Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

JZUS

Application of least squares vector machines in modelling water vapor and carbon dioxide fluxes over a cropland*

QIN Zhong (秦 钟)†1, YU Qiang (于 强)2, LI Jun (李 俊)2, WU Zhi-yi (吴志毅)3, HU Bing-min (胡秉民)†4

(1*Institute of Ecology, School of Life Science, Zhejiang University, Hangzhou 310029, China*)
(2*Institute of Geographic Sciences and Natural Resources Research*, *Chinese Academy of Sciences*, *Beijing 100101, China*)
(3*Institute of Applied Entomology, School of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310029, China*)
(4*School of Science, Zhejiang University, Hangzhou 310029, China*)
†E-mail: q_breeze@126.com; bmhu@mail.hz.zj.com

Received Oct. 7, 2004; revision accepted Feb. 28, 2005

Abstract: Least squares support vector machines (LS-SVMs), a nonlinear kemel based machine was introduced to investigate the prospects of application of this approach in modelling water vapor and carbon dioxide fluxes above a summer maize field using the dataset obtained in the North China Plain with eddy covariance technique. The performances of the LS-SVMs were compared to the corresponding models obtained with radial basis function (RBF) neural networks. The results indicated the trained LS-SVMs with a radial basis function kernel had satisfactory performance in modelling surface fluxes; its excellent approximation and generalization property shed new light on the study on complex processes in ecosystem.

Key words: Least squares support vector machines (LS-SVMs), Water vapor and carbon dioxide fluxes exchange, Radial basis function (RBF) neural networks
doi:10.1631/jzus.2005.B0491 **Document code:** A **CLC number:** S1; TP1.18

INTRODUCTION

Modelling the variation in surface-atmospheric exchange of water vapor and carbon dioxide fluxes and how they are influenced by a complex combination of environment variables and plant physiology is crucial for assessing the annual water and carbon budget for cropland. Biophysical or process-based models such as soil-vegetation-atmosphere transfer (SVAT) (Franks *et al*., 1997; Franks and Beven, 1999; Mo and Beven, 2004), Simultaneous Heat and Water (SHAW) (Flerchinger *et al*., 1996; Flerchinger and Pierson, 1991) has been developed to quantify these fluxes in different time scales and their interrelationship with biotic and abiotic factors based on field experiments. In their simulation, detailed processes of photosynthesis, respiration from vegetation and soil carbon components, evapotranspiration, hydrological cycling, etc. are required to be explicitly modelled and more realistic parameters such as soil, plant and vegetation characteristics, surface resistance are required to be specified (Unland *et al*., 1996; Arora, 2003), which makes fluxes estimation complex and carbon or water budget construction relatively inconvenient.

Artificial neural network (ANN), a tool used to process information in a non-liner manner, enables completely unconstrained optimization and estimation of input-output responses without a predefined mathematical model (Kosko, 1992; Demuth and Beale, 1994; Schulz and Härtling, 2003), has recently been applied to model fluxes exchanges at the

* Project supported by the National Science Fund for Outstanding Youth Overseas (No. 40328001) and the Key Research Plan of the Knowledge Innovation Project of the Institute of Geographic Sciences and Natural Resources, Chinese Academy of Sciences (No. KZCXI-SW-01)

land-atmosphere interface recently and been proved to have a higher accuracy compared to classical methods (Huntingford and Cox, 1997; Van Wijk and Bouten, 1999), though physical processes and parameters are introduced into the model structure. Despite their advantages, a number of drawbacks still remain like non-convex training problem with multiple local minima, dependence on quantity and quality of training dataset, choice of the number of hidden units, etc. (Suykens, 2001).

A breakthrough was obtained at this point when a new powerful machine learning method−support vector machines (SVMs), was developed on the basis of statistical learning theory in the last decade. Many successful applications in nonlinear classification and function estimation have shown that SVMs can handle higher dimensional data better even with relatively fewer training samples and that they exhibit very good generalization ability for complex models (Vapnik, 1995; 1998). The standard SVM is solved using complicated quadratic programming methods, which are often time consuming and difficult to implement adaptively, whilst LS-SVM is solved by a set of linear equations amenable to solution via on-line adaptive methods (Suykens, 2001), so we hereby use LS-SVMs to model surface water vapor and carbon dioxide fluxes. The goal of this paper is to investigate applicability of LS-SVMs in modelling the dynamics of water vapor and carbon dioxide fluxes over a cropland.

