Energy Conversion and Management 75 (2013) 311–318

Contents lists available at SciVerse ScienceDirect

Energy Conversion and Management

journal homepage: www.elsevier.com/locate/enconman

Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration

Ji-Long Chen [a,b,c,d], Guo-Sheng Li [c,d,*], Sheng-Jun Wu [a,b]

[a] Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 401122, China
[b] Three Gorges Institute of Ecological Environment, Chinese Academy of Sciences, Chongqing 401122, China
[c] Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[d] Key Laboratory of Coastal Wetland Biogeosciences, China Geologic Survey, Qingdao 266071, China

ARTICLE INFO

Article history:
Received 29 March 2012
Accepted 19 June 2013

Keywords:
Daily solar radiation
Sunshine duration
Models
Support vector machine
Liaoning province

ABSTRACT

Estimation of solar radiation from sunshine duration offers an important alternative in the absence of measured solar radiation. However, due to the dynamic nature of atmosphere, accurate estimation of daily solar radiation has been being a challenging task. This paper presents an application of Support vector machine (SVM) to estimation of daily solar radiation using sunshine duration. Seven SVM models using different input attributes and five empirical sunshine-based models are evaluated using meteorological data at three stations in Liaoning province in China. All the SVM models give good performances and significantly outperform the empirical models. The newly developed model, SVM1 using sunshine ratio as input attribute, is preferred due to its greater accuracy and simple input attribute. It performs better in winter, while highest root mean square error and relative root mean square error are obtained in summer. The season-dependent SVM model is superior to the fixed model in estimation of daily solar radiation for winter, while consideration of seasonal variation of the data sets cannot improve the results for spring, summer and autumn. Moreover, daily solar radiation could be well estimated by SVM1 using the data from nearby stations. The results indicate that the SVM method would be a promising alternative over the traditional approaches for estimation of daily solar radiation.

1. Introduction

Solar radiation arriving on the earth's surface plays an important role in energy balance of the Earth–Atmosphere system. A good knowledge of daily solar radiation is essential for many applications, such as agricultural, ecological, hydrological and soil–vegetation–atmosphere transfer models [1]. However, solar radiation is measured at a very limited number of meteorological stations in the world. This is mainly due to the cost of measuring equipment and its difficult maintenance and calibration [2]. Numerous methods have been developed to estimate solar radiation for the places where no measured values are available. The most common method is to determine this parameter by empirical correlations using other measured meteorological variables such as sunshine duration [3,4] and air temperature [5,6]. These variables are easily and reliably measured, and data are widely available. It is generally recognized that models based on sunshine duration are more accurate than models involving other standard meteorological observations [7–11].

The first attempt at estimation of solar radiation from sunshine duration is carried out by Angstrom [3] who proposed a simple linear relationship between the ratio of actual global radiation to the corresponding value on a clear day and the ratio of actual sunshine duration to the maximum possible sunshine duration. By replacing the solar radiation on a clear day with the extra-terrestrial radiation, Prescott [4] proposed a modified version with the following form (hereafter A–P model):

$$\frac{Rs}{Ra} = a\frac{S}{So} + b \quad (1)$$

where Rs is actual solar radiation, Ra is extra-terrestrial solar radiation, S is actual sunshine duration, So is potential sunshine duration, a and b are empirical parameters which were suggested as 0.54 and 0.22, respectively, by Prescott [4].

Several modified versions were subsequently made since the development of A–P model. Ögelman et al. [12] developed a quadratic form which performed well in Adana and Ankara, Turkey. Bahel et al. [13] proposed a cubic form based on data of 48 stations around the world, with varied meteorological conditions and a wide distribution of geographic locations. Almorox and Hontoria [14] presented an exponential model that fitted the data adequately and can be used to estimate monthly mean daily solar r- a- d-

* Corresponding author at: Key Laboratory of Coastal Wetland Biogeosciences, China Geologic Survey, Qingdao 266071, China. Tel./fax: +86 010 6488 8338.
E-mail address: ligscas@126.com (G.-S. Li).

0196-8904/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.enconman.2013.06.034

iation in Spain. Bakirci [15] added an exponential term to the original A–P equation to obtain a linear exponential model. Ampratwum and Dorvlo [16] presented a logarithmic equation which performed well in an arid region of Oman. Newland [17] revised the original A–P model and suggested a linear logarithmic equation. In these modifications, the basic structure of A–P model is changed from linear to nonlinear form. Although some authors claimed that their revised models performed better than the original A–P model [12,15,17]. It is noted that they are only slightly better than the latter. Many comparative studies preferred the simple A–P model due to its greater simplicity, wider application and very similar performances to its revised versions [1,10,14,18,19].

Accurate estimation of solar radiation has been being a major goal for solar energy practitioners, climatologists and all concerned scientists. Great efforts have been made to model the relationship between solar radiation and sunshine duration since the pioneering work of Angstrom [3]. It seems that A–P equation and its revised versions, as well as other empirical models, have far overreached its predictive limits. Therefore, it is important and urgent to explore new approaches that could produce more accurate long-term solar radiation data. This is perhaps more critical now than ever before, because of the more and more concerns on solar energy utilization and energy conservation, and the increasing needs of accurate long-term solar radiation data for studying the global climate change [20]. Recently, a novel machine learning method, support vector machine (SVM), has been widely applied in computer, environment and hydrology researches [21–23]. A number of studies have proved that SVM shows better performance than Neural Networks and traditional statistical models [24–27]. Despite successes in many fields, there is no application of SVM in estimation of daily solar radiation using sunshine duration. With respect to the rising popularity of SVM, the main objectives of this study are (1) to examine the feasibility of SVM and explore its potential in estimation of daily solar radiation using sunshine duration; and (2) to compare the SVM models with empirical models.

