ISSN 2616-5805 Vol. 5, Issue 2: 1-7, DOI: 10.25236/AJMS.2024.050201

Humidity research based on multiple regression and time-varying exponential smoothing model

Linfeng Lou[1,a,*], Hua Xu[1,b]

[1]*School of Mathematics and Statistics, Liaoning University, Shenyang, 110000, China*
[a]*20210902317@smail.lnu.edu.cn,* [b]*3400638665@qq.com*
[*]*Corresponding author*

Abstract: *Humidity, as a key indicator in meteorology, is related to the production and sustainability of urban areas. This paper aims to forecast the humidity data of Shenyang City in 2024. We examined four explanatory variables closely related to humidity and developed a multiple regression equation. Considering the mutual influence of these variables, we regard them as variable relations in multidimensional space and determine the weight of each factor by equation coefficients. On this basis, we further develop a smoothing index prediction model. The model incorporates a smoothing coefficient that varies over time and is flexibly adjusted based on actual data to accurately predict the humidity data at the next time point. The regression equation and time series model constructed in this paper comprehensively consider the factors affecting humidity from both spatial and temporal dimensions, aiming to enhance the accuracy and flexibility of the prediction results. This research has far-reaching significance for environmental monitoring, production, and daily life.*

Keywords: *Multiple linear regression; Time-varying exponential smoothing model; Humidity prediction*

1. Introduction

As an important index in meteorology, weather humidity prediction has a far-reaching impact on urban planning, agricultural production, and environmental protection. With the continuous development of numerical models, statistical methods, and artificial intelligence technology, research on humidity prediction has gradually formed a robust theoretical system and methodological framework. This paper aims to discuss the current research status of weather humidity forecasting, introduce several mainstream forecasting methods, and outline their advantages and disadvantages. This information aims to provide a reference for future research on humidity forecasting.

The research on weather humidity prediction has made significant advancements worldwide. Dynamic model forecasting, empirical model forecasting, and intelligent model forecasting are three main forecasting methods. Dynamic model prediction is based on physical processes and has high accuracy, but it is computationally intensive and sensitive to initial conditions. Empirical model prediction and intelligent model prediction utilize statistical relationships in historical data and machine learning algorithms, which possess the characteristics of fast computing speed and strong adaptability. In recent years, with the rapid development of machine learning and deep learning, intelligent pattern prediction based on neural networks has shown great potential in humidity prediction.

In order to thoroughly investigate the correlation between humidity and various factors and accurately predict the trend of humidity, many scholars have conducted in-depth research. Zuo Zhiyu et al. developed a temperature prediction model using time series analysis and achieved precise greenhouse temperature prediction through differential processing and autocorrelation analysis[1]. Guo Qingchun proposed a relative humidity prediction model based on a BP artificial neural network[2]. By training the neural network to capture the changing trend of relative humidity, we achieved high prediction accuracy. Huang Tianyi et al. combined a multi-modal data-driven method and introduced a low-complexity double convolutional layer to extract potential features from images. This approach helps identify the humidity change trend and build the Bayesian ISTM prediction model, leading to improved prediction performance of greenhouse environmental humidity[3]. In addition, the research team at Henan University developed a parallel 3D neural network for strong convection weather prediction based on MPI. This method reads radar echo data from the original Doppler weather radar and generates a 3D neural network training dataset, effectively enhancing prediction accuracy and program load balance [4].