THEORY

Given a training set $\{x_k, y_k\}_{k=1}^{N}$ with input data $x_k \in \mathrm{R}^n$ and output data $y_k \in \mathrm{R}$, the LS-SVM model for function estimation has the following representation in feature space,

$$y(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x})+b \tag{1}$$

Here the nonlinear function $\varphi(\cdot)$: $\mathrm{R}^n \rightarrow \mathrm{R}^{n_k}$ maps the input space to a higher dimension feature space. The dimension n_k of this space is only defined in an implicit way; b is a bias term; $\boldsymbol{w} \in \mathrm{R}^{n_k}$ is weight vector; $\boldsymbol{e}_{\mathrm{k}} \in \mathrm{R}$ is error vector; γ is the regularization parameter. The optimization problem is defined as:

$$\min_{w,b,e} J(\boldsymbol{w},\boldsymbol{e}) = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N}\boldsymbol{e}_i^2 \tag{2}$$

Subject to the equality constraints:

$$y_i = \boldsymbol{w}^{\mathrm{T}}\varphi(x_i) + b + \boldsymbol{e}_i \qquad i=1,\cdots,N \tag{3}$$

The solution is obtained after constructing the Lagrangian,

$$L(\boldsymbol{w},b,\boldsymbol{e};\boldsymbol{\alpha}) = J(\boldsymbol{w},\boldsymbol{e}) - \sum_{i=1}^{N}\alpha_i\left\{\boldsymbol{w}^{\mathrm{T}}\varphi(x_i) + b + \boldsymbol{e}_i - y_i\right\} \tag{4}$$

where α_i are Lagrangian multipliers. Application of the conditions for optimality yields the following linear system (5):

$$\begin{bmatrix} 0 & \mathbf{1}^{\mathrm{T}} \\ \mathbf{1} & \boldsymbol{\Omega} + \gamma^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{y} \end{bmatrix} \tag{5}$$

where $\boldsymbol{y}=[y_1,\ldots,y_N]$, $\mathbf{1}=[1,\ldots,1]$, $\boldsymbol{\alpha}=[\alpha_1,\ldots,\alpha_N]$, Mercer's condition is applied in the $\boldsymbol{\Omega}$ matrix

$$\boldsymbol{\Omega}_{il} = \psi(x_i,x_l) = \varphi(x_i)^{\mathrm{T}}\varphi(x_l),\ i,l=1,\cdots,N \tag{6}$$

The resulting LS-SVM model for function estimation becomes,

$$y(\boldsymbol{x}) = \sum_{i=1}^{N}\alpha_i\psi(\boldsymbol{x},x_i) + b \tag{7}$$

where α_i, b comprise the solution to the linear system.

In Eq.(6), $\psi(x_i,x_l)$ is the so-called kernel function with which the input vector can be mapped implicitly into a high-dimension feature space. The most usual kernel functions are polynomial, Gaussian-like or some particular sigmoids (Suykens, 2001).

SIMULATION AND RESULTS

Collection of the dataset

The dataset used in this work was obtained after continuous measurements of carbon dioxide and water vapor fluxes above a winter wheat and summer maize rotation field with eddy covariance technique from Nov. 2002 to Oct. 2003 at Yucheng Comprehensive Experiment Station (36°57′ N, 116°36′ E, 20 m a.s.l) in the North China Plain. Only data collected during the summer maize growth stage from the day of sowing [day of year (DOY165)] to harvest (DOY275) was applied for this study.

Of 105 days for fluxes and meteorological measurements on half hourly basis, the missing 5.67% data in observed fluxes was due to instrument maintenance, calibration, malfunction of the sensors and supporting equipment. The 0.06%, 0.48% and 0.73% of unreasonable data for carbon dioxide flux (F_c), water vapor flux (LE) and sensible heat flux (H_s) respectively were rejected; about 3.06% and 3.41% of soil heat flux data recorded at depth of 0.03 m and 0.05 m were eliminated from the dataset for the same reason. Measurements within 24 h after a rain event were also removed from the database.