2. Materials and method

2.1. Study sites and data

Liaoning province with an area of 145.900 km^2, situated in the southern part of the Northeast China, is a leading province in respect of agricultural productivity in China. It is characterized by a continental monsoon climate, with a hot, rainy summer; and a short, windy spring. The mean annual temperature is 4–10 °C. The rainfall is rather concentrated, with a mean annual precipitation of 400–1000 mm. The topography, soils and the climate in the province are quite favorable to agriculture, and hence the crop simulations are widely studied [28–30]. However, only 3 meteorological stations, Chaoyang (41°33′N, 120°27′E), Dalian (38°54′N, 121°38′E) and Shengyang (41°44′N, 123°31′E), are measuring daily solar radiation, while about 50 stations have records of sunshine duration. Therefore, solar radiation estimation using sunshine duration is of vital importance and significance. However, only a few works have been reported for this area. Chen et al. [31,32] validated the A–P model and the cubic form using daily solar radiation and sunshine duration data from 1994 to 1998 at Shengyang, the results showed that the two models returned same Nash–Sutcliffe equation (NSE). Zhou et al. [33] compared the the A–P model, quadratic and cubic form using long-term daily data sets at Chaoyang, Dalian and Shengyang, and also reported that the differences between the errors of the three models were very small. These results further indicate that the A–P model and its revised versions have far overreached the predictive limits, and suggest the need to explore new approaches that could produce more accurate long-term solar radiation data.

Therefore, the feasibility of SVM and its potential in estimation of daily solar radiation using sunshine duration are explored for this area. All the three stations with available daily solar radiation and sunshine duration are used in this work. For each station, a total of 35 years (1976–2010) daily sunshine duration and solar radiation data are used. Solar radiation is measured by pyranometer, and sunshine duration is measured by Jordan sunshine recorder, all the instruments are calibrated periodically and all the measurements are made following the guide of the World Meteorological Organization. The data were obtained from the National Meteorological Information Center, China Meteorological Administration. Preliminary quality control tests were conducted by the suppliers. We further checked the data and removed days according to the following criterions: (a) records with missing data which were replaced by 32,766, (b) daily solar radiation larger than the daily extra-terrestrial solar radiation, (c) daily sunshine duration larger than daily potential sunshine duration. Subsequently, two data sets were created for each station. About 70% (1976–2000) of the total records were used for training SVM models and calibrating the empirical models, and the remainder (2001–2010) for validation.

2.2. Theory of support vector machine

A brief introduction to the theory of SVM is presented in this section. More detailed information could be found in Vapnik [34,35]. The SVM, developed by Vapnik and his coworkers, is a supervised learning model with associated learning algorithm that analyzes data and recognizes patterns, widely used in classification, regression and forecasting [36–39]. The advantages of the SVM technique can be summarized as follows. SVM is based on statistical learning theory and principle of structural risk minimization, which attempts to minimize an upper bound of generalization error rather than minimize the local training error. This is the most significant difference from the commonly used principle of empirical risk minimization, which is used by the traditional learning machine methods [35]. SVM estimates the regression using a set of kernel functions which are defined in a high dimensional feature space, the kernel functions implicitly contain non-linear transformation, no assumptions about the functional form of the transformation, which makes data linearly separable, is necessary. SVM delivers a unique solution, since the optimality problem is convex. This is an advantage compared to Neural Networks, which have multiple solutions associated with local minima [40].

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 data size), SVM approximates the function using the following form:

$$f(x) = w\varphi(x) + b \tag{2}$$

where $\varphi(x)$ is the high dimensional feature space which is mapped from the input space x. w and b are estimated by minimizing the regularized risk function below [34]:

$$R_{SVMs}(C) = C\frac{1}{n}\sum_{i=1}^{n} L(d_i, y_i) + \frac{1}{2}\|w\|^2 \tag{3}$$

where $C\frac{1}{n}\sum_{i=1}^{n} L(d_i, y_i)$ is empirical error (risk) and measured by function $L\varepsilon$ given below:

$$L\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon |d - y| \geqslant \varepsilon \\ 0 \quad \text{otherwise} \end{cases}$$

The term $\frac{1}{2}\|w\|^2$ is the regularization term. C is the penalty parameter of the error and used to control 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. ε 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 ζ_i and ζ_i^* as follows [35]:

$$\text{Minimize } R_{SVMs}(w,\zeta^{(*)}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\zeta_i+\zeta_i^*) \tag{5}$$

$$\text{Subjected to } \begin{matrix} d_i - w\varphi(x_i) - b_i \leqslant \varepsilon + \zeta_i \\ w\varphi(x_i) + b_i - d_i \leqslant \varepsilon + \zeta_i^*, \zeta_i^* \geqslant 0 \end{matrix}$$

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

$$f(x,a_i,a_i^*) = \sum_{i=1}^{n}(a_i-a_i^*)K(x,x_i)+b \tag{7}$$

The term $K(x_i,x_j)$ is called kernel function, the value of kernel function $K(x_i,x_j)$ is equal to the inner product of two vectors x_i and x_j in the feature space $\varphi(x_i)$ and $\varphi(x_j)$, that is, $K(x_i,x_j) = \varphi(x_i)^*\varphi(x_j)$. The elegance of using the kernel function is that one can deal with feature spaces of arbitrary dimensionality without having to compute the map $\varphi(x)$ explicitly. Any function satisfying Mercer's condition can be used as kernel function.

One of the key tasks in SVM estimation of global solar radiation is the selection of the input attributes. In this paper, 7 SVM models (Table 1) using different combinations of actual and potential sunshine duration as input attributes are explored. The model (SVM1) using sunshine ratio is tested firstly, since sunshine ratio is widely recognized as the most useful variable for solar radiation estimation. SVM4-7 have the same variable expressions to the corresponding exponential, quadratic, linear exponential and cubic forms of the revised A–P models, therefore, they can be compared to the corresponding empirical model to assess the superiority of SVM model. Furthermore, comparison of the 7 SVM models can propose an optimal SVM model for estimation of daily solar radiation using sunshine duration.

