ISSN: 2664-9640

Self-checking of Aviation Meteorological Observation Messages Based on Multiple Linear Regression Method

Ruiting Xu[1], Jinxu Cai[1], Yajing Xu[1], Jingcheng Zeng[2], Hongda Tai[1]

[1] Department of Air Traffic Management, Civil Aviation University of China, Tianjing, China, 300300, China

[2]Department of Computer Science and Technology, Civil Aviation University of China, Tianjing, China, 300300, China

Abstract

The quality of meteorological observation reports is one of the most important factors to ensure the safety of aviation activities. Along with the development of civil aviation in China, it is of practical importance to control the quality of meteorological observation reports at a better level. In this paper, we use multiple linear regression method and SPSS and Excel to study and analyze the Meteorological Terminal Aviation Routine Weather Report (METAR) of about 100 airports in eight regions of China and Tianjin Binhai International Airport in autumn 2021 since September 2021. The Meteorological Terminal Aviation Routine Weather Report (METAR report) is used to analyze the wind speed, temperature, dew point, modified sea pressure, visibility and cloudiness data from eight airports in China and Tianjin Binhai International Airport in the fall of 2021. The reporting system automatically compares the manually reported reports with the calculated model values to judge the quality of meteorological observation reports, and achieves the self-checking function, which greatly reduces human errors when sending reports. The results show that the multiple linear regression method can be better applied to the quality inspection of meteorological data related to METAR reports; the inspection model built for the local area using local meteorological data is more applicable to the quality inspection of meteorological data in the local area than the regression model built using national meteorological data.

Keywords

METAR message; Self-check; Quality inspection of meteorological reports; Multiple linear regression.

1. Introduction

With the continuous development of China's civil aviation industry, airplanes have become one of the common ways for people to travel on a daily basis, and the flight safety of aircrafts has become especially important. There are many factors that affect aircraft flight safety, but meteorological factors are the most common and important factors that affect aircraft flight. As an important part of aviation meteorological information, timely and accurate release of meteorological observation reports is of great significance for aircraft take-off and landing [1].

The Quality Management Measures for Meteorological Work of Civil Aviation Air Control System has made strict regulations on the quality of meteorological observation reports, and many scholars in China have conducted a lot of research on such issues: Yang Fan studied the causes of meteorological report compilation and transmission errors [2], and concluded that there are two main forms of errors, compilation errors and transmission errors, and most of them belong to human errors; Hu Ye Xiao further pushed the research to the quality of reports [3]. The analysis of conflicts in the structure of civil aviation meteorological observation reports

was completed in accordance with the relevant regulations, and the analysis of the report structure, conflicts within and between report items was summarized, which is a good reference for the timely correction of report errors; Qiu Hui et al. compared and examined conventional meteorological observations with automatic station data [4], and monitored the data for minute data jumps, minute data missing measurements, and minute data abnormalities to reduce the station Zhu Guodong et al. focus on the control of message quality [5], and develop and design the quality control algorithm for civil aviation meteorological messages to realize simple logical judgments on ground average wind speed and gusts, dominant visibility and weather phenomena, distribution of different cloud groups, and combinations of maximum and minimum temperatures, which further improve the message quality; Zhu Qing et al. studied the monitoring system of meteorological observation messages for civil aviation [6], which effectively reduces the problems of human "errors, forgetting, and omissions" in the distribution of observation messages and the omission of messages due to equipment failure. Even though some scholars have developed and designed algorithms for the quality control of meteorological reports in civil aviation, the judgments of meteorological data values are limited to simple logical judgments, which lack statistical significance and cannot be used simultaneously. It is not possible to conduct more accurate quality checks based on multiple meteorological data at the same time.

