Comparison of Support Vector Machine and Generalized Additive Model for Evapotranspiration Prediction

Rigele Te and Sunil Kumar Bhandari

Kansas State University

Abstract

The accurate prediction of reference evapotranspiration (ET) is of great importance for the planning of regional water resources and irrigation system design. In this study, machine learning approach Support Vector Machine (SVM) and statistical method Generalized Additive Model (GAM) as a function of spatial and temporal data were compared using different statistical metrics for the accurate prediction of reference evapotranspiration across the Continental United States. The study shows that Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for SVM was 0.248 mm/day and 0.196 mm/day respectively and for GAM was 0.606 mm/day and 0.583 mm/day respectively. The computational time for SVM and GAM was around 540 and 3 seconds respectively. Overall the performance of SVM model was good over than GAM model but with the high computational time. Although the SVM performance was good with the smaller amount of data, as the data increases the computational time for SVM would be very high compared with the GAM. The study used smaller data set and hence the result obtained may not be apply to the current situation but the idea is still applicable when used with larger data set.

Keywords: Evapotranspiration; GAM; SVM

1 Introduction

(Allen, Pereira, Raes, Smith, et al., 1998) defined Evapotranspiration (ET) as “the combination of water loss from the soil surface by evaporation and from crop by transpiration” and is essential component for studies such as hydrological water balance, efficient irrigation planning, and water resources planning and management. Evapotranspiration can be either measured with a direct method such as lysimeter (Shrestha & Shukla, 2015) or indirect methods estimated from meteorological data. Nevertheless, the computation with a lysimeter is expensive, time consuming, and required precisely planned experiments (Chia, Huang, & Koo, 2020). Therefore, indirect methods, which vary from empirical relationships to the complex function, based on different meteorological data are used for ET estimation.

Evapotranspiration is a non-linear and complex phenomenon which depends on integrating climatological factors and hence is difficult to estimate using empirical approach. Therefore, researchers have put forward the application of Artificial Intelligence (AI) by applying different

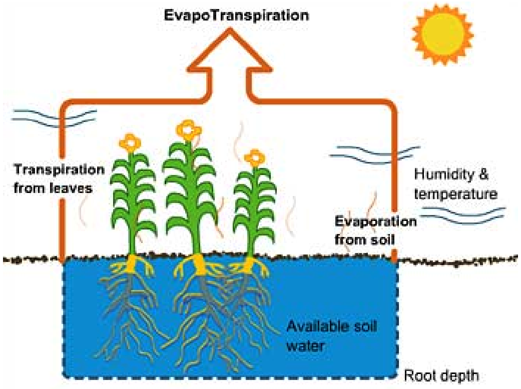

Figure 1: Evapotranspiration process.

(Source:http://ecoursesonline.iasri.res.in/mod/page/view.php?id=2000) machine learning algorithms and their application in ET estimation has increased since 2011 (Chia et al., 2020). Artificial Intelligence is a black box which maps the non-linear relationship between input and output variable and hence can significantly improve the prediction accuracy of ET.

The Support Vector Machine algorithm developed by (Yang et al., 2006) performance was better than Multiple Layer Perceptron with Back Propagation for the non -forest sites across the Continental US. The SVM model with radial basis function as the kernel function algorithm performance at the Chahnimeh reservoirs, Sistan region was found good with the gamma test (Moghaddamnia, Ghafari, Piri, & Han, 2009). The SVM model with only solar radiation and mean temperature as an input variable works best than input with other variables in the tropical climate in Malaysia (Chia et al., 2020).

Most of the Machine Learning model are inherently unstable and may yield less accuracy when new dataset is used for the prediction. The models so far works well for particular season in particular ecosytem only. The performance of model largely depends on time and space and hence it is great idea to built the model with spatial- temporal data inconsideration for the prediction of Evapotranspiration.

The Generalized Additive Models is a mode of inductive reasoning that has been applied in a lot of scientific disciplines which permits the researcher to use sample data and prior information in a logically consistent manner in making inferences. It's specified by giving a symbolic description of the additive predictor and a description of the error distribution.

However, the applications of GMA model to estimate evapotranspiration are currently limited and the knowledge on the topic is still partial and fragmented. The comparison of Machine Learning approach such as SVM and statistical GAM model has not been comprehensively conducted yet for the prediction of ET with meteorological data as an input under different climatic condition.

