Forecast model of airport haze visibility and meteorological factors based on SVR-RBF model

Dazheng Liu 1,a, Tao Jiang 2,b, Yangchun Zhang 1,c, Yingjie Wang 2,d, Xiaotian Pan 1,e, Jianping Wu 1,f*

1College of Meteorology and Oceanography, National University of Defense Technology, Changsha, Hunan, China
2College of Computer Science and Technology, National University of Defense Technology, Changsha, Hunan, China

aemail:liudz@nudt.edu.cn,bemail:jiangtt18@126.com,cemail:zyc19950306@sina.com
demail:wangyingjie20@nudt.edu.cn,eemail: pxt20@nudt.edu.cn

*Correspondence Author: femail:wjp@nudt.edu.cn

Abstract: All kinds of meteorological elements near the airport directly affect the dissipation process of fog, which also affects the visibility. The correlation analysis of temperature, humidity, dew point temperature, air pressure, wind speed and other meteorological factors with visibility Runway Visual Range (RVR) and Meteorological Optical Visual Range (MOR) was carried out by SPSS. It was found that the four variables, TEMP, RH, WS2A and PAINS, had great influence. In this paper, the above four variables are selected as inputs to nonlinear fit the RVR value of visibility. Combined with Radial Basis Function (RBF), it is found that SVM-RBF model has better generalization ability and fitting degree.

1. Introduction

With the development of industrialization, more and more pollutants are emitted, and the frequency of haze is more frequent. Haze not only seriously affects people's health and quality of life, but also brings huge problems to the transportation industry, especially to the safety of flights. With the popularity of airports and the increasing number of flights taken by people, the safety of navigation is particularly important. Therefore, it is of great significance for the development of aviation industry and people's travel safety to accurately predict the laws of fog dispersal and the changing trend of visibility at airports.

Since the 1950s, people at home and abroad have begun to pay attention to the prediction methods of visibility. From the initial linear regression to nonlinear regression, the prediction effect has been greatly improved, and the prediction effect of neural grid model has been improved a lot. However, most of the previous studies selected PM2.5, PM10 and relative humidity which are directly related to visibility as input variables. The measurement of PM2.5 and PM10 has a large error. Therefore, the prediction effect of the obtained model for other input data may not be very ideal. Therefore, it is necessary to screen a variety of meteorological elements to find out the ones that have great impact on visibility. How to select the relevant variables and establish a simulation prediction model of airport visibility is the focus of this paper.

As we all know, haze is related to near surface meteorological factors such as temperature, pressure, humidity and wind, and haze determines the magnitude of theoretical visibility, so visibility is closely
related to near surface meteorological factors. In this paper, the AMOS data of Changsha Huanghua Airport on March 13 and 14, 2020 are used. Ten kinds of ground observation data and two kinds of visibility are shown in Table 1.

| Major categories | Subclass | Meaning | Remarks |
|------------------|----------|---------|---------|
| pressure         | PAINS    | air pressure of the station QFE aircraft landing area QNH maximum pressure QNH corrected sea level pressure |
| temperature      | TEMP     | temperature DEWPOINT dew point |
| humidity         | RH       | relative humidity RVR average RVR value per minute ≤2000 |
| visibility       | MOR      | average MOR value per minute ≤10000 |
| brightness       | LIGHT    | Light data WS2A average wind speed in 2 minutes WD2A average wind direction in 2 minutes CW2A average vertical wind speed in 2 minutes |
| windspeed        | CREATEDATE | create time for observation record, universal time LOCALDATE is the corresponding Beijing time |
| site             | SITE     | is the name of the observation point |

2. Modeling

2.1 Data processing and analysis
The direct cause of visibility change is the change of atmospheric aerosols, and meteorological conditions play a significant role in the formation, distribution, maintenance and change of aerosols, especially the formation and change of secondary aerosols are greatly affected by meteorological conditions. It is difficult to determine the size of each type of aerosol under different weather conditions due to the influence of various meteorological factors on the development of visibility. Therefore, we
first use the linear regression method commonly used in mathematics to find the influence model of different meteorological elements on visibility.[1]