In this paper, we propose a method to automatically compare the calculated values of models and manually published reports by the sending system to check whether there are any anomalies that deviate from the predicted values, and then judge the quality of meteorological observation reports, which greatly reduces human errors in sending reports and makes the quality control of meteorological prediction reports better, and effectively improves the quality of meteorological work and meteorological services. In this paper, we compare and analyze the test models obtained from different data calculations, and study the test capability of a single forecast model and the adaptation effect of different forecast models in different regions. The findings of the study are of great significance to the airport meteorological forecasters in terms of reducing forecast errors and teaching and training in compilation training.

2. Information and Methods

In this paper, we use the Hong Kong Underground Astronomy (https: //www.weather.org.hk/) collection of about 100 airports in eight regions of China since September 2021 and the NOAA download (https://www.ncei.noaa.gov/maps/hourly/) of Tianjin Binhai International Airport for the fall of 2021 from September to We use statistical methods to correlate and multivariate linearly fit the six meteorological data of wind speed, temperature, dew point, modified sea pressure, visibility and cloudiness in the routine observation reports (METAR reports) to calculate a meteorological report quality check model applicable to the whole country and Tianjin region.

In many practical problems in the field of atmospheric meteorological science, there are often multiple independent variables affecting the dependent variable, so regression analysis between a dependent variable and multiple independent variables, i.e., multiple regression analysis, is required. In this paper, we adopt a multiple linear regression approach to set up and build a model, first establishing the explanatory variable, i.e., the dependent variable, and then using correlation analysis to identify and eliminate the explanatory variables, At the significance level of 0.01 or 0.05, if the *P*-value of the correlation coefficient is less than 0.01 or 0.05, the explanatory variables are significantly correlated with the explanatory variables and are introduced into the regression equation; conversely, the corresponding explanatory variables are excluded. The regression equation is established by taking the factor that is significantly correlated with the explanatory variable and has the largest contribution.

The general form of the multiple linear regression model.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \tag{1}$$

Multicollinearity among sample factors reduces the predictive power of the model [7], and the severity of multicollinearity in the multiple linear regression model is tested using the variance inflation factor *VIF* [8] , when *VIF* < *5*, it means that there is no multicollinearity among the factors, and when *VIF* > *5*, there is strong covariance among the factors and should be removed from the model [9]. The *D-W* value is used to test whether there is autocorrelation in the model samples, and the estimation error of the model coefficients is large when autocorrelation exists, and the reliability and accuracy of the model are reduced. The *D-W* test result is considered to be near 2, which means that the basic assumption of no autocorrelation between the linear regression samples is satisfied and the regression model is valid [10]. The regression effect of the model is usually evaluated by the coefficient of determination R^2, which indicates the proportion of the total variance of the dependent variable y that can be explained by the independent variable, and the closer the value of R^2 is to 1, the better the fit of the regression equation to the actual observations, and conversely the closer R^2 is to 0, the worse the fit is [11].

3. Correlation Analysis of Meteorological Elements

Correlation analysis is used to reflect the direction and degree of correlation between elements, and the correlation coefficient is a statistical measure of the relationship between any two elements [12]. The Pearson correlation coefficient method was used to analyze the correlation and its significance level of six meteorological elements: wind speed, temperature, dew point, modified sea pressure, visibility, and cloudiness in the national METAR messages.

For the correlation coefficient between two variables the following formula is calculated.

$$\rho_{XT} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \tag{2}$$

A total of 423 national reports and 4367 Tianjin reports were collected, and the six meteorological elements of wind speed, temperature, dew point, modified sea pressure, visibility and cloudiness were selected as the factors of this data quality test regression model by combining the elements of routine observations in METAR reports. The Pearson correlation coefficient method was used to analyze the correlation and significance levels of the elements in the national METAR reports. The correlation is significant when the significance level sig (two-tailed) of two variables is less than 0.01 or 0.05, and a positive Pearson correlation coefficient indicates a positive correlation between variables, and a negative Pearson correlation coefficient indicates a negative correlation [13].

