*Journal of the Korean Data & Information Science Society*
*2013,* **24**(2), 401–408

http://dx.doi.org/10.7465/jkdi.2013.24.2.401
한국데이터정보과학회지

# Soil moisture prediction using a support vector regression[†]

Danhyang Lee[1] · Gwangseob Kim[2] · Kyeong Eun Lee[3]

[1]Department of Statistics, Kyungpook National University
[2]Department of Civil Engineerings, Kyungpook National University
[3]Department of Statistics, Kyungpook National University

Received 23 February 2013, revised 20 March 2013, accepted 25 March 2013

**Abstract**

Soil moisture is a very important variable in various area of hydrological processes. We predict the soil moisture using a support vector regression. The model is trained and tested using the soil moisture data observed in five sites in the Yongdam dam basin. With respect to soil moisture data of of four sites-Jucheon, Bugui, Sangieon and Ahncheon which are used to train the model, the correlation coefficient between the esimtates and the observed values is about 0.976. As the result of the application to Cheoncheon2 for validating the model, the correlation coefficient between the estimates and the observed values of soil moisture is about 0.835. We compare those results with those of artificial neural network models.

*Keywords*: Kernel function, soil moisture, support vector regression.

## 1. Introduction

Soil moisture is a very important variable in various area of hydrological processes since it plays key role in land atmosphere interaction, rainfall runoff process and water energy balance in the basins. Therefore, last decades many studies were conducted to understand soil moisture characteristics. Recent advances of remote sensing technique of soil moisture allow us to get global data set and temporal measurements of soil moisture in ground networks also have established. The statistical characteristics of soil moisture fields and their variability were analyzed relating to the heterogeneity of precipitation, soil properties, vegetation, and topography (Kim and Barros, 2002; Rodriguez-Iturbe *et al.*, 1995). Soil moisture prediction models were proposed using physical approaches (Sheikha *et al.*, 2009) and statistical approaches (Laio, 2006).

[†] This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0030840).
[1] Ph.D. student, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea.
[2] Associate professor, Department of Civil Engineering, Kyungpook National University, Daegu 702-701, Korea.
[3] Corresponding author: Assistant professor, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea. E-mail: artlee@knu.ac.kr

Time series analysis is an active research issue in machine learning and data engineering (Van Gestel *et al.*, 2001). We perform the time series analysis for soil moisture prediction. Time series prediction can be regarded as the following simple numerical problem:

Time series data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ can be divided as data whose dimension is $k$, in a view of $(\boldsymbol{x}_i, \ldots, \boldsymbol{x}_{i+m-1})$. In other words, the prediction of time series can be briefly considered as the problem of finding a function $f : R^k \to R$ such that $f(\boldsymbol{x}_i, \ldots, \boldsymbol{x}_{i+m-1}) = \boldsymbol{x}_{i+m}$ , for every $i$ $(i = 1, 2, \ldots, M-m)$ . Other learning tasks such as classification or similarity computation of time series data are also formulated as these simple numerical problems. Support vector regressions have successfully solved these kinds of learning problems (Ruping, 2001).

Support vector regression, one of support vector machine technique applied to regression problems, is introduced by Vapnik *et al.* (1997). While support vector machines are generally used for classification problem, they can be applied to the regression problem. Because of their prediction accuracy and modeling conveniences, support vector regression has been applied to a variety of fields. Ahmad *et al.* (2010) developed a support vector machine model for estimating soil moisture using remote sensing data and Pasolli *et al.* estimated soil moisture using the support vector regression technique.

In this paper, we predict soil moisture in Yongdam dam basin in Korea using the support vector regression. We briefly review the support vector regression in Section 2 and apply the support vector regression to Yongdam dam data in Korea in Section 3. And we will conclude with some remarks in Section 4.

## 2. Support vector regression

The main goal of regression analysis is to estimate a functional relationship between input variables $\mathbf{X}$ and output variable $Y$. We want to find a function which fits the data well but is not much complex. So we can find the best function to optimize the risk function with a loss function and some appropriate regulularizers. For examples, the quadratic loss function with $l_2$ regulization is used in the ridge regression and the quadratic loss function with $l_1$ regulization is considered in the LASSO. Appropriate loss function and regulizers can be matched for a particular purpose.