First of all, in the two-day data, the meteorological elements related to wind and visibility are four groups per minute, and the other data are one group per minute. Therefore, the amount of these twelve kinds of data is different. It is necessary to delete redundant data to normalize the quantity. The initial time of each minute is used as the screening standard, and the meteorological elements containing four groups of data in one minute are reserved as one of the initial time. In order to ensure the synchronization and equivalence of all meteorological data, the two-day data are integrated together for more flexible processing and fitting. There are many meteorological factors that affect visibility, which need to be screened. We use the correlation analysis in SPSS to select the main influencing factors, and fit the correlation between various meteorological elements and visibility as follows:

| Table 2 | Correlation between visibility and meteorological elements |
|---------|-----------------------------------------------------------|
| RVR    | MOR  | WD2A  | CW2A  | LIGHTS | PAINS | QFE  | TEMP | RH (%) | DEWPOINT | WS2A | QNH |
| RVR    | 1    | 0.767 | -0.351 | 0.563  | 0.001  | 0.182 | -0.015 | 0.668   | -0.6     | -0.044 | 0.536 |
|        |      |       |        |        |        |       |       |        |          |       |     |
|        |      |       |        |        |        |       |       |        |          |       |     |
| MOR    | 0.767 | 1     | -0.073  | 0.617  | 0.222  | 0.159 | -0.055 | 0.866   | -0.485   | 0.628  | -    |
|        |       |       |        |        |        |       |       |        |          |       |     |
|        |       |       |        |        |        |       |       |        |          |       |     |

It can be seen from Table 2 that the correlation between RVR and MOR is 0.767, and the correlation between each variable and the two has the same characteristics, which is easy to simplify the problem. RVR is used as the judgment standard of visibility. There were significant correlations between RVR and TEMP (0.668), Rh (-0.600), CW2A (0.563), WS2A (0.563) and WD2A (-0.351), and the significance of these variables was less than 0.05. LIGHTS (0.953), QFE (0.432), QHN (0.450) were more significant than 0.05, which were not suitable for input. Although the significance of DEWPOINT was less than 0.05, the correlation (-0.044) was too small to be used as an input variable. The correlation of PAINS was 0.182. Although it was small, it was only a variable that could be screened in the barometric elements, so it was also used as an input variable. From Figure 1, it can be seen that WD2A oscillates violently and is not suitable as input variable. The correlation between WS2A and CW2A, which are both wind speed variables, is 0.887. It can be seen from figure 3 that the change trend of the two is basically consistent. Considering that the influence of average wind speed is more obvious, WS2A is selected as the input variable. Considering the above factors, TEMP, RH, WS2A and PAINS were selected as input variables and RVR as output variables. Only data with RVR < 3000, that is, low visibility, were considered in model selection.
2.2 Model selection
First of all, the linear regression model is considered, with four variables of TEMP, RH, WS2A and PAINS as the input and RVR as the output. The accuracy of SPSS linear regression model is only 55.6%, which indicates that the linear model is not suitable for prediction because of its poor fitting degree. Therefore, the nonlinear relationship between the variables and RVR must be considered.

Support Vector Machine (SVM) is a kind of generalized linear classifier which classifies data by supervised learning. Its decision boundary is the maximum margin hyperplane to solve the learning samples.[2]

SVM can be extended to regression problems to obtain Support Vector Regression (SVR), and the kernel function is introduced to linearize the nonlinear variables to obtain the regression results.
Among them:

\[ f(x) = \sum_{i=1}^{M} \alpha_i h_i k(x, y) + b \]  

(1)

Among them:

- \( M \): The number of support vector machines
- \( \alpha_i \): Lagrange coefficient of the \( i \)th support vector
- \( h_i \): class identification of the \( i \)th support vector
- \( k(x, y) \): kernel function

For kernel function, Radial Basis Function (RBF) kernel, also called Gaussian kernel function, is chosen. Gaussian kernel function is to get new samples by changing the characteristic data of samples according to certain rules, and the new samples can be classified better according to the new characteristic data. Because the characteristic data of new samples and the characteristic data of original samples have certain regular corresponding relationship, according to the distribution and classification of new samples, the classification of original samples is obtained. The expression of Gaussian kernel function is as follows:

\[ k(x, y) = \exp(-\gamma \| x - y \|^2) \]  

(2)

Among them:

- \( x, y \): for sample or vector
- \( \gamma \): is a super parameter
- \( \| x - y \| \): the norm of \( x - y \), which can be understood as a module when the dimensionality is low

The \( k(x, y) \) function represents the relationship between \( x \) and \( y \) and returns a value. The Gaussian kernel function first maps the original data point \( (x, y) \) to a new sample \( (x', y') \); then the new feature vector dot product \( (x' \cdot y') \), return the result of the dot product; first map the \( m \times n \) data set to the \( m \times m \) data set, \( m \) represents the number of samples, \( n \) represents the feature type of the original sample. The number is large, therefore, the samples of the new data set obtained are also high-dimensional, and the process of ascending makes the linearly inseparable meteorological data linearly separable. Finally, the relationship between RVR and TEMP, RH, WS2A and PAINS is the SVR-RBF model:\n
\[ f(x) = \sum_{i=1}^{M} \alpha_i h_i \exp(-\gamma \| x - y \|^2) + b \]  

(3)

Among them:

- \( f(x) \): here is the visibility RVR
- \( x \): the four-dimensional vector of (TEMP, RH, WS2A, PAINS)
- \( y \): denotes the correlation vector with inner product 1 of \( X \)
- \( b \): constant
M: there are 754 support vector machines and 754 corresponding $\alpha$ and $h$ groups.