The results of correlation analysis (Table 1) show that: temperature as the object of study, dew point and temperature are positively correlated, modified sea pressure, wind speed, cloudiness and temperature are negatively correlated, where wind speed, dew point and modified sea pressure are significantly correlated at 0.01 level (two-tailed). For dew point, temperature and dew point are positively correlated, and modified sea pressure and dew point are negatively correlated, and temperature and modified sea pressure are significantly correlated at the 0.01 level (two-tailed). For the modified SST, wind speed, visibility and modified SST are positively correlated, and temperature, dew point and modified SST are negatively correlated, with temperature and dew point correlating significantly at the 0.01 level (two-tailed). The correlation conclusions between different elements are consistent and also indicate the reliability of the regression model to some extent [14].

ISSN: 2664-9640

Table 1. Correlation analysis

		Wind Speed	Temperature	Dew point	Correction of sea pressure	Visibility	Cloud Volume
Wind Speed	Pearson Correlation	1	-.190**	.026	.106*	.127**	.035
	Sig. (two-tailed)		.000	.598	.032	.009	.614
Temperature	Pearson Correlation	-.190**	1	.246**	-.704**	-.060	-.152*
	Sig. (two-tailed)	.000		.000	.000	.229	.032
Dew point	Pearson Correlation	.026	.246**	1	-.567**	-.061	.038
	Sig. (two-tailed)	.598	.000		.000	.220	.589
Correction of sea pressure	Pearson Correlation	.106*	-.704**	-.567**	1	.120*	.017
	Sig. (two-tailed)	.032	.000	.000		.016	.807
Visibility	Pearson Correlation	.127**	-.060	-.061	.120*	1	-.112
	Sig. (two-tailed)	.009	.229	.220	.016		.103
Cloud Volume	Pearson Correlation	.035	-.152*	.038	.017	-.112	1
	Sig. (two-tailed)	.614	.032	.589	.807	.103	

**. Significant correlation at the 0.01 level (two-tailed).

*. At the 0.05 level (two-tailed), the correlation is significant.

Considering the default of the original data, wind speed and cloudiness were excluded from the study as dependent variables, and four items of visibility, air pressure, dew point and temperature in the national reports were selected as the four test elements established by the self-testing rule. Combined with the judgment of significance condition sig (two-tailed), the meteorological elements with relatively large Pearson correlation coefficients were selected as the independent variable factors of the regression model, respectively.

4. Model Building and Testing

A multiple linear regression model was established using SPSS24 to determine the regression equation coefficients, and the specific data obtained (including regression coefficients and corresponding *P*-values, coefficient of determination R^2, model *F*-values, and D-W-values) are shown in Table 3 and Table 4, and the regression standardized residuals are plotted on the vertical axis using the standardized residuals ZRESID as the vertical axis and the standardized predicted values ZPRED as the horizontal axis in Figure 1.

Since the methods and steps for constructing multiple linear regression models are the same for the four elements, and the methods for model testing and analysis are the same, to avoid repetitive elaboration and analysis, dew point is used as an example for detailed analysis, and the regression models and parameter results for the remaining three elements are reflected in the graphs(

Table 2).

Dew point is used as the explanatory variable to establish the multiple linear regression equation, TMP and *QNH* represent temperature and modified sea level pressure, respectively, and β is the regression coefficient, thus establishing the linear regression equation as follows.

ISSN: 2664-9640

$$DEW = \beta_1 * TMP + \beta_2 * QNH \quad (3)$$

The regression coefficients as well as the results of the significance analysis (Table 3) show that the p-values (Sig) of temperature and corrected sea pressure are less than 0.05 and pass the significance test. The Pearson correlation coefficient between dew point and temperature is 0.246, which indicates that dew point is significantly correlated with temperature at the 0.01 level. The Pearson correlation coefficient between dew point and modified sea pressure is -0.567, which indicates that dew point is significantly correlated with modified sea pressure at the 0.01 level. Through the model summary, the model fit (Table 4) illustrates that the coefficient of determination R^2 = 0.369, indicating that the independent variables modified sea pressure and temperature can explain 36.9% of the variation in dew point, and the Durbin-Watson value is 1.931, which can be considered as independent and reliable between samples, satisfying the basic assumptions, and the model F value is 1.064 significance level is less than 0.05, the model passes the test, and the independent factor has a significant effect on the dependent variable.