Let a training data be $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_M, y_M)\}$, where $\boldsymbol{x}_i \in R^k$, $y_i \in R$, $M$ is the number of the training data points, and $R^k$ is the space of input data.

Generally, fitting the function in support vector regression is based on the linear form. Let the true regression model be the linear function as follows.

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + \beta \tag{2.1}$$

where $\boldsymbol{w} \in R^k$ and $\beta \in R$.

A $\epsilon$-insensitive loss function is proposed in the support vector regression and defined as follows:

$$I_\epsilon(y - f(\boldsymbol{x})) = \begin{cases} 0 & \text{if} \quad |y - f(\boldsymbol{x})| \le \epsilon \\ |y - f(\boldsymbol{x})| - \epsilon & \text{if} \quad |y - f(\boldsymbol{x})| > \epsilon \end{cases}.$$

This loss function ignores all error within $\pm\epsilon$ of the true regression function, $f(\boldsymbol{x})$ and it is proportional to its value out of the range of $\pm\epsilon$. This loss function enables a sparse set of

support vectors.

The optimal regression function can be obtained by minimizing the risk of the regression, $Risk(f)$, defined as follows:

$$Risk(f) = \phi[f] + C\sum_{i=1}^{M} I_\epsilon(f(\boldsymbol{x}_i) - y_i).$$

where $\phi[f]$ is the structure risk which controls smoothness or complexity of the function, $I_\epsilon(\cdot)$ is the loss function and $C$ is a pre-specified trade-off value. Generally, $\phi[f]$ takes the form of $\|w\|_1$ or $\frac{1}{2}\mathbf{w}^T\mathbf{w}$ in support vector regression (which is called $l_1$-SVR, $l_2$-SVR, respectively).

In addition, the above linear regression model can be generalized to non-linear regression model by using Mercer's kernel (Mercer, 1909). This could be achieved by mapping the data from the input space $R^k$ into a high-dimensional feature space $\chi$ by a function

$$\Pi : R^k \to \chi,$$

and solving the linear learning problem in $\chi$ . The interest is not to know the actual function $\Pi$, but to have a kernel function $\Psi$ which calculates the inner product in the feature space (Ruping, 2001).

$$\Psi(\boldsymbol{x}_1, \boldsymbol{x}_2) = \Pi(\boldsymbol{x}_1) \cdot \Pi(\boldsymbol{x}_2)$$

The kernel function most in use is the Gaussian radial basis function (RBF) which takes the form $\Psi(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-\gamma\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2)$. In this paper, support vector regression model with the RBF kernel is developed to predict soil moisture.

The optimization of support vector regression can be considered as follows (more specifically, the optimization of $l_1$-SVR),

$$min_{\mathbf{w},\beta,\xi_i,\xi_i^*} \|\mathbf{w}\|_1 + C\sum_{i=1}^{M}(\xi_i + \xi_i^*)$$

and the constraints are

$$\begin{aligned} y_i - f(\boldsymbol{x}_i) &\le \epsilon + \xi_i, \\ f(\boldsymbol{x}_i) - y_i &\le \epsilon + \xi_i^*, \\ \xi_i \ge 0, \xi_i^* &\ge 0, i = 1, \ldots, M, \end{aligned}$$

where $\xi_i$ and $\xi_i^*$ indicate the positive error and the negative error at the $i$th data point, respectively.

