Estimating atmospheric visibility using synergy of MODIS data and ground-based observations

HOSSEIN KOMEILIAN[2], S. MOHYEDDIN BATENI[3], TONGREN XU[1] & JEFFREY NIELSON[3]

1 State Key Laboratory of Remote Sensing Science, Research Center for Remote Sensing and GIS, and School of Geography, Beijing Normal University, Beijing, 100875, China
xutr@bnu.edu.cn

2 Department of Civil and Environmental Engineering, Tarbiat Modares University, Tehran Province, 11369, Iran

3 Department of Civil and Environmental Engineering and Water Resource Research Center, University of Hawaii at Manoa, Honolulu, HI, 96822, USA

Abstract Dust events are intricate climatic processes, which can have adverse effects on human health, safety, and the environment. In this study, two data mining approaches, namely, back-propagation artificial neural network (BP ANN) and supporting vector regression (SVR), were used to estimate atmospheric visibility through the synergistic use of Moderate Resolution Imaging Spectroradiometer (MODIS) Level 1B (L1B) data and ground-based observations at fourteen stations in the province of Khuzestan (southwestern Iran), during 2009–2010. Reflectance and brightness temperature in different bands (from MODIS) along with *in situ* meteorological data were input to the models to estimate atmospheric visibility. The results show that both models can accurately estimate atmospheric visibility. The visibility estimates from the BP ANN network had a root-mean-square error (RMSE) and Pearson's correlation coefficient (R) of 0.67 and 0.69, respectively. The corresponding RMSE and R from the SVR model were 0.59 and 0.71, implying that the SVR approach outperforms the BP ANN.

Key words atmospheric visibility; MODIS data; back-propagation artificial neural network; supporting vector regression

1 INTRODUCTION

Various adverse impacts on air quality, the environment, and human health have been attributed to dust events (Shao *et al*., 2011). Remote sensing can provide a valuable source of data for dust storm studies due to its temporally- and spatially-wide coverage (Zhao *et al*., 2010; Li *et al*., 2011). Dust events can significantly affect atmospheric visibility, therefore, the intensity of dust events can be characterized by determining atmospheric visibility via remote sensing (Shao *et al*., 2011).

Generally, nonlinear regression and semi-empirical approaches have been used to estimate dust storm parameters, including particulate matter (PM) and atmospheric visibility (Esmaili *et al*., 2006). These methods are not robust and typically lead to erroneous results. In this study, two data mining approaches, namely, back-propagation artificial neural network (BP ANN) and supporting vector regression (SVR), were used to estimate atmospheric visibility at 14 stations in the province of Khuzestan (southwestern Iran) during 2009–2010, through the synergistic use of Moderate Resolution Imaging Spectroradiometer (MODIS) Level 1B (L1B) data and ground-based observations.

2 DATA, METHODS and MODELS

2.1 Data

Data from MODIS were used to estimate atmospheric visibility. Since some meteorological variables (e.g. wind speed, direction, and relative humidity) significantly affect the characteristics of dust events (Wu *et al.*, 2012), MODIS data and ground-based observations were used, synergistically, to increase the accuracy of the estimations.

2.1.1 MODIS The MODIS reflectance in bands 1, 2, 3, 4, 5, 7, 17, 18, 19 and 26, along with brightness temperature in bands 20, 22, 25, 29, 31 and 32, contain useful information about dust events (Hansell *et al*., 2007; Baddock *et al*., 2009; Karimi *et al*., 2011; Shahrisvand *et al*., 2013; Komeilian *et al*., 2014). MODIS L1B data used in this study were obtained from the Atmosphere Archive and Distribution System (LAADS, http://ladsweb.nascom.nasa.gov).

2.1.2 Ground-based observations Meteorological variables, including air temperature (Ta), relative Humidity (Rh), wind speed (Ws) and wind direction (Wd) influence dust event characteristics (Wu *et al*., 2012). These four variables were used to characterize dust events more accurately. Synoptic-scale visibility data were used to train and test the models.

2.2 Study site

Khuzestan Province (32°46′13″N and 48°32′55″E) is located in southwestern Iran. It covers an area of 64 055 km² and has a population of 4 345 607 (2005). Khuzestan Province was chosen as the study region because: (a) it has recently encountered severe dust events; (b) among other provinces in Iran that frequently experience dust events, it is politically and economically the most important one; and (c) ground-based meteorological data were available.