Table 3 gives the values of the coefficients of the regression equations, the results of the statistical significance tests of the coefficients and the variance inflation factor VIF. because the corresponding p-values of the coefficients in the estimated directions are less than 0.01, and the VIFs are less than 5, indicating that there is no multicollinearity between the respective variables, which verifies the rationality of the independent variable factors. Under the condition of satisfying the significance test, the meteorological elements with relatively large Pearson correlation coefficients were selected as independent variables, and the effects of temperature and modified sea pressure on dew point were negatively correlated, with increasing temperature and decreasing dew point under the condition of certain modified sea pressure, and with increasing modified sea pressure and decreasing dew point under the condition of constant other variables. Therefore, the multiple linear regression model can be obtained as.

$$DEW = 678.943 - 0.178 * TMP - 0.647 * QNH \quad (4)$$

The plot of standardized residuals in Figure 1 (Fig. 1) shows that the regression standardized residuals fluctuate randomly around 0 within a certain range of variation with no significant positive or negative bias, indicating that the residuals of this regression model obey a normal distribution, the model conforms to the basic assumptions, and the model regression is relatively ideal.

Table 2. Regression coefficient significance analysis

Models		Unstandardized coefficient		Standardization factor	Significance	Covariance statistics	
		B	Standard Error	Beta		Tolerances	VIF
TMP	(Constant)	1194.303	59.967		.000		
	QNH	-1.151	.059	-.815	.000	.663	1.507
	WID		.169	-.096	.005	.977	1.023
	DEW	-.365	.070	-.214	.000	.670	1.492
DEW	(Constant)	678.943	47.044		.000		
	QNH	-.647	.046	-.781	.000	.505	1.982
	TMP	-.178	.033	-.304	.000	.505	1.982
VIS	(Constant)	-24400.048	15318.960		.112		
	QNH	33.070	15.198	.107	.030	.989	1.011
	WID	128.297	53.280	.119	.016	.989	1.011

ISSN: 2664-9640

	(Constant)	1030.319	.847		.000		
QNH	TMP	-.425	.021	-.601	.000	.940	1.064
	DEW	-.507	.036	-.420	.000	.940	1.064

Table 3. Model summary

Models	R	R2	After adjustment R2	Durbin Watson	F
DEW	.607a	.369	.365	1.931	117.894
VIS	.168a	.028	.023	1.920	5.884
TMP	.734a	.539	.536	2.123	157.219
QNH	.813a	.661	.659	1.786	393.655