The above optimization problem can be solved by using the linear programming method. When the structure risk, $\phi[f]$ takes $\frac{1}{2}\boldsymbol{w}^t\boldsymbol{w}$ as $l_2$-SVR, the optimization problem becomes a quadratic programming problem (Yang *et al.*, 2009). This quadratic optimization problem can be equivalent to the problem of minimizing the Lagrangian

$$\begin{aligned} L &= \frac{1}{2}\boldsymbol{w}^t\boldsymbol{w} + C\sum_{i=1}^{M}(\xi_i + \xi_i^*) - \sum_{i=1}^{M}(\eta_i\xi_i + \eta_i^*\xi_i^*) \\ &- \sum_{i=1}^{M}\alpha_i(\epsilon + \xi_i - (y_i - f(\boldsymbol{x}_i)) - \sum_{i=1}^{M}\alpha_i^*(\epsilon + \xi_i^* - (f(\boldsymbol{x}_i) - y_i)), \end{aligned}$$

and the constraints are

$$\eta_i, \eta_i^*, \alpha_i, \alpha_i^* \geq 0,$$

where $L$ is the Lagrangian and $\eta_i, \eta_i^*, \alpha_i, \alpha_i^*$ are Lagrange multipliers.

For optimality, the following partial derivatives of $L$ with respect to the variables $(w, \beta, \xi_i, \xi_i^*)$ have to be zero.

$$\begin{aligned} \partial_\beta L &= \sum_{i=1}^{M}(\alpha_i^* - \alpha_i) \\ \partial_w L &= w - \sum_{i=1}^{M}(\alpha_i^* - \alpha_i)\boldsymbol{x}_i \\ \partial_{\xi_i} L &= C - \alpha_i - \eta_i \\ \partial_{\xi_i^*} L &= C - \alpha_i^* - \eta_i^*. \end{aligned}$$

Then it yields the dual optimization problem as follows (Smola and Schölkopf, 2004):

$$\begin{aligned} \text{maximize} &= \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{M}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\boldsymbol{x}_i^t\boldsymbol{x}_j \\ -\epsilon\sum_{i,j=1}^{M}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{M} y_i(\alpha_i - \alpha_i^*) \end{cases} \\ \text{subject to} &\quad \sum_{i=1}^{M}(\alpha_i - \alpha_i^*) = 0 \quad \text{and} \quad \alpha_i,\ \alpha_i^* \in [0, C]. \end{aligned}$$

We can also consider the non-linear case through the kernelization of support vector regression. The optimization problem can be written as follows:

$$min_{\mathbf{w},\beta,\xi_i,\xi_i^*} \|\mathbf{w}\|_1 + C\sum_{i=1}^{M}(\xi_i + \xi_i^*) \tag{2.2}$$

and the constraints are

$$\begin{aligned} y_i - (\boldsymbol{w}^T\pi(\boldsymbol{x}_i) + \beta) &\leq \epsilon + \xi_i, \\ (\boldsymbol{w}^T\pi(\boldsymbol{x}_i) + \beta) - y_i &\leq \epsilon + \xi_i^*, \\ \xi_i \geq 0, \xi_i^* &\geq 0, i = 1, \ldots, M. \end{aligned} \tag{2.3}$$

## 3. Application to soil moisture prediction at Yongdam dam basin

Figure 3.1 shows the target watershed located at the upstream Kum River basin. The area of the target watershed is $1164km^2$ and it consists of agriculture area (13.6%), forest (79.1%) and other type of land use (7.3%). The target watershed is selected since it has a reliable ground soil moisture monitoring network.