2.3. Implementation of SVM models

The SVM models are implemented by LIBSVM which is an integrated software for support vector classification, regression and distribution estimation [41]. The procedures include data scaling, selection of kernel function, cross-validation and training, and testing, more detailed information could be found in Hsu et al. [42].

Data scaling can avoid values in greater ranges dominate those in smaller ranges and avoid calculation difficulties; it can improve data fitting and estimation performance, because kernel values usually depend on the inner products of feature vectors. In this work, both the training and testing data are scaled to the range [0,1] using the same linearly scaling method before applying SVM.

Four basic kernel functions are provide by SVM, namely, linear, polynomial, radial basis function and sigmoid. The linear kernel function is a special case of radial basis function [43]. The polynomial kernel function has more hyperparameters which influence the complexity of SVM model [42]. The sigmoid kernel function is not valid under some conditions [34]. Therefore, the radial basis function $K(x_i,x_j) = \exp(-\gamma\|x_i-x_j\|^2)$ is a reasonable first choice of kernel function, where γ isthe kernel parameter. Radial basis function nonlinearly maps the input data into a high dimensional feature space, so it can handle the nonlinear relation between the independent and dependent variables. In addition, radial basis function has less calculation difficulties [44,45].

Two critical parameters, the penalty parameter C and the kernel parameter γ, have significantly effect on the accuracy of SVM model, and therefore it is important to select the proper values for C and γ. In this work, the optimum ranges of C and γ are determined by grid search [46], and the optimum values of parameters are then obtained from range using cross validation [42,47], in which the training data are divide into several subsets with equal size. Sequentially, each subset is tested using the trained SVM model on the remaining subsets. Thus, each instance of the whole training set is estimated once so the cross validation accuracy is the percentage of data which are correctly estimated. Finally, after the optimum values of C and γ are found, the final model, which is used for testing and for estimating the new data, is then trained on the whole training data set using the selected parameters [42]. Subsequently, the final estimating models are used to predict the daily solar radiation using the testing data, the estimation and the observation can be compared and the errors can be calculated. The summary of results of SVM models is presented in Table 4.

Table 1
The studied SVM models with different input attributes.

Model	Input attributes[a]
SVM1	S/So
SVM2	S, So
SVM3	S, 1/So
SVM4	Exp(S/So)
SVM5	S/So, S/So^2
SVM6	S/So, exp(S/So)
SVM7	S/So, S/So^2, S/So^3

[a] S and So are daily sunshine duration and potential sunshine duration, respectively.

2.4. Empirical models

A number of formulas have been developed to estimate solar radiation using sunshine duration. The first one (A–P model) was proposed by Angstrom [3], which was further modified by Prescott [4]. Subsequently, several modified versions of A–P model centered on improvement in estimation accuracy have been developed. In this work, five of the models (Table 2), which were previously used by researchers commonly, are used to estimate daily solar radiation. The empirical parameters of them are calibrated by least square regression of the dependent (Rs/Ra) on independent variables to determine the parameters of the model that best describes the relationship between expected and measured data sets by minimizing the sum of the squared residuals.

A common feature of these empirical models is that they account for latitude, solar declination, elevation, day length and atmospheric transmissivity by including the extra-terrestrial radiation (Ra) and potential sunshine duration (So) in the model. They are calculated using the equations detailed by Allen et al. [48].

$$\text{Ra} = 37.6d(\omega\sin\varphi\sin\delta + \cos\varphi\cos\delta\sin\omega) \tag{8}$$

$$d = 1 + 0.033\cos\left(\frac{2\pi}{365}n\right) \tag{9}$$

$$\delta = 0.4093\sin\left(\frac{2\pi}{365}n - 1.39\right) \tag{10}$$

$$\omega = \arccos(-\tan\varphi\tan\delta) \tag{11}$$

$$So = 24\omega/\pi \tag{12}$$

where d is the relative distance between the sun and the earth, ω is sunset hour angle (rad), φ is latitude (rad), δ is solar declination

Table 2
Regression models proposed in literatures.

Models	Equation[a]	Parameter	Source
Linear	$Rs = Ra(a_1 S/So + b)$	a_1, b	[3] and [4]
Quadratic	$Rs = Ra(a_1 S/So + a_2 S/So^2 + b)$	a_1, a_2, b	[12]
Cubic	$Rs = Ra(a_1 S/So + a_2 S/So^2 + a_3\ S/So^3 + b)$	a_1, a_2, a_3, b	[13]
Exponential	$Rs = Ra(a_4 \exp(S/So) + b)$	a_4, b	[14]
Linear exponential	$Rs = Ra(a_1 S/So + a_4 \exp(S/So) + b)$	a_1, a_4, b	[15]

[a] Rs, Ra, S and So are daily solar radiation, extra-terrestrial solar radiation, sunshine duration and potential sunshine duration, respectively.

angle (rad), n is the number of the day of year starting from the first of January.

2.5. Performance criteria

Root mean square error (RMSE), relative root mean square error (RRMSE) (%) and coefficient of determination (R^2) are used to assess the performance of models. R^2 is commonly calculated based on the calibration dataset, RMSE and RRMSE based on the validation dataset. The metric R^2 varying between 0 and 1 is adopted to measure the fit of model, higher the value, better the fit. While the lower the values of RMSE, and RRMSE, the better is the model's performance. RMSE and RRMSE are calculated by the following equations.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{13}$$

$$\mathrm{RRMSE} = \frac{100}{\bar{y}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{14}$$

where n, y, $\hat{y}$ and $\bar{y}$ represent the number of testing data, the observed value, the estimated value and the average value of the observation, respectively.