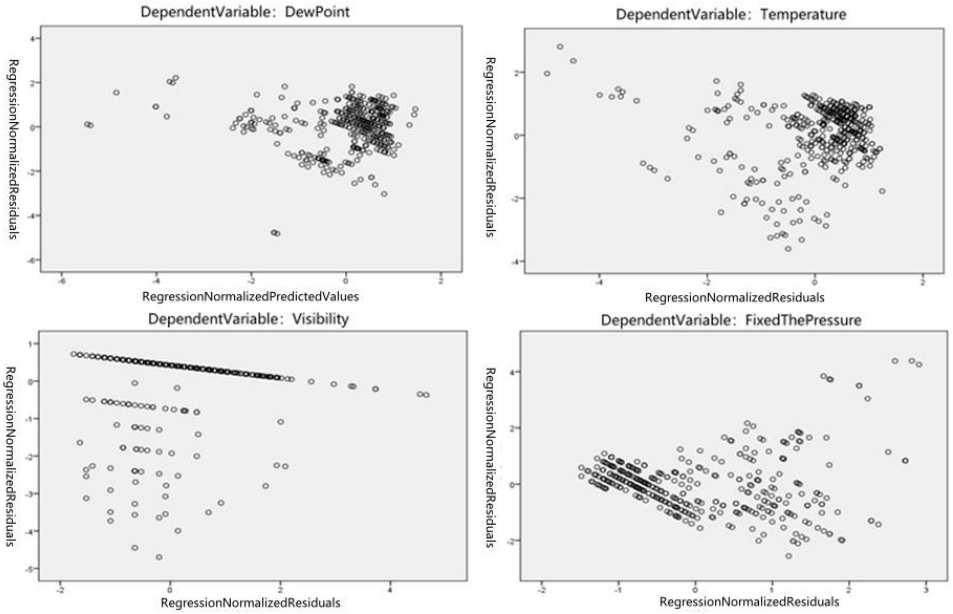

Figure 1. Standardized residual plot

The same method was used to establish multiple linear regression equations for the remaining three elements separately and passed the test with the following results.

$$QHN = 1030.319 - 0.425 * TMP - 0.507 * DEW \quad (5)$$

$$VIS = -24455.229 + 125.985 * WND + 33.132 * QNH \quad (6)$$

$$TMP = 1194.303 - 1.151 * QNH - 0.477 * WND - 0.365 * DEW \quad (7)$$

5. Discussion

The results are summarized in Table 5. All the equations fit the actual values well, and the results pass the significance test, reflecting the good linear correlation. Figure 2 shows the comparison between the actual and calculated values of dew point, modified sea pressure, visibility and temperature in the national sample in 2021 based on the multiple linear regression equation, which indicates that the calculated values of the multiple linear regression model correlate well with the actual values and can be used for data quality testing.

Table 4. Summary of multiple linear regression equations

Regression equation

QNH	Y=1030.319-0.425x1-0.507x2
DEW	Y=678.943-0.178x1-0.647x3
TMP	Y=1194.303-0.365x2-1.151x3-0.477x4
VIS	Y=-24455.229+33.132x3+125.985x4

Note: Y is the dependent variable, and according to the multiple linear regression model, the relationship equation between Y and the four influencing factors temperature x_1, dew point x_2, modified sea pressure x_3, and wind speed x_4 can be obtained.

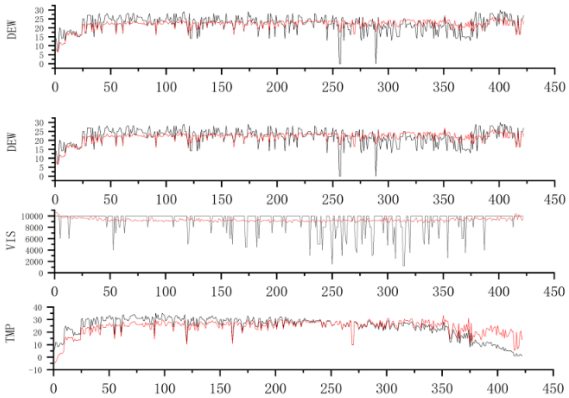

Figure 2. Actual values versus model calculated values

In order to further analyze the testing ability of the model, the relative errors of the calculated values were studied in 10% intervals, as shown in Table 6. The relative errors of 97.87% of the data were within 1%; the errors of visibility models were within 10% in 301 cases, accounting for 71.16% of the total. Overall, most of the test errors are concentrated below 10%, and the test results are reliable, indicating that the self-test model has good testing ability.

Table 5.Model predictions versus actual values relative to error results

Meteorological elements	Relative Error	Percentage of
DEW	＜10%	48.46%
	＜20%	78.96%
QNH	＜1%	94.33%
	＜10%	96.22%
VIS	＜5%	17.49%
	＜10%	71.16%
TMP	＜10%	36.17%
	＜20%	71.39%

To further analyze the testing capability of the model, scatter plots were generated using the actual values of the four test elements as horizontal coordinates and the corresponding relative errors as vertical coordinates (Figure 3) to illustrate that the accuracy of the self-test model varies when different values of the elements to be tested are taken, i.e., the range of application of the test model varies with different accuracy requirements. Combining the data and scatter plot analysis, the model developed in this paper is best tested at TMP between 27°C and 29°C, DEW between 22°C and 24°C, VIS=9999, and QNH around 1005hpa.