In this study, two different data-driven models were developed for forecasting the evapotranspiration. The aims of the present study were to: Comparison of the prediction accuracy of Support Vector Machine and GAM model for the prediction of Evapotranspiration across the diverse climate of Contitental United States.

The evapotranspiration is calculated using the ETRHEQ method available at (Rigden & Salvucci, 2015) which acts as an output i.e. response variable to the model. The input for the model include the four climatic parameter with longitude, latitude, elevation and date.

2 Study area

Our study area covers the central USA, which is generally designated as Continental United States. The area of the region is extended from the Canadian border in the north to Texas in the south, Wyoming and Colorado on the west and Iowa on the east. The marginal area consists of nine states (North Dakota, South Dakota, Nebraska, Iowa, Wyoming, Colorado, Kansas, Oklahoma and Texas), which comprise of total 834 counties. About 30% of the terrestial area of the USA is acquired by the marginal study area in which the topography varies from flat-to-rolling plains. The highest elevation is in the Rocky Mountains in Colorado and the lowest elevation is at the southern coastline in Texas. The blue dots in the figure 2 represent the location of 80 stations across the Continental United States.The data is acquired from 80 stations distributed uniformly across the study area (Fig. 2).

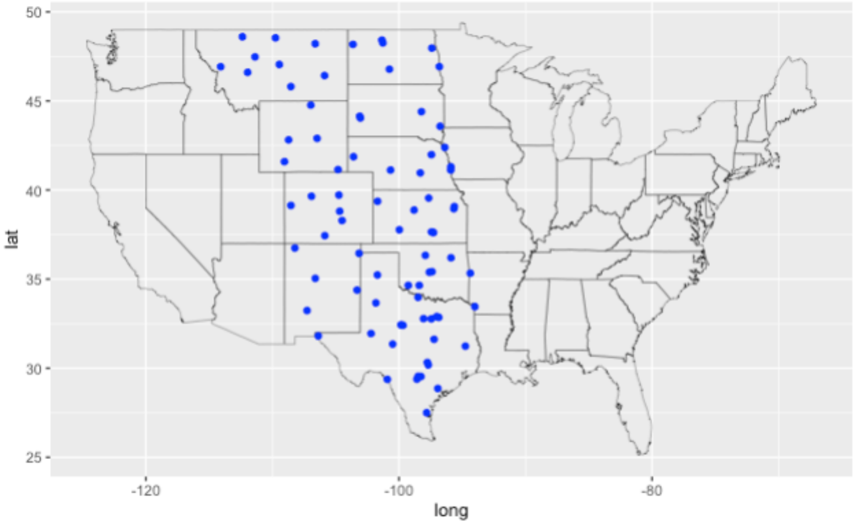

Figure 2: Study Area

3 Methodology

3.1 Support vector machines

Support Vector Machines are supervised learning algorithms developed by (Vapnik, 2013) which deals with the classification and regression problems. The Support Vector Machine (SVM) works by transforming the lower-dimensional input data into higher-dimensional feature space and map their relationship with the output data. The transformation of input vector is controlled by the various kernel functions such as linear, polynomial, radial basis function (RBF), and sigmoid. The type of kernel function use depends on the field of application. In general, the SVM can be expressed mathematically as shown in equation:

$y = w * \phi(x) + b$

in which y represent output vector, w is the weight vector, ϕ is the kernel function, x indicate the input vector and b is the bias term. The bias term and weight are estimated by minimizing the loss function as shown in equation below:

$0.5||w||^2 + C\frac{1}{n}\sum_{i=1}^{n} L_\varepsilon(y_{predicted}, y_{actual})$

where

$$L_\varepsilon(y_{predicted}, y_{actual}) = \begin{cases} 0 & \text{if } |y_{predicted} - y_{actual}| < \varepsilon \\ |y_{predicted} - y_{actual}| - \varepsilon & \text{otherwise} \end{cases} \tag{1}$$

in which L_ε is the ε-insensitive error function, ε is the margin of the SVM, C is a penalty parameter of the error term and $||w||$ is the regularization term, $y_{predicted}$ and y_{actual} are the predicted and actual value of output ET respectively.