The days that were selected for this study were the days during 2009–2010 when at least half of the meteorological stations in Khuzestan Province recorded a dust-induced reduction in visibility, to less than 1 km (the WMO definition of a dust storm) (Baddock *et al*., 2009). Additionally, only daytime events were used, allowing the influence of surface reflectance on the intensity of dust events to be explored. From this selection process, 27 days with dust events were identified in the study region, and were used in the models (DOY 3, 42, 43, 53, 54, 62, 75, 160, 170, 180, 185, 187, 195, 211, 213 in 2009, and DOY 54, 55, 56, 82, 95, 139, 159, 175, 184, 203, 213, 272 in 2010).

Due to the availability of a large number of studies on the theory of the SVR approach (e.g. Vapnik, 1979; Shawe-Taylor *et al*., 2004; Hastie *et al*., 2008; Noori *et al*., 2011), only a brief explanation of the SVR model is given below. Here the ε-SVR model (also known as Regression SVM type 1) was used to estimate atmospheric visibility, because it is commonly used in many regression problems. SVR estimates the real function as follows:

$$y = f(x) + \delta \tag{1}$$

where δ, x, and y are an independent random noise (defined by ε error tolerance), a multivariable input, and a scalar output. f is a deterministic function of the regression, and is defined by the following equation:

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

where φ is a kernel function, w is a regression function coefficient, and $(.)^T$ denotes a transpose. The aim is to find a functional form of $f(x)$ by training the ε-SVR model (Hastie *et al*., 2008). In the ε-SVR model, the ε-insensitive loss function was used, so that the problem can be written as follows:

$$\text{Minimize } \frac{1}{2}\|\mathrm{W}\|^2 + \mathrm{C}\sum_{i=1}^{N}\xi_i + \mathrm{C}\sum_{i=1}^{N}\xi_i^* \tag{3}$$

$$\text{subject to } \begin{cases} \mathrm{W}^{\mathrm{T}}.\varphi(x_i) + b - y_i \le \varepsilon + \xi_i^* \\ y_i - \mathrm{W}^{\mathrm{T}}.\varphi(x_i) - b \le \varepsilon + \xi_i \\ \xi_i,\ \xi_i^* \ge 0\ ,\mathrm{i} = 1, 2, \ldots, \mathrm{N} \end{cases} \tag{4}$$

where parameter C > 0 determines the tradeoff between the model flatness and the degree to which deviations larger than ε can be tolerated (Shawe-Taylor *et al*., 2004). N, φ, are the sample size and the kernel function. ξ_i and ξ_i^* are slack variables (stating the upper and lower training error, subject to an error tolerance, ε) (Noori *et al*., 2011). To solve equation (3) (subject to the constraints in (4)), dual sets of Lagrange multipliers were used to solve for a and a^*. That allowed the optimization problem to be solved by maximizing equation (5), subject to equation (6):

$$\text{Maximize } \sum_{i=1}^{N} y_i(a_i - a_i^*) - \sum_{i=1}^{N}\varepsilon(a_i + a_i^*) - 0.5\sum_{i,j=1}^{N}(a_i - a_i^*)\left(a_j - a_j^*\right)\varphi(x_i)^{\mathrm{T}}.\varphi\left(x_j\right) \tag{5}$$

$$\text{subject to } \begin{cases} \sum_{i=1}^{N}(a_i - a_i^*) = 0 \\ 0 \le a_i \le \mathrm{C} \\ 0 \le a_i^* \le \mathrm{C}, \quad i = 1, 2, \ldots, N \end{cases} \tag{6}$$

Following the Karush-Kuhn-Tucker complementarity conditions, *w* and *b* in the SVR function can be calculated (Noori *et al*., 2011). Ultimately, the SVR function can be written in the form shown below:

$$W = \sum_{i=1}^{N}(a_i - a_i^*)\,\varphi(x_i)^T.\varphi(x) + b \quad (7)$$

Finally, the ε-SVR can be expressed as follows:

$$f(x) = \sum_{i=1}^{N}\bar{a}_i\varphi(x_i)^T.\varphi(x) + b \quad (8)$$

where $\bar{a}_i$ is a Lagrange multiplier term $(a_i - a_i^*)$. Since calculation of the kernel function in the feature space (equation (8)) is difficult, the commonly used kernel functions such as Linear, Polynomial, Gaussian, Sigmoid and Radial Basis Function (RBF) are used (Hastie *et al*., 2008; Noori *et al.*, 2011). Among the kernel functions, the RBF has the highest efficiency and has been reported to be the most effective kernel function. The RBF can be written as follows:

$$K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2) \quad (9)$$

where $\gamma > 0$ controls the amplitude of the Gaussian function, and determines the generalization ability of the SVM model. According to the SVM theory, the SVR model is highly sensitive to the tuning parameters (i.e. *C*, ε and γ). Using the LIBSVM toolbox in MATLAB, a K-fold cross-validation (K = 5) was used to optimize the free (tuning) parameters of the SVR.

2.3 BP ANN model

Many theoretical and experimental studies have shown that a single hidden layer is sufficient for ANNs to approximate any complex non-linear function (Fausett, 1993; Haykin, 1999). A BP ANN model has three layers of nodes: the input layer, hidden layer, and output layer. Twenty nodes in the input layer (reflectance in bands 1, 2, 3, 4, 5, 7, 17, 18, 19 and 26, brightness temperature in bands 20, 22, 25, 29, 31 and 32, along with relative humidity, air temperature, wind speed and direction) and one node in the output layer (atmospheric visibility) were used.

The feed forward ANN was used, with a single hidden layer. It is trained with the BP algorithm, and takes advantage of sigmoid and identical transfer functions, respectively, in its hidden and output layers; 21 different networks were explored with different numbers of nodes (5 to 25) in the hidden layer (called "hidden nodes"), to find the optimum number of hidden nodes.

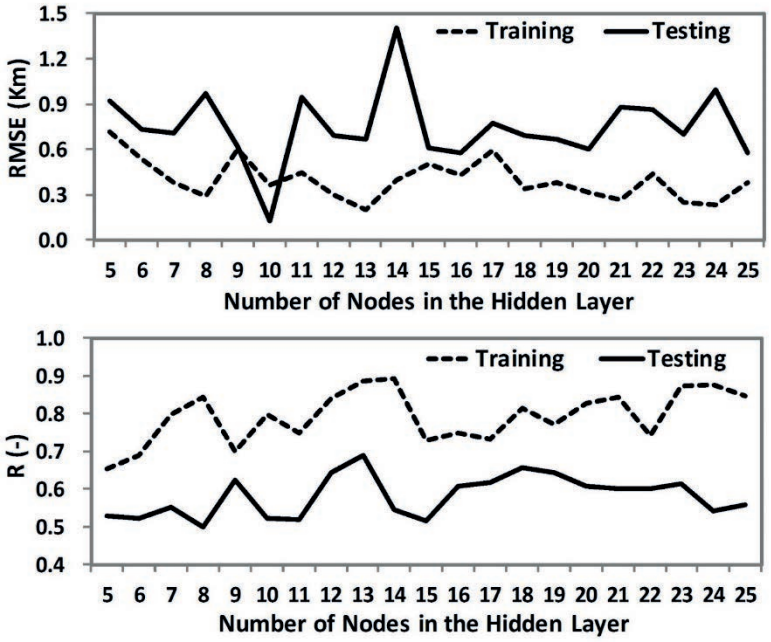

Fig. 1 The RMSE and correlation coefficient (R) of atmospheric visibility estimates for different numbers of nodes in the hidden layer.

3 RESULTS

Before training the BP ANN model, the data were normalized to vary between –1 and 1. The data were divided into training and testing datasets. Of the data points, 65% were used to train the model, and 35% were used to test the model.

Using the training dataset, the ε-SVR model with the RBF kernel was optimized through K-fold cross-validation. The tuned parameters (C, γ , ε) are 10, 0.10, and 0.22, respectively. The BP ANN method was tested using different numbers of hidden nodes. Figure 1 shows the R and RMSE of the atmospheric visibility estimates with different numbers of hidden nodes. A higher R and a lower RMSE indicate more accurate results. Among the 21 examined networks, the network with 13 hidden nodes (with RMSE = 0.67, and R = 0.69 in the testing step, and RMSE = 0.19, and R = 0.89 in the training step) showed the best performance. Thus, a network of 13 hidden nodes was used for the BP ANN method.

Table 1 Statistical indices of atmospheric visibility estimates from the BP ANN and ε-SVR models for both training and testing steps.