ISSN: 2664-9640

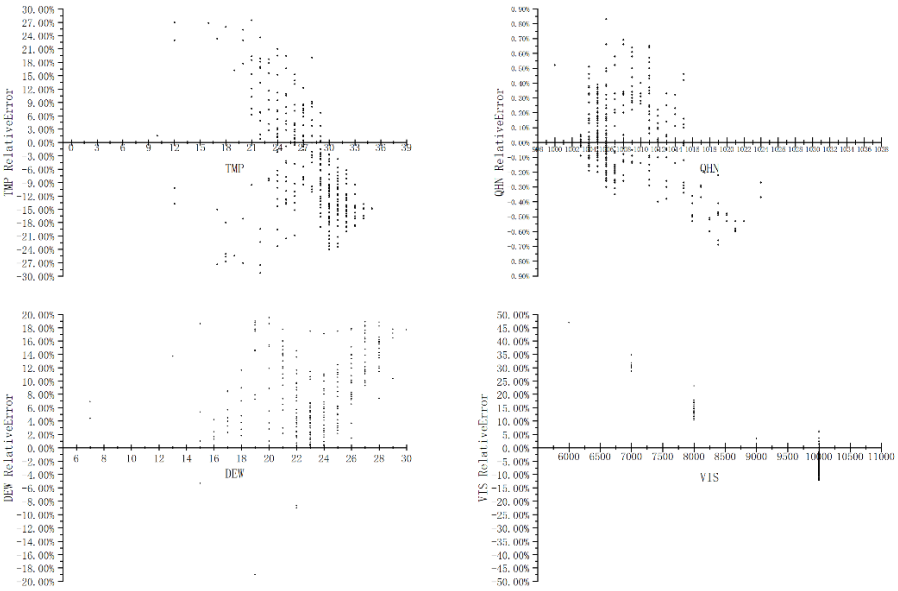

Figure 3. Test Feature Actuals - Relative Error Scatter Plot

In this paper, we use the ground MEATR messages from September to November 2021 at the fixed site Tianjin Binhai International Airport as the dataset used for validation analysis, with a total of 4367 messages, and establish a self-test model (the national sample model is denoted as model 1, and the Tianjin sample model is denoted as model 2) using the method above. If the accuracy of local site data model 2 is better than model 1, it means that the influence of regional differences in establishing the test model should be fully considered, and the message test model should be established for the national site partition based on the geographical location, climate characteristics and other factors to achieve optimization. The values of meteorological elements of MEATR messages from September to November 2021 at Tianjin Binhai International Airport were substituted into model 1 and model 2 for testing, and the results of the error analysis comparing the measured and calculated values were as follows.

Table 6. Prediction error comparison of models 1 and 2

Meteorological elements	Contrast items	Model 1	Model 2
DEW	Error <3℃share	20.2%	49.8%
	Error <6℃share	39.5%	80.9%
TMP	Error <3℃share	29.7%	52%
	Error <6℃share	55.1%	80.1%
VIS	Error <1800m share	57.3%	39.5%
	Error <3600m share	64.3%	80.2%
QNH	Error <4hpa share	51.7%	49.4%
	Error <8hpa share	80.7%	87.3%

The analysis and comparison show that, taking Tianjin Binhai International Airport as an example, the test model based on the national data can achieve some accuracy for the local meteorological data of a single station, but the test model based on the local meteorological data is relatively better.