3. Results

3.1. Performances of empirical models

The locally calibrated parameters and performances of the empirical models are presented in Table 3. All the models perform well with RMSE < 2.8 MJ m^{-2} (average 2.323 MJ m^{-2}), RRMSE < 20% (average 17.31%) and $R^2 > 0.6$ (average 0.759). The linear, linear exponential, quadratic and cubic models give very similar RMSE, RRMSE and R^2 at each station. The exponential model returns the highest RMSE (average 2.401 MJ m^{-2}), RRMSE (average 19.95%) and lowest R^2 (average 0.718).

Among the modified versions of A–P model, the exponential form returns relatively higher RRMSE, RRMSE and lower R^2 when compared to the others, indicating less suitable for the present sites. The poor performance was also reported by others [14,15,19,49]. Generally, for empirical modes such as those presented here, the more parameter a model has, the higher chance the model gives a better performance. However, in the present work, the linear exponential, quadratic and cubic models give very similar results to the linear A–P model. The largest differences between the RMSE, and R^2 of the best and A–P model are only 0.006 MJ m^{-2}, and 0.004 for Chaoyang, 0.032 MJ m^{-2}, and 0.002 for Shengyang, 0.019 MJ m^{-2}, and 0.008 for Dalian, respectively. Our results are consistent with those reported by Chen et al. [31,32] who reported that the A–P model and cubic form returned same NSE, and by Zhou et al. [33] who found the differences between the errors of A–P model, quadratic and cubic form were very small at Chaoyang, Dalian and Shengyang. Many comparative studies also reported similar results [1,10,14,49], for example, Yorukoglu and Celik [19] found the quadratic and cubic models returned similar values of R^2 and RMSE to those of the linear model. The cubic model performed even worst at 159 stations in Turkey [49]. Only a few works presented better performances of the revised A–P models, for example, Bakirci [15] compared the A–P model and its revised forms for provinces in different regions of Turkey and found the the linear exponential gave generally the best results.

Given the slightly differences between the variance explained by A–P model and its modified versions, the original A–P model is always preferred due to its greater simplicity and wider applications [1,10,14,18,19]. The reason for lack of significant improvement has not been widely discussed. We believe that one possible reason may be due to the highly significant correlations of different terms. For example, S/So correlates significantly with

Table 3
Calibrated parameters and statistical indicators (RMSE in MJ m^{-2}, RRMSE and R^2) of the studied empirical models.

Station	Model	b	a_1	a_2	a_3	a_4	R^2	RMSE	RRMSE (%)
Chaoyang	Linear	0.206	0.510				0.810	2.652	19.42
	Exponential	−0.059				0.301	0.751	2.729	19.95
	Linear exponential	0.297	0.677			−0.101	0.812	2.658	19.47
	Quadratic	0.194	0.598	−0.091			0.812	2.658	19.47
	Cubic	0.185	0.843	−0.771	0.476		0.814	2.646	19.44
Dalian	Linear	0.143	0.522				0.673	1.994	14.91
	Exponential	−0.132				0.308	0.632	2.101	15.70
	Linear exponential	0.173	0.577			−0.033	0.673	1.992	14.90
	Quadratic	0.139	0.556	−0.036			0.673	1.991	14.89
	Cubic	0.130	0.801	−0.712	0.468		0.681	1.975	14.77
Shengyang	Linear	0.165	0.550				0.820	2.281	17.27
	Exponential	−0.138				0.335	0.774	2.374	17.97
	Linear exponential	0.235	0.672			−0.076	0.821	2.271	17.20
	Quadratic	0.158	0.612	−0.068			0.821	2.269	17.18
	Cubic	0.151	0.795	−0.601	0.388		0.822	2.249	17.03

Table 4
performance of the SVM models (RMSE in MJ m^{-2}).

Model	Chaoyang		Dalian		Shengyang	
	RMSE	RRMSE (%)	RMSE	RRMSE (%)	RMSE	RRMSE (%)
SVM1	2.302	16.86	1.801	13.47	2.003	15.16
SVM2	2.378	17.42	2.199	16.44	2.126	16.10
SVM3	2.380	17.43	2.196	16.42	2.140	16.20
SVM4	2.307	16.90	1.808	13.52	2.009	15.21
SVM5	2.304	16.88	1.812	13.55	2.005	15.18
SVM6	2.305	16.88	1.802	13.47	2.007	15.20
SVM7	2.302	16.87	1.789	13.37	1.989	15.06

S/So^2 at Chaoyang ($r = 0.969$, $p < 0.001$), Dalian ($r = 0.971$, $p < 0.001$) and Shengyang ($r = 0.967$, $p < 0.001$).

3.2. Performances of SVM models

Table 4 displays the performances of the SVM models. Overall, all the SVM models give good performances with RMSE < 2.4 MJ m^{-2} (average 2.094 MJ m^{-2}) and RRMSE < 18% (average 15.6%), indicating that SVM shows good generalization. SVM1, 4-7 which include sunshine ratio in input attribute perform better than SVM2-3 with an average 9% lower RMSE, suggesting that the sunshine ratio is also an important variable for solar radiation estimation using SVM. Whereas SVM1, 4-7 give very similar performances at each station, mean values of RMSE are 2.035, 2.041, 2.039, 2.038 and 2.027 MJ m^{-2}, respectively; mean values of RRMSE are 15.16%, 15.21%, 15.2%, 15.18% and 15.1%, respectively. The largest differences between RMSE of the best and the worst are only 0.005 MJ m^{-2} for Chaoyang, 0.02 MJ m^{-2} for Shengyang, and 0.024 MJ m^{-2} for Dalian. SVM5-7 have more input parameters than SVM1, which are generally expected to give better results. While the similar performances of them indicate that more input parameters do not adequately contribute to the improvement in estimation accuracy. The reason for lack of improvement may be attributed to the significant correlations of different attributes, which is similar to the reason for lack of improvement of the revised A–P models. Therefore, SVM1 is proposed due to its accuracy and the simple input attribute.