	Training		Testing	
	RMSE (km)	R	RMSE (km)	R
BP ANN	0.19	0.89	0.67	0.69
ε-SVR	0.15	0.93	0.59	0.71

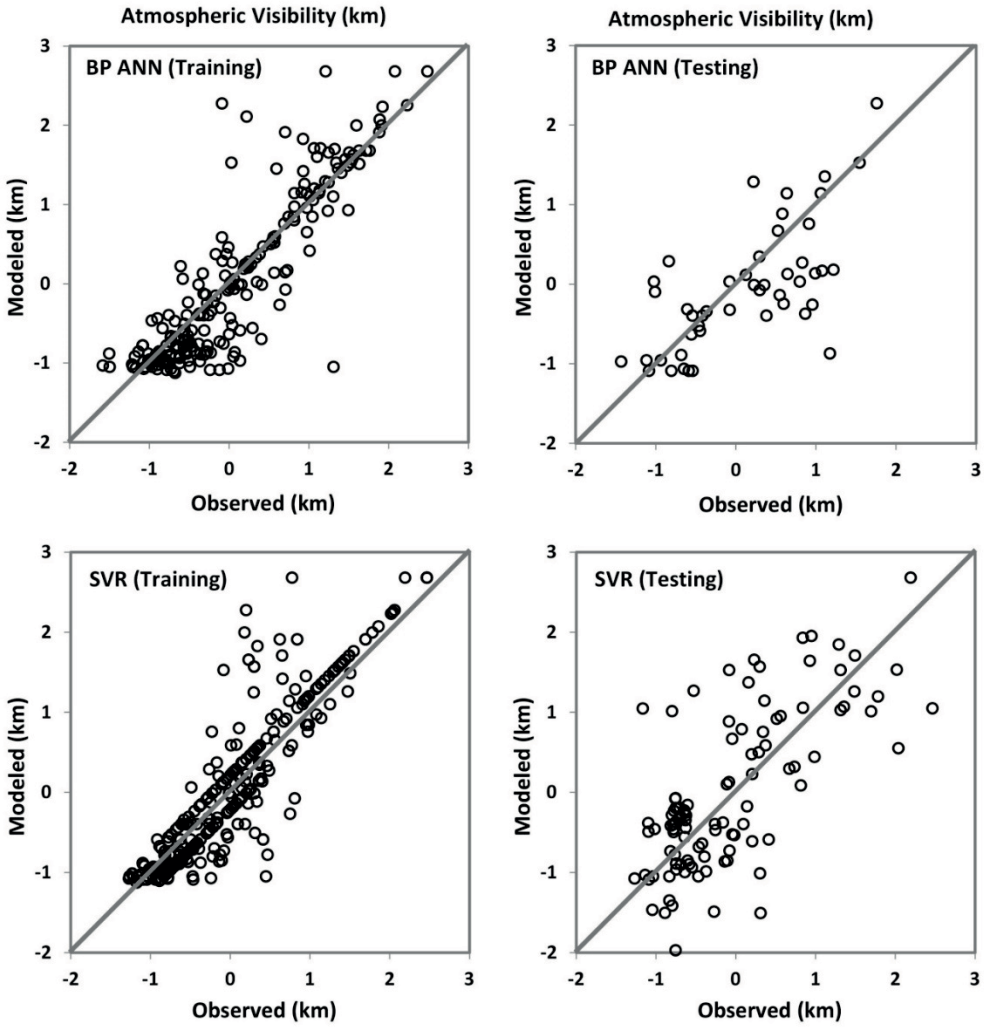

Fig. 2 Scatter plots of atmospheric visibility estimates from BP ANN (top) and ε-SVR (bottom) models *versus in situ* observations for training and testing stages.

Table 1 shows the atmospheric visibility estimates from the BP ANN and ε-SVR models. As indicated, the training results from the BP ANN and ε-SVR models have lower RMSE values and higher R values than the testing results. Both the BP ANN and ε-SVR models produced satisfactory results, which indicate that the training process was effective and reliable. The RMSE value of the ε-SVR model was lower than that of the BP ANN model, and R value of the ε-SVR model was higher than that of the BP ANN model, which means the ε-SVR model outperformed the BP ANN model.

Figure 2 shows scatter plots of the atmospheric visibility estimates from the BP ANN and ε-SVR models *versus* observations. As indicated, results from both models mainly fall around the 45 degree line for both training and testing stages. However, the estimates from testing of both models are more dispersed.