Nighttime fluxes have been reported to be underestimated by the eddy covariance approach during stable condition because of CO_2 storage in the layer below the eddy flux system. A u^* threshold (u^*>0.12 ms^{-1}) was determined and data with values below this threshold were removed from the dataset (Falge *et al.*, 2001; Anthoni *et al.*, 2004).

Consequently, there were 1951 half hourly complete data records available for modelling.

Determination of input variables

Land-atmosphere exchange of water vapor and carbon dioxide fluxes associated with complex ecophysiological processes including assimilation and transpiration, which are driven by external factors such as solar radiation, atmospheric conditions, soil water status and internal factors such as plant physiological and biometrical conditions (Bosveld and Bouten, 2001). Many previous studies on assessing the environmental constraints on carbon and water vapour exchange in forest, grassland or crop field (Huntingford and Cox, 1997; Kelliher *et al.*, 1997; Valentini *et al.*, 1996; Granier *et al.*, 2000a; 2000b; Van Wijk and Bouten, 1999; Baldocchi and Wilson, 2001; Bosveld and Bouten, 2001) proved this very well. Based on the results of study with artificial neural networks using the same dataset, photosynthetically active radiation (PAR), vapor pressure deficit (VPD), air temperature (*T*), and leaf area index (LAI) were taken as the inputs in modelling both carbon dioxide and water vapor fluxes exchange. Besides these variables, wind velocity (U) and soil volumetric water content (W) were selected for computing F_c and LE respectively.

Preparation of training dataset

One thousand and nine hundred fifty-one data records were randomly divided into two subsets, one used exclusively for training and the other exclusively for testing, were applied to the development of SVMs. All the variables were rescaled to be included within the interval [−1, 1] by using the following equation:

$$x_{\text{scale}} = \frac{x - x_{\min}}{x - x_{\max}} \tag{8}$$

where x and x_{scale} were the old and new value of the variable for a sampling point respectively, $x_{\min}$ and $x_{\max}$ were the minimum and maximum values of that variable in the original dataset.

SVM training

In this study, we take radial basis function (RBF) kernals with LS-SVMs,

$$\psi(x, x_i) = \exp(-\frac{\|x - x_i\|^2}{\sigma^2}) \tag{9}$$

where σ>0 is a constant defining the kernel width.

It should be noted that predetermined parameters in LS-SVMs algorithms with RBF kernel are γ and σ^2, which are less than those in standard SVMs (Jemwa and Aldrich, 2003). Moreover, ε-insensitive formulation in Vapnik's standard SVMs are modified by introducing a squared error term and equality constraints so that one solves a linear system instead of a quadratic programming problem, thus greatly reduce the computing complexity. More detailed information on SVMs can be found (Vapnik, 1995; 1998; 1999).

A 5-fold cross-validation procedure (Duan *et al.*, 2001; Witten and Frank, 2000) was used to train and test the LS-SVMs under various model and parametric settings. Neural network models were implemented with software package LS-SVMlab1.5. (Pelckmans *et al.*, 2002). Parameters in LS-SVMs for

F_c and LE are given in Table 1. Since the results of modelling water vapor and carbon dioxide fluxes were similar, we only showed the former in this paper (Fig.1, Table 2). As shown in Table 2, LS-SVM with RBF kernel for modelling water flux had greater learning mean squared errors (LMSE) and less generalization means squared errors (GMSE) compared with those in RBF neural network no matter what the proportion of dataset for analysis was. The GMSEs of the two algorithms increased with the training dataset size at different rate with the RBF neural network having the larger. LS-SVM was less dependent on the size of the training dataset and having better generalization ability than that of the RBF neural network.

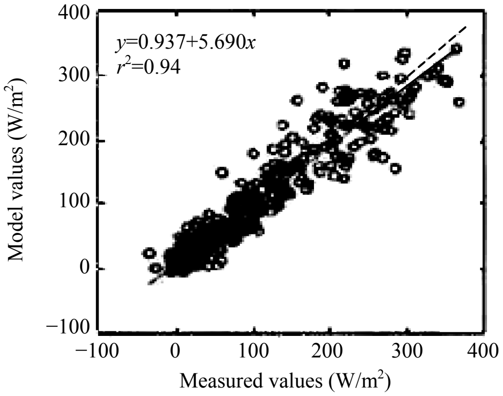

Fig.1 Analysis of dataset-estimated values with LS-SVM for water vapor flux (unit: W/m^2) (σ^2=4.513, γ=279.568)

