vector machines for time series prediction. *Chemometrics and Intelligent Laboratory Systems*. 2003, 69: 35-49
- Vapnik V B, Golowich S E, Smola A J. Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*. 1996, 9: 281-287
- Vapnik V N. *The nature of statistical learning theory*. New York: Springer, 1995
- Xian G L, Luo X C, Xiao Y F. Statistics learning theory and support vector machine. *China science and technology information*. 2005, 12: 178-181 (in Chinese)
- Yan H, Zhang X G, Li Y D. Support vector machine methods in pattern recognition of sedimentary facies. *Computing techniques for geophysical and geochemical exploration*. 2000, 2: 158-164
- Zhang G P. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*. 2003, 50: 159-175