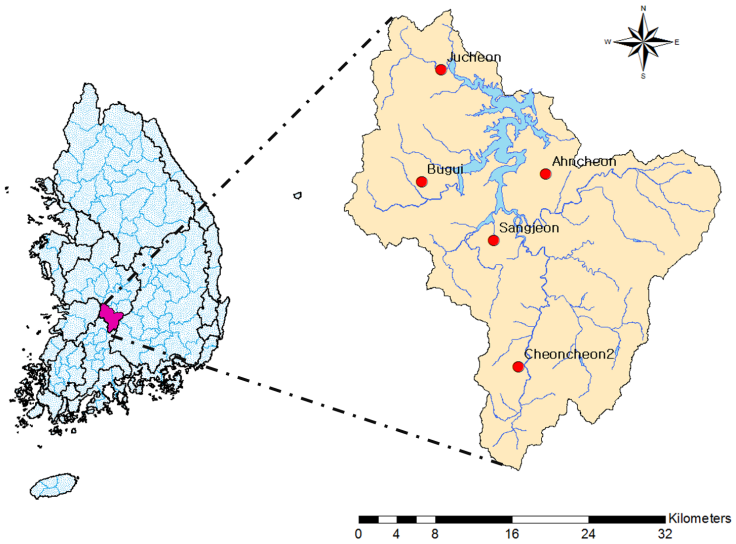



**Figure 3.1** Target watershed area for the Yongdam dam basin

For modeling the support vector regression to predict soil moisture, we use some ancillary information - rainfall (rf), surface temperature (st), the normalized distance vegetation index (NDVI) derived from moderate resolution imaging spectroradiometer (MODIS) and soil moisture (sm) itself as input variables. And these variables are measured at five sites ; Jucheon, Bugui, Sangieon and Ahncheon of Yongdam dam basin from 16 May 2008 to 19 Aug 2008. We can express more specifically our input variable vector as follows:

$$\boldsymbol{x}_m = (rf_m, st_m, NDVI_m, sm_{m-1}), \qquad m = 1, \ldots, M$$

where $sm_0 = \frac{\sum_{t=1}^{M-1} sm_t}{M-1}$.

Since soil moisture is greatly influenced by the previous observed soil moisture, we consider the soil moisture observed at the day before$(m-1)$ as the input variable. This can improve the performance of our model to predict soil moisture.

When there is no prior information about observed data, RBF is primarily used as a kernel function in support vector regressions (Karatzoglou *et al.*, 2006). So, we also use Gaussian RBF to estimate the model to predict soil moisture. And after dividing data into training data (Jucheon, Bugui, Sangieon, Ahncheon) and testing data (Cheoncheon2), the support vector regression model is fitted using Gaussian kernel with $C = 0.5$ and $C = 0.25$. Even though we fit the corresponding model with various trade-off values ($C \in (0.01, 20)$), the results are very similar.

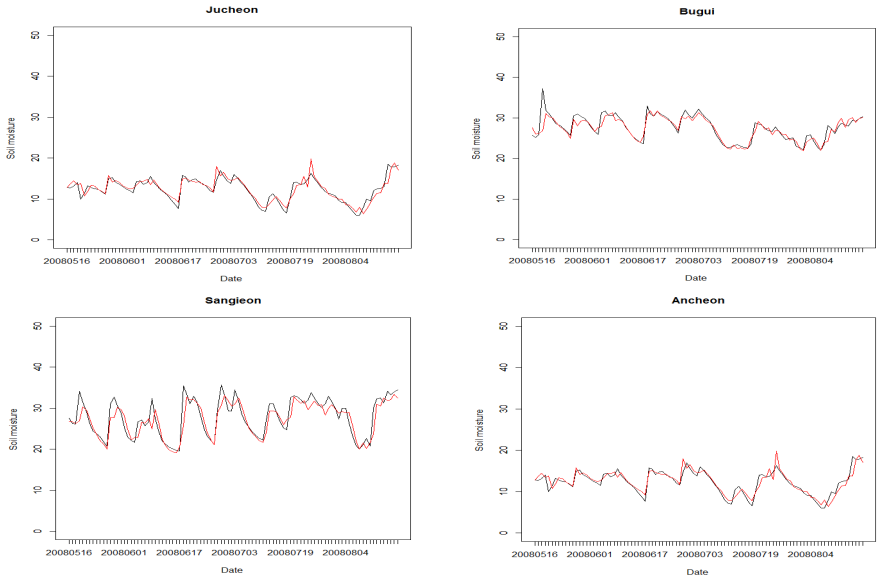



**Figure 3.2** Comparison between the observed soil moisture data (black line) and estimated soil moisture data (red line) of four sites - Jucheon, Bugui, Sangieon, Ahncheon