3.3. Comparison of SVM between empirical models

To demonstrate the superiority of SVM, comparison of SVM models between empirical models is made. Overall, SVM models produce an average RMSE of 2.094 MJ m^{-2} which is 10% lower than that of 2.323 MJ m^{-2} produced by empirical models, suggesting that SVM outperforms the empirical models. The improvements of SVM differ at different sites, at Chaoyang SVM models produce an average RMSE of 2.325 MJ m^{-2} which is 12.9% lower than that of 2.668 MJ m^{-2} produced by empirical models, while the average RMSE of SVM are 10.8% and 7.7% lower than those of empirical models at Shenyang and Dalian., respectively. The improvements of SVM also differ from model to model. The best improvements are obtained by SVM4 which shows 15.4%, 13.9% and 15.1% lower RMSE than the exponential model at Chaoyang, Dalian and Shengyang, respectively. SVM1 also makes better improvements with 13.2%, 9.7% and 12.2% lower RMSE than the widely used A–P model at Chaoyang, Dalian and Shengyang, respectively. While similar improvements are observed for SVM5-7 which produce an average RMSE of 2.04, 2.038, and 2.027 MJ m^{-2}, corresponding to 11.5%, 11.7%, and 11.4% lower than 2.306, 2.307, and 2.29 MJ m^{-2} produced by quadratic, linear exponential, and cubic models, respectively. These results further confirm the superiority of SVM over the empirical models. The remarkable improvement indicates that the SVM method would be a promising alternative over the traditional approaches for estimation of daily solar radiation using sunshine duration.

3.4. Performances of SVM1 with season-dependent data sets

As SVM1 is preferred to other SVM models, the following analysis will be limited to the results of SVM1. Further investigation is carried out to investigate the effect of the training timescale on the results of the estimation. The original training data sets were divided into 4 groups corresponding to four seasons: spring (March, April and May), summer (June, July and August), autumn (September, October and November) and winter (December, January and February). The seasonal data sets for validation were created in the same way. Subsequently, the seasonal data sets were used for training SVM1 (hereafter season-dependent model). The seasonal validation data sets were used for testing the corresponding season-dependent model and the SVM model (hereafter fixed model) trained on the data sets of all seasons. The results are presented in Fig. 1.

Obviously, both season-dependent and fixed models perform differently in different seasons. In terms of RMSE, they perform better in winter at each station. The highest RMSE and RRMSE are obtained in summer, which may be attributed to the large daily fluctuation of the weather variables, since this season also corresponds to the rainy season in Liaoning province.

The season-dependent model provides better results than the fixed model in winter, with 11.6%, 10.5% and 10.9% lower RMSE for Chaoyang, Dalian and Shengyang, respectively. While they give similar results in other seasons. Mean values of RRMSE obtained by fixed model are 13.83%, 16.01% and 13.33% for spring, summer and autumn, respectively. Which are very similar to the corresponding values of 13.89%, 15.91% and 13.34% by the seasonal-dependent model. It is interesting that at Shengyang the fixed model even obtains slightly better results for spring and summer. Similar results are also found in autumn for Dalian and in summer for Chaoyang. These results indicate that the time-dependent SVM model has no advantage over the fixed model for estimation of daily solar radiation in spring, summer and autumn, In other words, consideration of seasonal variation of the training data sets cannot improve the model performance, and thus unnecessary.

The season-dependent characteristic of empirical models, in particular the A–P model, is also widely discussed. A few studies reported the advantage of the season-dependent models. Podestá et al. [8] suggested that the seasonal parameters seemed to be required to avoid systematic temporal patterns in estimation residuals in the Argentine Pampas. Tymvios et al. [50] stresses the significance of seasonality as an important factor in estimating solar radiation in Athalassa. While there are many others presented different conclusions. Iziomon and Mayer [18] found that the parameters of the A–P model did not show any particular trend with respect to season at Bremgarten and Feldberg in south-west Germany. Almorox and Hontoria [14] reported that seasonal partitioning did not significantly improve the estimation. Recent studies of Liu et al. [51] and Liu et al. [52] explicitly demonstrated that consideration of the season-dependent characteristics of the A–P model had little predictive value and thus unnecessary. This is partly consistent with the result of time-dependent SVM model which gives similar performance to the fixed model in spring, summer and autumn.

3.5. Analysis of interchangeability for SVM1 of different stations

In the case of unavailable daily solar radiation data for a site, the predictability of SVM1 is further investigated, and the results are presented in Fig. 2. As it can be seen, good agreement has been ob-

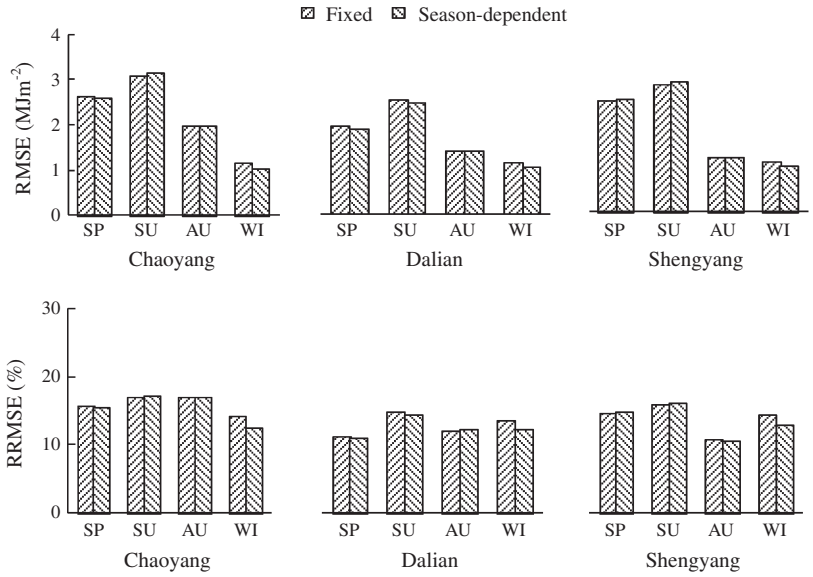

Fig. 1. Performance of the fixed and season-dependent SVM models. SP, SU, AU and WI represent for spring, summer, autumn and winter, respectively.