ISSN 2616-5805 Vol. 5, Issue 2: 1-7, DOI: 10.25236/AJMS.2024.050201

Although significant progress has been made in the field of weather humidity forecasting, many challenges remain. Traditional forecasting methods are not flexible enough to accurately capture the dynamic changes in humidity when dealing with complex meteorological data. Therefore, based on previous studies, we propose an innovative prediction method that combines linear regression and a time-varying α exponential smoothing model. The linear regression model can quantify the influence of various explanatory variables on humidity and construct a multi-dimensional relationship between humidity and other meteorological factors. This helps to understand the internal mechanism of humidity change more comprehensively. At the same time, the time-varying α exponential smoothing model is based on the characteristics of time series data. It constructs a two-dimensional function relationship between humidity and time, adapting to the short-term fluctuations of humidity by dynamically adjusting the smoothing coefficient. By combining these two methods, we aim to enhance the flexibility and accuracy of the forecast model. This will enable it to better adapt to complex and changeable meteorological data, providing a more reliable and precise solution for predicting weather humidity.

2. Regression analysis of humidity and its influencing factors

2.1. Descriptive statistics

After a comprehensive review of meteorological data, we identified four factors that have a significant impact on humidity: temperature, wind speed, visibility, and cloud cover. In order to further understand the impact of these factors on humidity, we will conduct detailed descriptive statistical analysis of these four factors to reveal the basic statistical characteristics of each factor, such as average, median, standard difference, etc. The results are shown in Table 1.

In this paper, the collected panel data, which includes factors affecting humidity (from the Hui data), are used for descriptive statistics and to create box plots. At the same time, outlier removal and missing value imputation are performed. The overall distribution of the data can be visually observed in Figure 1, which is discrete and relatively uniform.

Table 1: Descriptive statistics of factors.

Factor	Maximum	Minimum	Mean	Standard deviation	Coefficient of variation
Temperature	14.9	-19.3	-0.862	8.063	-9.354
Wind Speed	5.5	0.9	2.289	0.914	0.399
Visibility	30	6.8	22.452	6.326	0.282
Cloud Cover	94	0	23.65	22.866	0.967

Descriptive statistical analysis indicates that the four groups of influencing factors follow a normal distribution. It can be seen from the coefficient of variation that the temperature has a discrete distribution due to seasonal factors. From the standard deviation data, it can be concluded that the difference in cloud cover data group is large, the difference in wind speed data group is the least, and the data is relatively concentrated.

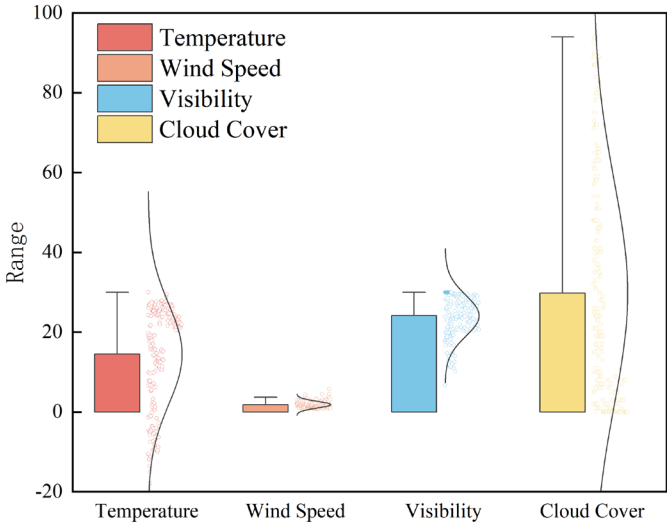

Figure 1: Box diagram and normality test.

ISSN 2616-5805 Vol. 5, Issue 2: 1-7, DOI: 10.25236/AJMS.2024.050201

2.2. Regression model of humidity and influencing factors

2.2.1. Establishment of multiple linear regression model

In this paper, the daily humidity y (%) in Shenyang in 2023 was selected as the response variable. Four influencing factors were considered as explanatory variables, denoted as temperature x1, wind speed x2, visibility x3 and cloud cover x4, respectively.

To further reveal the precise degree of influence of these factors on humidity change in Shenyang, a multiple linear regression model was established. This model clearly expresses the linear relationship between humidity and the four explanatory variables in the form of a mathematical formula, enabling us to quantify the specific contribution and impact of each factor on humidity.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \tag{1}$$

ε is a perturbation term with $\varepsilon \sim N(0,1)$ holds.