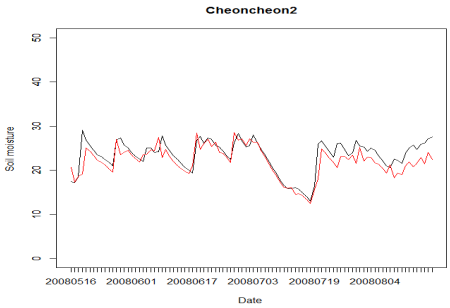


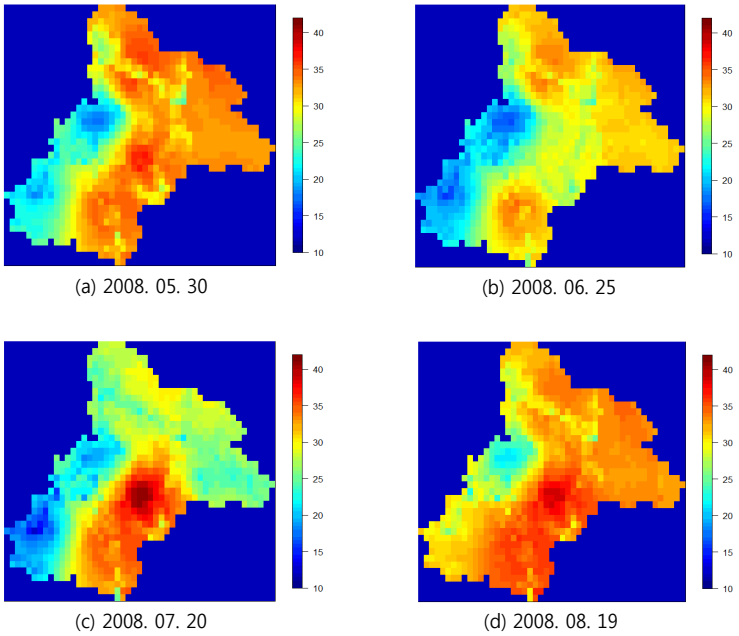

**Figure 3.3** Comparison between the observed soil moisture data (black line) and estimatedsoil moisture data (red line) of Cheoncheon2

To validate the feasibility of the model, four statistical criteria such as correlation coefficients (R), root mean squared error (RMSE) and mean absolute error (MAE) are considered. In addition, we compare these statistical indices values obtained from the support vector regression (SVR) model with those of the artificial neural network (ANN) model. Table 3.1 shows that in case of the SVR model, the Rs are from about 0.87 to 0.90, RMSEs are from about 1.46 to 2.31 and MAEs are from about 0.81 to 1.56 in estimating at four sites; Jucheon, Bugui, Sangieon and Ahncheon. And the results of test site-Cheoncheon2 are that the R is about 0.84, RMSE is about 2.41 and MAE is about 1.73. Meanwhile, the values of R obtained from the ANN model are from about 0.92 to 0.96, RMSEs are from about 1.00 to 1.88, and MAEs are from about 0.75 to 1.45 in estimating at four sites. In case of the test site, the R is about 0.91, RMSE is about 3.19 and MAE is about 2.72. These results show that even though the ANN model is fitted better to estimate the soil-moisture at training sites, the SVR model performs better for soil-moisture prediction than the ANN model.

**Table 3.1** Statistical indices of the soil moisture estimates for each site

| (**SVR**) | | R | RMSE | MAE | Mean | Min. | Max. |
|---|---|---|---|---|---|---|---|
| Training | Jucheon | 0.901 | 2.008 | 1.246 | 14.509 | 7.407 | 25.686 |
| | Bugui | 0.883 | 1.462 | 0.813 | 27.295 | 21.924 | 31.751 |
| | Sangieon | 0.871 | 2.305 | 1.556 | 27.180 | 19.121 | 33.347 |
| | Ahncheon | 0.905 | 1.208 | 0.877 | 12.453 | 6.378 | 19.804 |
| Validation | Cheoncheon2 | 0.835 | 2.408 | 1.732 | 21.970 | 12.442 | 28.522 |
| (**ANN**) | | R | RMSE | MAE | Mean | Min. | Max. |
| Training | Jucheon | 0.960 | 1.393 | 1.153 | 15.067 | 4.687 | 26.298 |
| | Bugui | 0.961 | 1.025 | 0.746 | 27.398 | 20.246 | 34.712 |
| | Sangieon | 0.915 | 1.878 | 1.449 | 4.067 | 20.246 | 34.712 |
| | Ahncheon | 0.942 | 1.004 | 0.769 | 12.009 | 6.322 | 17.439 |
| Validation | Cheoncheon2 | 0.911 | 3.192 | 2.720 | 20.731 | 11.897 | 28.472 |

Using our model based on data of five sites - Jucheon, Bugui, Sangieon, Ahncheon and Cheoncheon2, we apply it to predict soil moisture at the Yongdam dam basin. The following figures (Figure 3.2 and Figure 3.3) are some parts of predicting results.