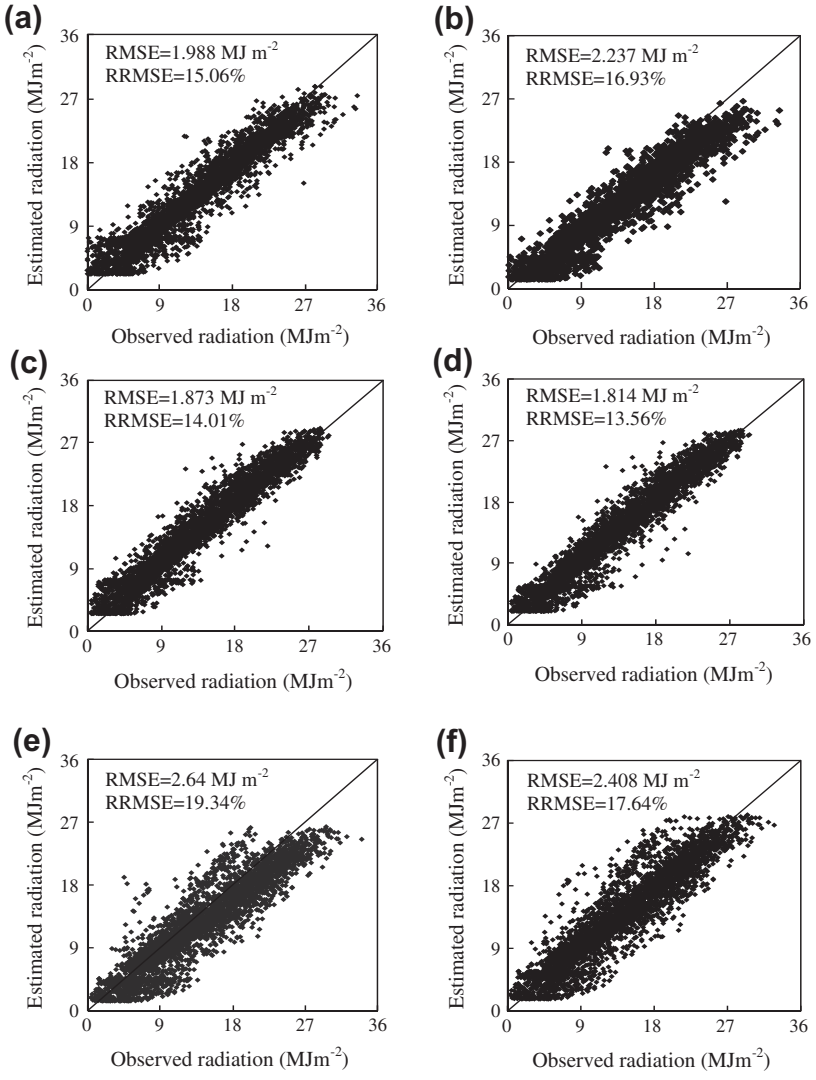

Fig. 2. Scatter plots of the observed vs. estimated solar radiation by SVM1 using the data from other stations (a and b are results for Shengyang using the data from Chaoyang and Dalian, respectively, c and d for Dalian using the data from Chaoyang and Shengyang, respectively, e and f for Chaoyang using the data from Dalian and Shengyang, respectively).

served between estimated and measured values, as points tend to line up around the 1:1 line. These results illustrate a good reliability of SVM1 in estimating daily solar radiation using the training data sets from other stations.

At Chaoyang (Fig. 2f) and Dalian (Fig. 2d) SVM1 using the data from Shengyang gives good results with the RRMSE of 17.64% and 13.56%, respectively. Which are similar to the corresponding values of 16.86% and 13.47% derived by their own data (Table 4). Similar results are also found when the data from Chaoyang are applied to Dalian (Fig. 2c). It is interesting that at Shengyang (Fig. 2a) SVM1 using the data from Chaoyang provides slightly better results than using its own data (Table 4). At Chaoyang (Fig. 2e) and Shengyang (Fig. 2b) SVM1 using the data from Dalian return relatively higher RMSE and RRMSE, but reasonable results. These results indicate that daily solar radiation could be well estimated by SVM1 using the data from other stations although the distances between stations are up to 340 km. This would be of significance because all area of the province is covered by the 3 meteorological stations.

It is an alternative to substitute solar radiation by those from nearby stations [2,9,53]. It was proposed by Allen et al. [48] as a possible way of replacing missing data when calculating the daily evapotranspiration. The precision of this method decreases with the increase in distance between the two stations. Hunt et al. [2] reported that the RMSE between replacement and measured values increased with increase in distance in a curvilinear manner. The study of Trnka et al. [9] also presented same conclusion. In our study, when the training data from other stations are applied to Shengyang, RMSE obtained by SVM1 increases from 1.988 MJ m^{-2} (Fig. 2a) to 2.237 MJ m^{-2} (Fig. 2b) with increase in distance from 301 km (between Shengyang and Chaoyang) to 340 km (between Shengyang and Dalian). At Dalian using the data form Chaoyang gives the RMSE of 1.873 MJ m^{-2} (Fig. 2c), which is similar to 1.814 MJ m^{-2} (Fig. 2d) obtained using the data from Shengyang although the distance increases from 285 km (between Dalian and Chaoyang) to 340 km. While RMSE obtained at Chaoyang decreases from 2.64 MJ m^{-2} (Fig. 2e) to 2.408 MJ m^{-2} (Fig. 2f) as increase in distance from 285 km to 301 km. It seems that RMSE shows no generic trend with the distance between stations providing data and the stations for which the data are used for SVM1 to estimate daily solar radiation. Admittedly, as the solar radiation data are spatially very sparse in the study area, and the range of distances between stations is narrow, it is difficult to plot the precision vs. distance relationship. Further study will focus on the reliability of SVM models to estimate solar radiation on large spatial scales with denser network of solar radiation observations, and reveal the precision vs. distance relationship.