The Radial Basis Function is one of the most commonly deployed kernel functions ((Chia et al., 2020); (Fan et al., 2018); (Dou & Yang, 2018); (Granata, 2019);(Shrestha & Shukla, 2015)) in predicting the evapotranspiration. RBF has advantages of good generalization and strong tolerance to input noise. Thus, the Radial Basis Function (RBF) was used for this study as shown below:

$$K(x_n, x_i) = exp(-\gamma||x_n - x_i||^2 + C)$$

where x_n & x_i are the n^{th} and i^{th} term of input vector and γ & C are the parameters to be tuned in the SVM model.

The "kernlab" package included in the "R" software (Team et al., 2013) was used for the SVM to model the evapotranspiration in (Shrestha & Shukla, 2015). The same package is also used for our study to predict the evapotranspiration.

3.2 Generalized Additive Models

Under this method, all the calculations were conducted with the programming language R (R Core Team, 2013). We first considered Markov Chain Monte Carlo (MCMC) method but it takes huge amount of effort in finding the accurate conditional distributions of parameters. Even establishing Metropolis algorithms can be inefficient sometime. However, generalized additive model (GAM) provides a much more flexible and simpler way in explaining the relationships between response and predictors. It is a generalized linear model (GLM) in which the linear predictor depends linearly on unknown smooth functions of some predictor variables (Bender, Groll, & Scheipl, 2018). GAM process can be carried out in R by using the MGCV [1] package.

The original thinking was to set Evapotranspiration be the response variable and Elevation, Temperature, Wind speed, Relative Humidity, Solar Radiation along with location and time will be the predictors. The formulas of GAM function are shown below:

$$\mu_i = (\beta_0 + \beta_i x_i + z_j)^{1/\lambda}, i = 1, 2, 3, 4, 5; j = 1, 2 \tag{2}$$

$$[y_i|\beta, z, X] = Tweedie(\mu_i, \phi, p) \tag{3}$$

$$z_1 \sim N(0, \kappa) \tag{4}$$

Link function is the power link function $\eta = \mu^\lambda$. In this case we choose $\lambda = 0.1$. z_1 is the spatial random effect follows a normal distribution with mean 0 and covariance matrix k. Low rank Gaussian process is used to provide inference on the variation. z_2 is the time random effect with thin plate regression splines. These are low rank isotropic smoothers of any number of covariates. By isotropic is meant that rotation of the covariate co-ordinate system will not change the result of smoothing. By low rank is meant that they have far fewer coefficients than there are data to smooth.

Tweedie is an exponential family distribution for which the variance of the response is given by the mean response to the power $p = 1.25$. ϕ is a dispersion parameter. Tweedie distribution has a support great or equal to 0 which suits the characters of the response variables. $x_1, ..x_5$ represents Elevation, Temperature, Wind speed, Relative Humidity and Solar Radiation respectively, with coefficients $\beta_1, ..\beta_5$.

4 Results and Discussion

We cleaned data set by removing unreasonable none positive ET values. Time series analysis has been the first step to perform data exploration so we randomly choosed four stations and for each location, the evapotranspiration in first 10 years has been plotted (Fig. 3).

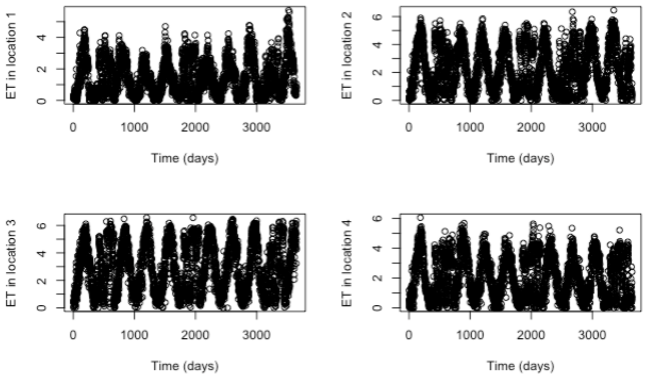

Figure 3: ET trend

The figure shows that a yearly trend among all the station: ET started increasing from the beginning of the year, reached it's peak in summer, finally decreasing gradually during fall and winter, cycling over and over again annually. This trend reminded us to exam the time series correlation of ET for each stations. One common way for examining the coefficient of correlation between two values in a time series is called the autocorrelation function (ACF). Again, we randomly picked four stations and calculated ACF of the evapotranspiration. We note that the majority of the stations chooses a AR(1) for a temporal margin (Fig. 4) based on these commonly used model selection criteria, indicating that time is a very important factor we need to consider.