(a) 2008. 05. 30 (b) 2008. 06. 25

(c) 2008. 07. 20 (d) 2008. 08. 19

**Figure 3.4** A sample of soil moisture prediction at the Yongdam dam basin

## 4. Conclusion

In this study, we proposed a model for soil moisture prediction using a support vector regression. The model was applied to the Yongdam dam, since the target basin has reliable soil moisture observations for model calibration and validation. For training the model, we considered rainfall, soil moisture at $(t-1)$, surface temperature at Jucheon, Bugui, Sangieon, Ahncheon sites and NDVI derived from MODIS. The results show that the correlation coefficient is about 0.976 and the value of $R^2$ is about 0.951. And data of Cheoncheon2 was used to validate the model. As the results, the correlation coefficient is about 0.835 and the value of $R^2$ is about 0.717, which reflects well the fluctuations of soil moisture. Results demonstrated that the model for soil moisture prediction using a support vector regression method and ancillary data such as soil temperature, vegetation index and rainfall should be useful to generate soil moisture fields. We comared those results with them of ANN model and ANN model performed better at the training sites but SVR model performed better at the test site.

## References

Ahmad, S., Kalra, A. and Stephen, H. (2010). Estimating soil moisture using remote sensing data: A machine learning approach. *Advances in Water Resources*, **33**, 69-80.

Karatzoglou, A., Meyer, D. and Hornik, K. (2006). Support vector machine in R. *Journal of Statistical Software*, **15**, 1-28.

Kim, G. and Barros, A. P. (2002). Space-time characterization of soil moisture from passive microwave remotely sensed imagery and ancillary data. *Remote Sensing of Environment*, **81**, 393-403.

Laio, F. (2006). A vertically extended stochastic model of soil moisture in the root zone. *Water Resources Research*, **42**, W02406, doi:10.1029/2005WR004502.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, **209**, 415-446.

Pasolli, L., Ntarnicola, C. and Bruzzone, L. (2011). Estimating soil moisture with the support vector regression technique. *IEEE Geosicence and Remote Sensing Letters*, **8**, 1080-1084.

Rodriguez-Iturbe, I., Vogel, G. K., Rigon, R., Entekhabi, D., Castelli, F. and Rinaldo, A. (1995). On the spatial organization of soil moisture fields. *Geophysical Research Letters*, **22**, 2757-2760.

Ruping, S. (2001). SVM kernels for time series analysis. *LLWA 01 - Tagungsband der GI-Workshop-Woche Lernen - Lehren - Wissen - Adaptivitt*, 43-50.

Sheikha, V., Saskia Visserb, S. and Stroosnijderb, L. (2009). A simple model to predict soil moisture: Bridging event and continuous hydrological (BEACH) modelling. *Environmental Modeling & Software*, **24**, 542-556.

Smoal, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, **14**, 199-222.

Van Gestel, T., Suykens, J., Baestaens, D., Lambrechts, A., Lanckriet, G., Vandaele, B., De Moor, B. and Vandewalle, J. (2001). Financial time series prediction using least squares support vector machines within the evidence framework. *IEEE Transactions on Neural Networks*, **12**, 809-821.

Vapnik, V., Golowich, S. and Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processing Systems*, **9**, edited by M. Mozer and M. Jordan and T. Petsche, MIT Press, Cambridge, MA, 281-287.

Yang, H., Huang, K., King, I. Lyn, M. R. (2009). Localized support vector regression for time series prediction. *Neurocomputing*, **72**, 2659-2669.