β_i $(i = 1,2,3,4)$ is each intercept term of the regression model.

70% of the sample data $\{(y_i, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}); i = 1, 2, ..., 255\}$ is randomly selected and the parameters are estimated by using the principle of least square method.

Python was used to estimate and analyze the parameters of each explanatory variable and constant term in the model, and the results were shown in Table 2.

Table 2: OLS Regression Results.

Variable	Coefficient	Std Error	t-Statistic	Prob
Const	87.3376	2.762	31.623	0.000
Temperature	0.5369	0.045	11.864	0.000
Wind Speed	-7.1343	0.653	-10.929	0.000
Visibility	-1.0574	0.107	-9.837	0.000
Cloud Cover	0.2644	0.022	11.887	0.000

Therefore, the multiple linear regression equation as shown in equation (2) is valid.

$$y = 87.3376 + 0.5369x_1 - 7.1343x_2 - 1.0574x_3 + 0.2644x_4 + \varepsilon \tag{2}$$

2.2.2. Testing of multiple linear regression model

- Economic significance test

On the premise that other variables remain unchanged, the coefficients of x1 and x4 are positive. This indicates that x1 and x4 are positively correlated with the response variable y. The coefficient of the relationship between x2 and x3 is negative, indicating that when x2 and x3 change, y changes in the opposite direction. The above changes are consistent with theoretical analysis and empirical judgment.

In real life, the higher the temperature of a certain volume of air, the more water vapor it can contain. Consequently, the absolute humidity increases, which aligns with the earlier conclusion of a positive correlation. Cloud formation is primarily driven by the condensation of water vapor in the atmosphere. According to the Clausius-Clapeyron equation[5], the increase in average temperature leads to an increase in saturated water vapor pressure, resulting in a decrease in relative humidity and cloud formation. It is concluded that cloud cover is positively correlated with humidity. On the contrary, an increase in wind speed promotes the dispersion of fog and aerosols, leading to improved atmospheric visibility, which is inversely correlated with humidity.

- Statistical test

The coefficient of determination R^2=0.638 and the corrected coefficient of determination R^2=0.634 indicate that the selected model fits the selected samples well.

Assume null hypothesis H_0 is a wireless relationship between *y* and each item *x*. At the significance level of $\alpha = 0.05$, F=153.2, p=0.000, rejecting the null hypothesis, the influence of the four explanatory variables on the response variable humidity is significant.

- Econometric test

ISSN 2616-5805 Vol. 5, Issue 2: 1-7, DOI: 10.25236/AJMS.2024.050201

To test the econometric properties of the model, a multicollinearity test is needed for explanatory variables, and the correlation coefficients between explanatory variables are calculated. Since the descriptive statistics showed that each explanatory variable was normal, Pearson correlation analysis was used to assess the strength of the linear relationship between these variables. The results are shown in Table 3.

Table 3: Explanatory variable phase relationship table.

	Temperature	**Wind Speed**	**Visibility**	**Cloud Cover**
Temperature	1.000	-0.117	0.146	0.020
Wind Speed	-0.117	1.000	0.204	0.174
Visibility	0.146	0.204	1.000	-0.217
Cloud Cover	0.020	0.174	-0.217	1.000

According to the correlation coefficients of each variable shown in Table 3, it can be observed that the correlation between each explanatory variable is maintained below 0.25, which indicates that there is a weak correlation between them and passes the multicollinearity test.

- Model prediction test

In order to evaluate the sensitivity of the model when the sample size changes, we need to test the hyperparameters property of the model, that is, the model is applied to a certain period outside the sample, the predicted value is obtained, and the actual value is compared[6]. 30% panel data outside the model is taken as the validation set and its root-mean-square error is calculated.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_t - \hat{y}_t\right)^2} \tag{3}$$

n is the number of samples.;

y_t is the true value;

$\hat{y}_t$ is the predicted value.