4 CONCLUSION

In this study, two well-known data processing approaches namely, SVR and BP ANN were used to estimate atmospheric visibility, using the synergy of MODIS data and ground-based observations. Both approaches yielded satisfactory estimates of atmospheric visibility. The visibility estimates from the BP ANN approach were found to have an RMSE and R of 0.67 and 0.69, respectively. The corresponding RMSE and R from the SVR model were 0.59 and 0.71, implying that SVR outperformed the BP ANN.

Acknowledgements This project was funded by the High-Tech Research and Development Program of China (No. 2013AA12A301-5), the National Natural Science Foundation of China (41201330 and 41101331), the Water Resource Research Center (WRRC) at the University of Hawaii at Manoa, the Fundamental Research Funds for the Central Universities (2012LYB37), and Specialized Research Fund for the Doctoral Program of Higher Education (20120003120017).

REFERENCES

Baddock, M., Bullard, J. E. and Bryant, R. G. (2009), Dust source identification using MODIS: A comparison of techniques applied to the Lake Eyre Basin, Australia. *Remote Sensing of Environment* 113(7), 1511–1528.

Esmaili, O., Tajrishy, M. and Arasteh, P. D. (2006), Evaluation of dust sources in Iran through remote sensing and synoptical analysis. *AECRIS 2006 Proceedings*, UK, 136–143.

Hansell, R.A. *et al.* (2007) Simultaneous detection/separation of mineral dust and cirrus clouds using MODIS thermal infrared window data. *Geophysical Research Letters*, 34 (13), L11808, doi:10.1029/2007GL029388.

Hastie, T., Tibshirani, R. and Friedman, J. (2008) The elements of statistical learning: data mining, inference, and prediction. *The Mathematical Intelligencer* 27 (2), 83–85.

Karimi, K, *et al.* (2011) A new false color composite technique for dust enhancement and point source determination in Middle East. *Proc. SPIE* 8177, *Remote Sensing of Clouds and the Atmosphere* XVI, 81770X, doi:10.1117/12.897018.

Komeilian, H., Ganjidoust, H. and Khodadadi, A. (2014) Parametric analysis for dust plumes modeling using MODIS data over Khuzestan Province, Iran. *Journal of Middle East Applied Science and Technology* 20, 704–708.

Mei, D., *et al.* (2008) A dust-storm process dynamic monitoring with multi-temporal MODIS data. The International Archives of the Photogrammetry. *Remote Sensing and Spatial Information Sciences* 27, Part B7, Bejing 2008.

Noori, R., *et al.* (2011) Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. *Journal of Hydrology* 401, 177–189.

Schepanski, K., Tegen, I. and Macke, A. (2012) Comparison of satellite based observations of Saharan dust source areas. *Remote Sensing of Environment*, 123, 90–97.

Shao, Y., *et al.* (2011) Dust cycle: An emerging core theme in Earth system science. *Aeolian Research* 2, 181–204.

Shao, Y. and Dong, C.H. (2006), a review on East Asian dust storm climate, modelling and monitoring. *Global and Planetary Change*, 52, 1–22

Shawe-Taylor, J. (2004) *Kernel Methods for Pattern Analysis*, Cambridge UP, UK.

Li, C., Hsu, N. C. and Tsay, S. C. (2011) A Study on the potential applications of satellite data in air quality monitoring and forecasting. *Atmospheric Environment* 45, 3663–3675.

Ma, Y. and Gong, W. (2012) Evaluating the performance of SVM in dust aerosol discrimination and testing its ability in an extended area. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5(6), 1849 – 1858.

Shahrisvand, M. and Akhoondzadeh, A. (2013) A Comparison of empirical and intelligent methods for dust detection using MODIS satellite data. *Remote Sensing and Spatial Information Sciences* XL-1/W3, October 2013, Tehran, Iran.

Vapnik, V. (1979) *Estimation of dependences based on empirical data*. Nauka.

Wu, Y., *et al.* (2012) Synergy of satellite and ground based observations in estimation of particulate matter in eastern China, *Science of the Total Environment*, 433, 20–30.

Zhao, T., Ackerman, S. and Guo, W. (2010) Dust and smoke detection for multi-channel imagers, *Remote Sensing* 2, 2347–2368.