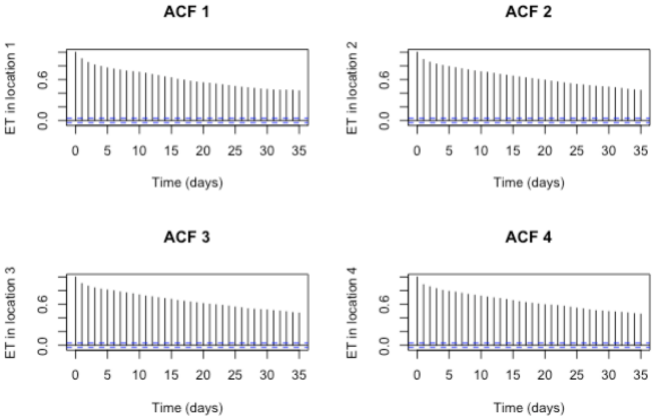

Figure 4: ACF

Multicollinearity is another thing to consider when we perform variable selection based on regression model. It occurs when independent variables in a regression model are correlated. This correlation is a problem because variables should be independent. If the degree of correlation between variables is high enough, it can cause issues when fitting the model and interpret the results. The correlation between remaining variables: Elevation, solar radiation, wind speed, relative humidity, and mean temperature has been checked. The correlation plot (Fig. 5) shows the predictors are not highly correlated, indicating that there is no need to consider multicollinearity problem before we move on.

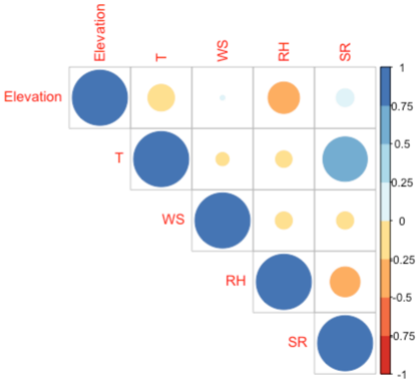

Figure 5: Correlation

Since the entire dataset is too large to handle with, based on the yearly trend of ET values given by data exploration result, it's enough to utilize all the information during a short period of time to fit the model. Thus, we set the training data as the weather data from September to November, 2014 including all 80 stations and testing data to be the weather data in December, 2014 among 80 stations.

As for the GAM function method, the estimated parameter values are $\beta_0 = 7.250e-01$, $\beta_1 = -2.915e-05$,$\beta_2 = 2.519e-03$,$\beta_3 = 1.819e-03$,$\beta_4 = 1.661e-03$,$\beta_5 = 7.877e-03$. The output showed that the p-value for the parametric coefficients as well as smooth terms are significant, which means all the component are important in the model. The tuned function in kernlab package in "R" was used to find the best parameters combination to fit the SVM model, the estimated tuning parameters for support vector machine are as follows: $\gamma = 0.1$, $\epsilon = 0.1$, and $C = 32$.

Table 1. ET RMSE Statistics

Measure	GAM	SVM
AVG. RMSE	0.606	0.248
STD. DEV.	0.098	0.123
95% DATA.I	(0.585, 0.628)	(0.220, 0.275)
Low Station	2	78

Table 2. ET MAE Statistics

Measure	GAM	SVM
AVG. MAE	0.583	0.196
STD. DEV.	0.114	0.105
95% DATA.I	(0.558, 0.608)	(0.173, 0.219)
Low Station	1	79

As in (Gneiting, Genton, & Guttorp, 2006), the root mean-square error (RMSE) and mean absolute error (MAE) are examined to compare the prediction performance of the models.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \tag{6}$$

Where y_i, $\hat{y_i}$ are the true value and estimated value respectively.(Demel & Du, 2015) generated a reasonable way of comparing multiple spatio-temporal model in different locations. Since RMSE is calculated for each of 80 stations and it is meaningful to consider the consistency of the superior performance. Table 1 gives the average RMSE, standard deviation over all stations. The 95% data interval gives empirical interval that covers the majority of RMSE across stations. The "low count" gives how many stations have the lowest RMSE per method.