When the difference between outliers and normal values is too large, the error will be greater than 1, and squared will further increase the error, so RMSE is relatively sensitive to outliers. The calculated value is 12.5947, which is within the accepted range and passes the model prediction test.

3. Time varying α exponential smooth prediction of humidity

3.1. Exponential smoothing time series model

The basic formula of exponential smoothing method is Formula (4).

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t \tag{4}$$

y_t is the actual value at time t ;

$\hat{y}_t$ is the exponential smoothing value at time t ;

$\hat{y}_{t+1}$ is the exponential smoothing value at time $t+1$;

α is the smoothing coefficient, whose value range is [0,1].

Exponential smoothing is a classical method used for time series prediction. It involves assigning more weight to recent observations to capture time series features[7]. The basic principle of this paper is that the exponential smoothing value of each period is determined by the weighted average of the actual value of the current period and the predicted value of the previous period. In other words, the result of each forecast is influenced by both the actual value of the current period and the last predicted value.

However, the smoothing coefficient of traditional exponential smoothing models is usually considered a constant in specific situations. May reduce the accuracy of the prediction [8].

ISSN 2616-5805 Vol. 5, Issue 2: 1-7, DOI: 10.25236/AJMS.2024.050201

3.2. Time varying exponential smoothing time series model

3.2.1. Establishment of time-varying exponential smoothing time series model

To avoid the problem that the prediction accuracy of traditional exponential smoothing model is reduced because the smoothing coefficient is regarded as constant, the time-varying α smoothing coefficient can be introduced and dynamically adjusted according to the actual situation of the data.

Based on equation (4), α is modified to time-varying smoothing coefficient α_t.

$$\hat{y}_{t+1} = \alpha_t y_t + (1-\alpha_t)\hat{y}_t \tag{5}$$

▪ Significance of time-varying smoothing coefficient

The significance of a time-varying smoothing coefficient in the exponential smoothing model is that it provides the model with increased flexibility and adaptability. By allowing the smoothing coefficient to change over time, the model can more accurately capture dynamic features of time series data, such as trends, seasonal fluctuations, or abrupt events.

Therefore, the selection of the time-varying smoothing coefficient is very important because it determines the weight distribution between the new data in the new predicted value and the original predicted value in equation (4). When the smoothing coefficient is larger, the proportion of new data in the predicted value is higher. This indicates that the model places greater emphasis on recent changes in the data. On the other hand, if the smoothing coefficient is small, the weight of the original predicted value will be larger, and the model is more inclined to maintain the historical trend.

The calculation formula is set as equation (6).

$$\alpha_t = 1 - \frac{|\triangle y_t|}{H},\ H \in [40, 60] \tag{6}$$

Given H= {40, 50 ,60}, we get the daily values α_{t1}, α_{t2}, and α_{t3}.

By adjusting the time-varying smoothing coefficient α_t, the corresponding exponential smoothing value $\hat{y}_t$ is calculated.

▪ Test of model

RMSE is used to test the advantages and disadvantages of a1,a2 and a3, and its formula is shown as equation (7).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_t - \hat{y}_t\right)^2} \tag{7}$$

$\hat{y}_t$ is the predicted value.

Compared with the other options, α_{t3} corresponds to the smallest root-mean-square error of 12.6035. This data shows that α_{t3}, as a time-varying smoothing index, has the best performance in terms of prediction accuracy, and its prediction results are more accurate and reliable.

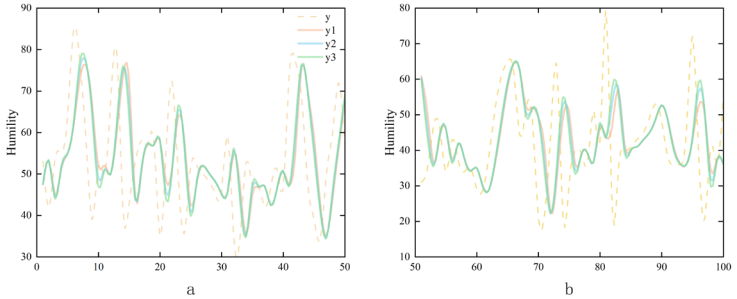

Figure 2: Comparison of three exponential smoothing values with actual humidity values.