3. Result analysis
Among them, $\gamma$ can be adjusted as the fitting parameter, and the fitting results of different $\gamma$ are shown in the figure below.

![Figure 3 different fitting results of $\gamma$ values from 0.005 to 10](image)

The neural network model is selected as the control test. The selected parameters are shown in Table 4-3, and the fitting diagram is shown in Figure 4.

| Table 3 Neural network model parameters |
|-----------------------------------------|
| Parameters  | Value |
| Input       | 1     |
| Units_1     | 10    |
| Units_2     | 10    |
| Units_3     | 10    |
| Output      | 1     |
| Hidden      | 3     |
| layers      |       |
It can be seen from Figure 3 that with the increase of $\gamma$ value, the fitting effect is closer to the actual value. When $\gamma$ is 10, the fitting result is very close to the actual value. It can be seen from Figure 4 that the simulation effect of the neural network is not very good and cannot be used as the calculation model. Therefore, we take $\gamma = 10$ as the coefficient of the final SVR-RBF model, and make further analysis.

The accuracy of SVR-RBF model is 96% after linear fitting, and the data correlation between them is 99.7%. It shows that the relationship between the predicted visibility and the selected meteorological elements is relatively accurate, but more than 700 parameters are selected, which results in a large amount of calculation and long calculation time. The fitting degree between the calculated data and the original data is shown in Table 4.

| V1 (Original value) | V1 (Calculated value) |
|---------------------|-----------------------|
| Pearson correlation | 1                     | 0.997                 |
| Significance        | 0.000                 |                       |
| Number of samples   | 777                   | 777                   |
| Pearson correlation | 0.997                 | 1                     |
| Significance        | 0.000                 |                       |
| Number of samples   | 777                   | 777                   |

4. Conclusion

In this paper, based on the two-day AMOS data of Changsha Huanghua Airport, the fitting prediction model between RVR and various meteorological elements is found through linear analysis and nonlinear analysis. Through linear analysis, TEMP, RH, WS2A and PAINS which are closely related to RVR are selected as input variables and RVR as output variables. Through the linear regression analysis, the fitting degree of the linear model is not high, and the accuracy rate is only 55.6%. In the nonlinear model using support vector machine, Gaussian kernel function (RBF), that is, SVR-RBF model, is selected as kernel function. It is found that when the parameter $\gamma = 10$, the accuracy rate of prediction results and real value reaches 96%, and the data correlation between them reaches 99.7%, indicating that the prediction model has a high degree of fitting. Compared with the hidden three-layer neural network, it is found that the prediction effect is not obvious. The prediction accuracy of RVR-RBF model is high, but the related calculation parameters are as high as 754 groups, which is the place to be solved and improved in the future.
Acknowledgments

This paper is one of the stage achievements of the National Natural Science Foundation of China's "Highly Scalable Parallel Computing Technology Research on global numerical weather forecasting spectrum model" (41875121).

References

[1] Jacobson M C, Hansson H C, Noone K J, et al. Organic atmospheric aerosols: Review and state of the science[J]. Reviews of Geophysics, 2000, 38(2): 267-294.
[2] Vapnik, V. . Statistical learning theory. 1998 (Vol. 3). . New York, NY: Wiley, 1998: Chapter 10-11, pp.401-492.
[3] Karatzoglou A, Smola A, Hornik K, et al. kernlab-an S4 package for kernel methods in R[J]. Journal of statistical software, 2004, 11(9): 1-20.
[4] Sarah Komasová. Airport Security and Visibility: Security as Visualization and Its In-Place Performance[J]. Journal of Applied Security Research, 2020, 15(3).
[5] Mateusz Taszarek, Sebastian Kendzierski, Natalia Pilguj. Hazardous weather affecting European airports: Climatological estimates of situations with limited visibility, thunderstorm, low-level wind shear and snowfall from ERA5[J]. Weather and Climate Extremes, 2020, 28.
[6] Environmental Health - Air Quality; Investigators at Nanyang Technological University Report Findings in Air Quality (Impact of Fine Particulate Matter On Visibility At Incheon International Airport, South Korea)[J]. Ecology, Environment & Conservation, 2020.
[7] S. K. Nayar, S. G. Narasimhan, Vision in bad weather, ICCV’99.
[8] R. T. Tan, Visibility in bad weather from a single image, CVPR, 2008.
[9] N. Hautiere, J-P Tarel, J Lavenant D. Aubert, Automatic fog detection , and estimation of visibility distance through use of onboard camera, Machine Vision and Application, 2006, 17 (1): 8-20.
[10] C. Sakaridis, D. Dai, L. V. Gool, Semantic foggy scene understanding.