4. Conclusions

This work presents the application of SVM to estimating daily solar radiation using sunshine duration. Seven SVM models using different input attributes and five empirical sunshine-based models are evaluated. SVM shows good generalization, and all the SVM models give good performances. The developed SVM models outperform the empirical models. SVM1 using sunshine ratio as input attribute is preferred due to its greater accuracy and simple input attribute. It performs better in winter, while highest RMSE and RRMSE are obtained in summer. The season-dependent SVM model is superior to the fixed one in estimation of daily solar radiation in winter, while consideration of seasonal variation of the data sets cannot improve the results in spring, summer and autumn. In the case of unavailable data for constructing SVM model, the daily solar radiation could be well estimated by SVM1 using the data from nearby stations although the distances are up to 340 km, so that all area of the province is covered by the existing meteorological stations.

Besides the sunshine duration, air temperature, relative humidity, precipitation and atmospheric pressure are also routinely measured at most meteorological stations. Many works have shown that these meteorological variables in combination with the sunshine duration can significantly improve the accuracy of the A–P model [31,54,55]. Lam et al. [56] investigated the Neural Networks for estimating solar radiation using sunshine duration in combination with air temperature and geographical variables, the RMSE is 1.73 MJ m^{-2} and RRMSE is 13.4% for Chaoyang. Apart from the effect of period of data sets, Neural Network even gives better performance than SVM model in this work, this indicates sunshine duration in combination with other routinely measured meteorological variables can also improve the accuracy of machine learning method. Therefore, future works will investigate the performance of SVM with more available meteorological variables and propose the optimal input attributes.

Acknowledgments

The work was supported by National Key Technology Research and Development Program (2012BAC21B01), the Geological Survey program of China Geological Survey (GZH201200503) and Special foundation for scientific research on public Interest (1212010611402, 201111023)and Chongqing Science and Technology Key project (cstc2012ggB20001). We thank the National Meteorological Information Center, China Meteorological Administration for providing the long-term data records. Many thanks go to the anonymous reviewers for the comments on the manuscript.

References

[1] Liu XY, Mei XR, Li YZ, Zhang YQ, Wang QS. Calibration of the Ångström–Prescott coefficients (*a*, *b*) under different time scales and their impacts in estimating global solar radiation in the Yellow River basin. Agric For Meteorol 2009;149:697–710.
[2] Hunt LA, Kuchar L, Swanton CJ. Estimation of solar radiation for use in crop modelling. Agric For Meteorol 1998;91:293–300.
[3] Ångström A. Solar and terrestrial radiation. QJR Meteorolog Soc 1924;50:121–6.
[4] Prescott JA. Evaporation from a water surface in relation to solar radiation. Trans R Soc S Aust 1940;64:114–8.
[5] Hargreaves GH, Samani ZA. Estimating potential evapotranspiration. J Irrig Drain Div 1982;108:225–30.
[6] Bristow KL, Campbell GS. On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agric For Meteorol 1984;31:159–66.
[7] Ertekin C, Yaldiz O. Comparison of some existing models for estimating global solar radiation for Antalya (Turkey). Energy Convers Manage 2000;41:311–20.
[8] Podestá GP, Núñez L, Villanueva CA, Skansi MA. Estimating daily solar radiation in the Argentine Pampas. Agric For Meteorol 2004;123:41–53.
[9] Trnka M, Zalud Z, Eitzinger J, Dubrovský M. Global solar radiation in Central European lowlands estimated by various empirical formulae. Agric For Meteorol 2005;131:54–76.
[10] Wu GF, Liu YL, Wang TJ. Methods and strategy for modeling daily global solar radiation with measured meteorological data – a case study in Nanchang station, China. Energy Convers Manage 2007;48:2447–52.
[11] Chen JL, Li GS. Estimation of monthly average daily solar radiation from measured meteorological data in Yangtze River Basin in China. Int J Climatol 2013;33:487–98.
[12] Ögelman H, Ecevit A, Taşdemiroğlu E. A new method for estimating solar radiation from bright sunshine data. Sol Energy 1984;33:619–25.
[13] Bahel V, Bakhsh H, Srinivasan R. A correlation for estimation of global solar radiation. Energy 1987;12:131–5.
[14] Almorox J, Hontaria C. Global solar radiation estimation using sunshine duration in Spain. Energy Convers Manage 2004;45:1529–35.
[15] Bakirci K. Correlations for estimation of daily global solar radiation with hours of bright sunshine in Turkey. Energy 2009;34:485–501.
[16] Ampratwum DB, Dorvlo ASS. Estimation of solar radiation from the number of sunshine hours. Appl Energy 1999;63:161–7.
[17] Newland FJ. A study of solar radiation models for the coastal region of South China. Sol Energy 1988;31:227–35.
[18] Iziomon MG, Mayer H. Assessment of some global solar radiation parameterizations. J Atmos Sol Terr Phys 2002;64:1631–43.