Although all the models have a average RMSE pretty small on ET, the SVM model does have the lowest average RMSE that is much smaller than GAM. 97.5 % of stations, are best fitted by the SVM in prediction. However, as for the variability, the GAM model results in a slightly smaller standard deviation.

Using the same techniques, we also use MAE to assess predictive accuracy. Table 2 shows the mean MAE of all stations and again the SVM method has the lowest value for ET. Also the SVM method has the most stations with the lowest MAE with 79 out of 80 stations for ET given by Low Count. Based on this analysis, the SVM preforms much better than GAM.

However, although SVM gives better results in prediction, it still has several drawbacks to be considered. First, it is not very efficient computationally for large datasets. The running time for SVM is 544.901 seconds comparing to 3.485 seconds for GAM model. When we increased training dataset, the SVM spent much longer time than GAM approach. Second, it 's difficult to interpret variable weights and individual impact of SVM based on lacking of model assumptions.

The major limitation of this study is that the training data set isn't big enough, so that the exact result concluded from this dataset may not be applied to the current situation, but the idea is still applicable. We need to train larger dataset to catch up yearly trend. Also, as we explored ACF, we might consider putting AR(1) type of correlation part into the model structure. Another improvable part is that we can try to combine the benefit of both SVM and GAM method to get result faster and more accurate.

References

Allen, R. G., Pereira, L. S., Raes, D., Smith, M., et al. (1998). Crop evapotranspiration-guidelines for computing crop water requirements-fao irrigation and drainage paper 56. *Fao, Rome*, *300*(9), D05109.

Bender, A., Groll, A., & Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, *18*(3-4), 299–321.

Chia, M. Y., Huang, Y. F., & Koo, C. H. (2020). Support vector machine enhanced empirical reference evapotranspiration estimation with limited meteorological parameters. *Computers and Electronics in Agriculture*, *175*, 105577.

Demel, S. S., & Du, J. (2015). Spatio-temporal models for some data sets in continuous space and discrete time. *Statistica Sinica*, 81–98.

Dou, X., & Yang, Y. (2018). Evapotranspiration estimation using four different machine learning approaches in different terrestrial ecosystems. *Computers and Electronics in Agriculture*, *148*, 95–106.

Fan, J., Yue, W., Wu, L., Zhang, F., Cai, H., Wang, X., … Xiang, Y. (2018). Evaluation of svm, elm and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of china. *Agricultural and Forest Meteorology*, *263*, 225–241.

Gneiting, T., Genton, M. G., & Guttorp, P. (2006). Geostatistical space-time models, stationarity, separability, and full symmetry. *Monographs On Statistics and Applied Probability*, *107*, 151.

Granata, F. (2019). Evapotranspiration evaluation models based on machine learning algorithms—a comparative study. *Agricultural Water Management*, *217*, 303–315.

Moghaddamnia, A., Ghafari, M., Piri, J., & Han, D. (2009). Evaporation estimation using support vector machines technique. *International Journal of Engineering and Applied Sciences*, *5*(7), 415–423.

Rigden, A. J., & Salvucci, G. D. (2015). Evapotranspiration based on equilibrated relative humidity (etrheq): Evaluation over the continental us. *Water Resources Research*, *51*(4), 2951–2973.

Shrestha, N., & Shukla, S. (2015). Support vector machine based modeling of evapotranspiration using hydro-climatic variables in a sub-tropical environment. *Agricultural and Forest Meteorology*, *200*, 172–184.

Team, R. C., et al. (2013). *R: A language and environment for statistical computing.* Vienna, Austria.

Vapnik, V. (2013). *The nature of statistical learning theory.* Springer science & business media.

Yang, F., White, M. A., Michaelis, A. R., Ichii, K., Hashimoto, H., Votava, P., … Nemani, R. R. (2006). Prediction of continental-scale evapotranspiration by combining modis and ameriflux data through support vector machine. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(11), 3452–3461.