ISSN 2616-5805 Vol. 5, Issue 2: 1-7, DOI: 10.25236/AJMS.2024.050201

3.2.2. Prediction of time-varying exponential smoothing time series model

When $t=1$, that is, for January 1, 2023, we use the exponential smoothing method and set the initial exponential smoothing value as follows.

$$\hat{y}_1 = \frac{y_1 + y_2}{2} \tag{8}$$

Then, we plug this value into equation (5), and through continuous iterative calculation, we get the predicted humidity on January 1, 2024. The humidity data for the first 15 days of 2023 is shown below.

Table 3: Humidity data for the first 15 days of 2023.

Date	y	t	Smoothing coefficient	Exponential smoothing value
2023-1-1	53	1	1.000	47.500
2023-1-2	42	2	0.817	53.000
2023-1-3	54	3	0.800	44.017
2023-1-4	55	4	0.983	52.003
2023-1-5	64	5	0.850	54.950
2023-1-6	85	6	0.650	62.643
2023-1-7	77	7	0.867	77.175
2023-1-8	58	8	0.683	77.023
2023-1-9	39	9	0.683	64.024
2023-1-10	52	10	0.783	46.924
2023-1-11	50	11	0.967	50.900
2023-1-12	68	12	0.700	50.030
2023-1-13	79	13	0.817	62.609
2023-1-14	41	14	0.367	75.995
2023-1-15	43	15	0.967	63.163

Finally, we get that the humidity in Shenyang on January 1, 2024 is 73.520%, which is consistent with the actual situation.

3.3. Improvement of time-varying exponential smoothing model

In the application of the time-varying exponential smoothing model, we observed a slight lag phenomenon in the prediction results, as depicted in Figure 2. Upon thorough analysis, we identified that this lag was attributed to the substantial impact of extreme humidity values on the smoothing coefficient in the subsequent time. To improve this situation, it is necessary to impose stricter constraints on the smoothing coefficient.

To achieve this goal, we draw upon the concept of activation functions in the field of machine learning for guidance and introduce innovative modifications to the calculation process of the smoothing coefficient. It is specified that the activation switch value is 20. When the humidity difference between the upper and lower moments exceeds this threshold, we consider it an extreme humidity change. In this case, we will reduce the smoothness coefficient at the next moment to mitigate the adverse impact of extreme humidity on the prediction results. The activation function of the time-varying smoothness index is set as shown in Equation (9).

$$a_t(|\Delta y_t|) = \begin{cases} 1 - \dfrac{|\Delta y_t|}{H}, & if\ |\Delta y_t| \le 20 \\ e^{-|\Delta y_t|}, & if\ |\Delta y_t| > 20 \end{cases} \tag{9}$$

After a series of scientific advancements, we have successfully integrated exponential smoothing method, time-varying smoothing coefficient and activation function to build a more comprehensive and accurate prediction model. Among them, the introduction of activation function is based on the application of nonlinear factors in neural networks[9], and its effect coincides with our expectation to reduce the adverse impact of extreme weather on the prediction results.

Through this innovative combination, we expect to be able to further optimize the predictive performance of the model to show greater robustness in the face of extreme changes in climate factors such as humidity.

ISSN 2616-5805 Vol. 5, Issue 2: 1-7, DOI: 10.25236/AJMS.2024.050201

The improved algorithm of time-varying exponential smoothing model is shown in Figure 3.