6. Conclusion

Data quality inspection of METAR messages is of great importance for civil aviation operations. In this paper, the observations of temperature, dew point, corrected sea pressure and visibility in METAR messages from 100 airports nationwide in September 2021 are studied based on statistical correlation analysis, multiple linear regression model, R^2 test, DW test and standardized residual analysis method. The results show that: (i) a series of self-testing rules based on multiple linear regression models can be constructed to better test the quality of METAR data, and the calculation error can be controlled within an acceptable range of less than 10%. (ii)self-test rules exist to test the best data range. (iii)METAR message self-test rules should be taken into account regional climate differences, the national site to discuss the study of zoning.

After modeling and predicting a large amount of data, a feasible method of automatic METAR message inspection rules was derived, and a preliminary set of self-inspection rules was summarized to achieve a better result of relative error within 10% for the four elements of temperature, dew point, corrected sea pressure and visibility inspection. Among them, the test for QNH can reach more than 97% of the data with relative error less than 1%, which is considered to meet the requirements of aviation weather practitioners' ability to compile accurate reports.

However, due to the time constraint of the project and limited original data collection, there are some shortcomings in the design of the test rules, for example, only the single station of Tianjin Binhai International Airport in North China in autumn 2021 METAR messages were applied for the test revision, but the design of the classification to study the effects of sunny and rainy weather changes on the QNH regression model and the effects of optical differences between day and night on the VIS model has not been resolved yet. failed to make the self-test rule with better physical interpretation. In this study, only four meteorological elements were selected for comparative analysis at a single site with limited samples, and further sub-regional studies are needed in later work.

References

[1] Xu Q, Zhang YF. Research on the quality detection system of civil aviation meteorological observation reports[J]. Technology and Innovation,2018(17):42-44.

[2] Yang, Fan. Analysis of the causes of errors in the compilation and distribution of meteorological reports and countermeasures to avoid them [J]. Guangxi Meteorology,2003(01):60-63.

[3] Hu Yecheng. Analysis and application of conflicting structures in civil aviation weather observation reports[J]. China Flight, 2020.

[4] Qiu Hui, Zhan Hongxia. Development and design of a system for comparing and verifying regular meteorological observations and automatic station data[J]. Desert and Oasis Meteorology, 2009,3(S1):42-43.

[5] Zhu G, Yang L, Sun Shaoming. Development and design of quality control algorithm for civil aviation meteorological reports[J]. Technology and Innovation, 2019(20):53-54+58.

[6] Zhu Qi, Tang M. Research on monitoring and feedback system of civil aviation meteorological observation messages[J]. Journal of Civil Aviation Flight Academy of China,2015,27(02):77-80.

[7] GIS-based landslide susceptibility evaluation using a novel hybrid integration approach of bivariate statistical based random forest method[J]. 2018

[8] Ma Xiongwei. Linear regression equation in multiple covariance diagnosis method and its empirical analysis[J]. Journal of Huazhong Agricultural University (Social Science Edition),2008(02):78-81+85.

[9] Alireza Arabameri, Sunil Saha, Jagabandhu Roy, et al. Landslide Susceptibility Evaluation and Management Using Different Machine Learning Methods in The Gallicash River Watershed, Iran[J]. Remote Sensing,2020,12(3).

[10] Liu, Jingjuan. Multiple linear regression model testing methods[J]. Journal of Hunan Taxation College,2005(05):49-50+60.

[11] Qian Ying, Fang Xiunan. Multiple linear regression model and its application[J]. China Science and Technology Information,2022(04):73-74.

[12] Huang Jiayou. Meteorological statistical analysis and forecasting methods (4th edition) [M]. Beijing: Meteorological Publishing House, 2016.5:21-37

[13] Liao Daiqiang, Wu Yao, Chai Bunching. Construction of nonlinear regression models for atmospheric pollutants[J]. Meteorological Science and Technology, 2020,48(06):871-876.

[14] Zhao Hailong, Zhang Dandan, Huang Song, et al. Analysis of the relationship between ground flash density and lightning fault in Hainan Province based on Pearson correlation coefficient[J]. High Voltage Electrical,2019,55(08): 186-192.