Table 1 Parameters in LS-SVMs with RBF kernel

Outputs	γ	σ^2
F_c	115.342	3.254
LE	279.568	4.513

Table 2 Mean squared errors (MSE) analysis of LS-SVM and RBF neural networks for water vapor flux simulation

Training set	Testing set	LS-SVM			RBF neural network	
		LMSE	*GMSE*	SVs*	*LMSE*	*GMSE*
390	1561	0.1749	0.2753	272	0.1256	0.2875
585	1366	0.1786	0.2684	472	0.1367	0.2742
780	1171	0.1843	0.2576	638	0.1534	0.2615
976	975	0.1735	0.2357	829	0.1610	0.2386
1171	780	0.1701	0.2415	948	0.1583	0.2274
1366	585	0.1773	0.2364	1026	0.1642	0.2382

SVs are support vectors

DISCUSSION AND CONCLUSION

Support vector machines (SVMs) have been proposed as a powerful machine learning approach. An improved SVMs model, least squares support vector machines (LS-SVMs), which enables solution of highly nonlinear and noisy black-box modelling problems was presented and applied to modelling the dataset obtained from the field measurements in this study. The results indicated that LS-SVMs could be used to model surface fluxes exchange without restrictive assumptions and parameters specification required by other models. Compared with RBF neural network, it has stronger learning ability, better generalization ability and is less dependent on the size of the training dataset.

Besides, SVM's ability to handle high-dimension and incomplete data allows successful extraction of information even when part of the data records was missing or unreasonable owing to the problems of instrument malfunction or maintenance, calibration and climate influences, so SVMs method is suitable to simulate land-atmosphere interaction in an efficient and stable way.

Although the proposed LS-SVM-based model may be superior to other modelling methods in some aspects, it has some potential drawbacks such as the underlying Gaussian assumptions related to a least squares cost function. Some researchers have made some efforts to overcome these by applying an adapted form called weighted LS-SVM (Thissen *et al.*, 2003).

We intend to continue the studies on the application of LS-SVMs in modelling energy and mass exchanges at the cropland-atmosphere interface using larger dataset obtained from different sites. Possibilities to model these fluxes independently of crop species or sites specification will be investigated. Furthermore, LS-SVMs application in prediction or gap-filling the missing data as well as its algorithm optimization will be the focus in our future work.

References

Anthoni, P.M., Freibauer, A., Kolle, O., Schulze, E.D., 2004. Winter wheat carbon exchange in Thuringia, Germany. *Agricultural and Forest Meteorology*, **121**:55-67.

Arora, V.K., 2003. Simulating energy and carbon fluxes over winter wheat using coupled land surface and terrestrial ecosystem models. *Agricultural and Forest Meteorology*,

118:21-47.

Baldocchi, D.D., Wilson, K.B., 2001. Modelling CO_2 and water vapor exchange of a temperate broadleaved forest across hourly to decadal time scales. *Ecological Modelling*, **142**:155-184.

Bosveld, F.C., Bouten, W., 2001. Evaluation of transpiration models with observations over a Douglas-fir forest. *Agricultural and Forest Meteorology*, **108**:247-264.

Demuth, H., Beale, M., 1994. Neural Network Toolbox for Use with MATLAB. Natick, The MathWorks, Inc.

Duan, K., Keerthi, S., Poo, A., 2001. Evaluation of Simple Performance Measures for Tuning SVM Hyperparameters (Tech. Rep. No. Control Division Technical Report CD-01-11). Department of Mechanical Engineering, National University of Singapore.

Falge, E., Baldocchi, D., Olson, R., Anthoni, P., Aubinet, M., Bernhofer, C., Burba, G., Ceulemans, R., Clement, R., Dolman, H., *et al*., 2001. Gap filling strategies for defensible annual sums of net ecosystem exchange. *Agricultural and Forest Meteorology*, **107**:43-69.

Flerchinger, G.N., Pierson, F.B., 1991. Modelling plant canopy effects on variability of soil temperature and water. *Agric. and Forest Meteor.*, **56**:227-246.