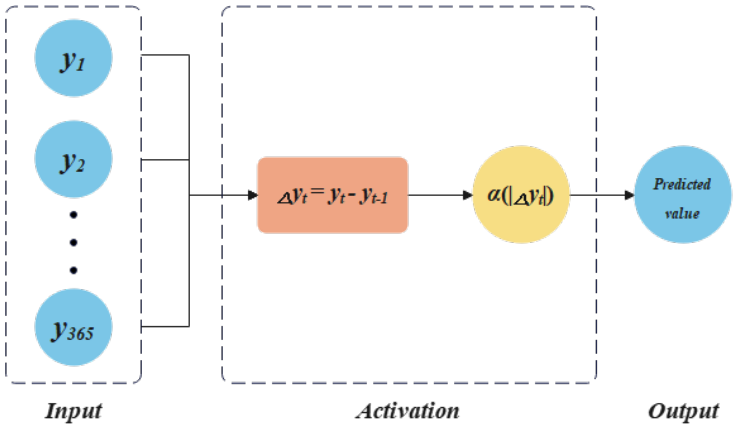

Figure 3: Improved algorithm of time-varying exponential smoothing model.

4. Conclusions

In this study, we conducted a multiple linear regression analysis using the least squares method to unravel the intricate relationships between humidity and pertinent climate factors. The results indicate that as the average temperature rises, the saturated water vapor pressure increases, resulting in a decrease in relative humidity and a reduction in cloud formation. This finding reveals a positive correlation between cloud cover and humidity. On the other hand, an increase in wind speed contributes to the diffusion of fog and aerosols, which, in turn, enhances atmospheric visibility, exhibiting a negative correlation with humidity.

Then, we establish a two-dimensional function to represent the relationship between humidity and time and apply the exponential smoothing model for analysis. By introducing the time-varying smoothing coefficient, we can accurately capture the dynamic characteristics of time series data, thereby enhancing the accuracy of predictions. In order to address the lag problem in the forecast results, we have introduced an activation function. This function can effectively mitigate the adverse impact of extreme weather on the forecast results.

Looking forward to the future, we will further explore the application of activation functions in setting the smoothing coefficient. We aim to deepen our understanding of relevant knowledge in the field of machine learning. Through continuous research and improvement, we aim to enhance our ability to offer more accurate and reliable model support for humidity prediction.

References

[1] Zuo Zhiyu, Mao Hanping, Zhang Xiaodong, Hu Jing, Han Greening, Ni Jing. Greenhouse temperature prediction model based on time series analysis[J]. Journal of Agricultural Machinery, 2010, 41(11): 173-177182

[2] Guo Qingchun, He Zhenfang, Hui Ying, Li Xue. Application of artificial neural network in relative humidity prediction [J]. Modern Food Science and Technology, 2013, 29(6):1297-1301

[3] Huang Tianyi, Wu Huarui, Zhu Huaji. A multimodal data-driven humidity prediction method for cucumber greenhouse [J]. Electronic Measurement Technology, 2023, 46(16):97-104

[4] Zhang Lei, Meng Kunying, Shen Xiajiong, et al. An MPI-parallelized three-dimensional neural network-based method for strong convective weather prediction: 202110832813

[5] Jia Pengqun, Li Jinghua. An immortal classic:The Clausius-Clapeyron equation after 190 years[J]. Advances in Meteorological Science and Technology, 2023, 13(2):2-4

[6] Li Zinai. Econometrics: Methods and Applications[M]. Beijing: Tsinghua University Press, 2020.

[7] Wang TX. Dynamic triple exponential smoothing method with dynamic smoothing coefficients and parameters for power generation forecasting in thermal power plants [J]. Power Equipment Management, 2024(1):150-152

[8] Zhao Tianze, Hu Xueyou, et al. Methanol price prediction based on adaptive exponential smoothing method [J]. Journal of Bengbu Institute, 2023, 2.

[9] Zhou Zhihua. Machine Learning[M]. Beijing. Tsinghua University Press. 2016.