[19] Yorukoglu M, Celik AN. A critical review on the estimation of daily global solar radiation from sunshine duration. Energy Convers Manage 2006;47:2441–50.
[20] Gueymard C, Jidra P, Eatrada CV. A critical look at recent interpretations of the Ångström approach and its future in global solar irradiation prediction. Sol Energy 1995;54:357–63.
[21] Lee SW, Verri A. Support vector machines for computer vision and pattern recognition. Int J Pattern Recognit Artif Intell 2003;3:331–2.
[22] Lu WZ, Wang W. Potential assessment of the support vector machine method in forecasting ambient air pollutant trends. Chemosphere 2005;59:693–701.
[23] Tirusew A, Mariush K, Mac M, Abedalrazq K. Multi-time scale stream flow predictions: the support vector machines approach. J Hydrol 2006;318:7–16.
[24] Thorsten J. Text categorization with support vector machines: Learning with many relevant features. Machine learning: ECML-98. Lect Notes Comput Sci 1998;1398:137–42.
[25] Mukkamala S, Janoski G, Sung AH. Intrusion detection using neural networks and support vector machines. In: Proceedings of IEEE international joint conference on neural networks. 2002. p. 1702–7.
[26] Mukkamala S, Sung AH. Feature selection for intrusion detection with neural networks and support vector machines. Transport Res Rec: J Transport Res Board 2003;1822:33–9.
[27] Sung AH, Srinivas M. Identifying important features for intrusion detection using support vector machines and neural networks. In: Proceedings of the 2003 symposium on applications and the internet. 2003. p. 209–216.
[28] Liu XB, Herbert SJ, Jin J, Zhang QY, Wang GH. Responses of photosynthetic rates and yield/quality of main crops to irrigation and manure application in the black soil area of Northeast China. Plant Soil 2004;261:55–60.
[29] Tao FL, Yokozawa M, Liu JY, Zhang Z. Climate-crop yield relationships at provincial scales in China and the impacts of recent climate trends. Clim Res 2008;38:83–94.
[30] Chen JL, Li GS. Parameterization and mapping of solar radiation in data sparse regions. Asia-Pacific J Atmos Sci 2012; 48(4): 423–31.
[31] Chen RS, Ersi K, Yang JP, Lu SH, Zhao WZ. Validation of five global radiation models with measured daily data in China. Energy Convers Manage 2004;45:1759–69.
[32] Chen RS, Lu S, Kang E, Yang J, Ji X. Estimating daily global radiation using two types of revised models in China. Energy Convers Manage 2006;47:865–78.
[33] Zhou J, Wu YZ, Yan G. General formula for estimation of monthly average daily global solar radiation in China. Energy Convers Manage 2005;46:257–68.
[34] Vapnik VN. The nature of statistical learning theory. New York: Springer Verlag; 1995.
[35] Vapnik VN. Statistical learning theory. New York: Wiley; 1998.
[36] Vapnik VN, Golowich SE, Smola AJ. Support vector method for function approximation, regression estimation and signal processing. In: Mozer M, Jordan M, Petsche T, editors. Advances in neural information processing systems. Cambridge (MA): MIT press; 1996. p. 281–7.
[37] Collobert R, Bengio S. SVMTorch: support vector machines for large-scale regression problems. J Mach Learn Res 2001;1:143–60.
[38] Tay FEH, Cao LJ. Application of support vector machines in financial time series forecasting. Int J Manage Sci 2001;29:309–17.
[39] Huang C, Davis LS, Townshend JRG. An assessment of support vector machines for land cover classification. Int J Remote Sens 2002;23:725–49.
[40] Auria L, Moro RA. Support vector machines as a technique for solvency analysis. vol. 811. Berlin: German Institute for Economic Research; 2008.
[41] Chang C, Lin C. LIBSVM: a library for support vector machines, 2001. <http://www.csie.ntu.edu.tw/∼cjlin/libsvm>.
[42] Hsu CW, Chang CC, Lin CJ. A practical guide to support vector classification. 2003. <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>.
[43] Keerthi SS, Lin CJ. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Comput 2003;15(7):1667–89.
[44] Dong B, Cheng C, Siew EL. Applying support vector machines to predict building energy consumption in tropical region. Energy Build 2005;37(5):545–53.
[45] Wu KP, Wang SD. Choosing the kernel parameters for support vector machines by the inter-cluster distance in the feature space. Pattern Recognit 2009;42(5):710–7.
[46] Gestel TV, Suykens JAK, Baesens B, Viaene S, Vanthienen J, Dedene G, et al. Benchmarking least squares support vector machine classifiers. Mach Learn 2004;54(1):5–32.
[47] Cherkassky V, Mulier F. Learning from data: concepts, theory, and methods. New York: Wiley; 1998.
[48] Allen RG, Pereira LS, Raes D, Smith M. Crop evapotranspiration-guidelines for computing crop water requirements. In: FAO Irrigation and drainage paper 56. Rome: Food and Agriculture Organization of the United Nations; 1998.
[49] Ertekin C, Evrendilek F. Spatio-temporal modeling of global solar radiation dynamics as a function of sunshine duration for Turkey. Agric For Meteorol 2007;145:36–47.
[50] Tymvios FS, Jacovides CP, Michaelides SC, Scouteli C. Comparative study of Ångström and artificial neural networks methodologies in estimating global solar radiation. Sol Energy 2005;78:752–62.
[51] Liu XY, Mei XR, Li YZ, Porter JR, Wang QG, Zhang YQ. Choice of the Ångström–Prescott coefficients: are time-dependent ones better than fixed ones in modeling global solar irradiance? Energy Convers Manage 2010;51:2565–74.
[52] Liu JD, Liu JM, Linderholm HW, Chen DL, Yu Q, Wu DR, et al. Observation and calculation of the solar radiation on the Tibetan Plateau. Energy Convers Manage 2012;57:23–32.
[53] Nonhebel S. The importance of weather data in crop growth simulation models and assessment of climate change effects. PhD Thesis. Wageningen Agriculture University; 1993. p. 144.
[54] Abdalla YAG. New correlation of global solar radiation with meteorological parameters for Bahrain. Int J Sol Energy 1994;16:111–20.
[55] Trabea AA, Shaltout MAM. Correlation of global solar radiation with meteorological parameters over Egypt. Renew Energy 2000;21:297–308.
[56] Lam JC, Wan KKW, Yang L. Solar radiation modelling using ANNs for different climates in China. Energy Convers Manage 2008;49(5):1080–90.