Flerchinger, G.N., Hanson, C.L., Wight, J.R., 1996. Modelling of evapotranspiration and surface energy budgets across a watershed. *Water Resour. Res.*, **32**(8):2539-2548.

Franks, S.W., Beven, K., 1999. Conditioning a multiple patch SVAT model using uncertain time-space estimates of latent heat fluxes as inferred from remotely sensed data. *Water Resour. Res.*, **35**:2751-2761.

Franks, S.W., Beven, K.J., Quinn, P.F., Wright, I.R., 1997. On the sensitivity of soil-vegetation-atmosphere transfer (SVAT) schemes: Equifinality and the problem of robust calibration. *Agric. Forest Meteor.*, **86**:63-75.

Granier, A., Ceschia, E., Damesin, C., Dufrêne, E., Epron, D., Gross, P., Lebaube, S., Le Dantec, V., Le Goff, N., Lemoine, D., *et al*., 2000a. The carbon balance of a young beech forest. *Funct. Ecol.*, **14**:312-325.

Granier, A., Biron, P., Lemoine, D., 2000b. Water balance, transpiration and canopy conductance in two beech stands. *Agric. Forest Meteor.*, **100**:291-308.

Huntingford, C., Cox, P.M., 1997. Use of statistical and neural network techniques to detect how stomatal conductance responds to changes in the local environment. *Ecol. Model.*, **97**:217-246.

Jemwa, G.T., Aldrich, C., 2003. Identification of Chaotic Process Systems with Least Squares Support Vector Machines. Neural Networks. Proceedings of the International Joint Conference on Volume 3, p.20-24.

Kelliher, F.M., Hollinger, D.Y., Schulze, E.D., Vygodskaya, N.N., Byers, J.N., Hunt, J.E., McSeveny, T.M., Milukova, I., Sogachev, A.F., Varlagin, A.V., *et al*., 1997. Evaporation from an eastern Siberian larch forest. *Agric. Forest Meteor.*, **85**:135-147.

Kosko, B., 1992. Neural Networks and Fuzzy Systems. A Dynamical Systems Approach to Machine Intelligence. New Jersey, Prentice-Hall, Inc, Englewood Cliffs, p.449.

Mo, X.G., Beven, K., 2004. Multi-objective parameter conditioning of a three-source wheat canopy model. *Agricultural and Forest Meteorology*, **122**:39-63.

Pelckmans, K., Suykens, J.A.K., Van Gestel, T., De Brabanter, J., Lukas, L., Hamers, B., Moor, B., Vandewalle, J., 2002. A Matlab/C Toolbox for Least Squares Support Vector Machines. Internal Report 02-44, ESAT-SISTA and K.U. Leuven, Belgium.

Schulz, H., Härtling, S., 2003. Vitality analysis of Scots pines using a multivariate approach. *Forest Ecology and Management*, **186**:73-846.

Suykens, J.A.K., 2001. Nonlinear Modelling and Support Vector Machines. Budapest, Hungary. IEEE Instruments and Measurement Technology Conference.

Thissen, U., van Brakel, R., de Weijer, A.P., Melssen, W.J., Buydens, L.M.C., 2003. Using support vector machines for time series prediction. *Chemometrics and Intelligent Laboratory Systems*, **69**:35-49.

Unland, H.E., Houser, P.R., Shuttleworth, W.J., Yang, Z.L., 1996. Surface flux measurement and modelling at a semi-arid Sonoran Desert site. *Agricultural and Forest Meteorology*, **82**:119-153.

Valentini, R., Deangelis, P., Matteucci, G., Monaco, R., Dore, S., Mugnozza, G.E.S., 1996. Seasonal net carbon dioxide exchange of a beech forest with the atmosphere. *Global Change Biol.*, **2**:199-208.

Van Wijk, M.T., Bouten. W., 1999. Water and carbon fluxes above European coniferous forests modeled with artificial neural networks. *Ecological Modelling*, **20**:181-197.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Sringer-Verlag, New York, p.311.

Vapnik, V., 1998. Statistical Learning Theory. John Wiley and Sons, New York.

Vapnik, V., 1999. The Nature of Statistical Learning Theory. 2nd Ed., Springer-Verlag, New York.

Witten, I. H., Frank, E., 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Diego, CA: Morgan Kaufmann.