An Analysis of Typhoon Paths and Intensity Based on Feature Engineering and Machine Learning

Dongming Ma *

College of computer science and technology, Huaqiao University, Xiamen, China, 361021

* Corresponding Author Email: m1692463291@163.com

Abstract. Given the increasing frequency of typhoon disasters and persistent prediction challenges, this study develops a typhoon classification model using machine learning to improve forecast accuracy. The research analyzed correlations between typhoon characteristics and environmental factors using Spearman coefficients, then visualized typhoon paths of different intensities. Recognizing multiple influencing factors, the study performed clustering analysis based on two key features: typhoon intensity and moving path, integrating these results for a comprehensive classification model. The model was validated by predicting typhoons in 2024. Results showed typhoon intensity significantly correlates positively with class and wind speed, while negatively with air pressure, and typhoon paths exhibit strong seasonal and geographical differences. K-means clustering determined the optimal number of clusters as 3 for intensity (silhouette coefficient: 0.545) and 4 for paths (silhouette coefficient: 0.615), indicating effective clustering. This model optimizes disaster response strategies and improves emergency management efficiency, offering substantial socioeconomic value for protecting lives and property in the context of increasing extreme weather events due to climate change.

Keywords: K-means clustering algorithm, Average profile factor, Spearman's correlation coefficient.

1. Introduction

Typhoons, as an extreme weather phenomenon, often affect countries and regions along the coast of the northwestern Pacific Ocean, causing severe loss of life and property. Tropical cyclone is a cyclonic vortex that occurs in the tropical or subtropical ocean, and this vortex can be divided into tropical depression, tropical storm, strong tropical storm, typhoon, strong typhoon and super typhoon according to its intensity [1], China is located in the west coast of the northwest Pacific Ocean, which is the most active region of the world's tropical cyclones, and also one of the countries in the world that suffered from typhoon disasters the most serious, according to the “China's Meteorological Utility Development Strategy Study General Theory”. Based on the overview of “China Meteorological Development Strategy Research General Introduction”, Qin et al. (2005) put forward the strategic idea of “adhering to the development direction of public meteorology, vigorously improving the ability of meteorological information to safeguard national security, and vigorously improving the ability of meteorological resources to support sustainable development” [2]. Since the “Eighth Five-Year Plan” scientific and technological research since the National Meteorological Center of the typhoon numerical prediction business system continues to develop [3], the numerical prediction of the typhoon more accurate. With the intensification of global climate change, the paths and intensities of typhoons have become more complex and unpredictable, so improving the accuracy of typhoon forecasts is crucial for disaster prevention and mitigation.

Current typhoon research primarily focuses on path and intensity prediction [4]. Traditional numerical weather prediction depends on physical models, but typhoon dynamics are influenced by complex atmospheric and oceanic factors that are difficult to fully characterize through physical models alone. Recent advances in machine learning and data-driven approaches have enabled researchers to enhance typhoon prediction capabilities using statistical models and artificial intelligence. Qian et al. (2021) developed an objective typhoon intensity estimation technique using pre-trained ResNet deep learning with satellite cloud imagery from 2005-2018 Northwest Pacific and South China Sea typhoons [5]. Zhou et al. (2023) addressed the multi-factorial nature of typhoon intensity by creating a rapid intensification trend judgment technique combining ResNet and LSTM models with life cycle indicators [6]. Typhoon path analysis has also advanced significantly. Zhao et al. (2022) introduced a path classification method based on similar threshold learning combined with an adaptive mode processing algorithm using fastDTW, improving classification intelligence while reducing traditional method randomness [7]. Zhou et al. (2020) proposed a neural network ensemble forecasting model for typhoon paths that combines hybrid ensemble forecasting with back-propagation multi-layer feed-forward training, offering a novel integration of numerical forecasting and artificial intelligence [8].

Previous typhoon studies typically focus on single dimensions (intensity or path) without integrated analysis models. These studies face challenges: inconsistent feature engineering standards leading to inefficient use of meteorological data, and supervised learning methods requiring costly labeled data that limit model generalization. This study innovatively develops an integrated typhoon classification model incorporating multiple environmental factors (intensity, path, wind speed, air pressure, monsoon) and analyzes relationships between these features using clustering methods. Unlike traditional supervised approaches, this research employs K-means clustering for typhoon intensity and path classification, reducing dependence on labeled data through unsupervised learning, thereby enhancing model flexibility, generalization capability, and prediction accuracy.

2. Data Source and Preprocessing

2.1. Data collection and processing

First of all, the data in this study come from the China Meteorological Network, from which the information on typhoon paths offshore China from 1945 to 2023 is collected, including typhoon path information, typhoon intensity, barometric pressure, central wind speed, moving speed, moving direction, Chinese and English names of typhoons, and numbers, etc., but some of the data are missing. The latitude, longitude, and time data in the typhoon track are very complete in the dataset. Considering that the moving speed and direction of the typhoon are very important for the typhoon track prediction model, the data are supplemented by using the half-positive vector formula and the azimuthal angle calculation formula, which are as follows:

$$d = 2 \cdot R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{d_{lat}}{2}\right) + \cos\left(lat_1\right) \cdot \cos\left(lat_2\right) \cdot \sin^2\left(\frac{d_{lon}}{2}\right)}\right), \tag{1}$$

$$a = a\tan 2\left(\sin\left(d_{lon}\right) \cdot \cos\left(lat_2\right), \cos\left(lat_1\right) \cdot \sin\left(lat_2\right) - \sin\left(lat_1\right) \cdot \cos\left(lat_2\right) \cdot \cos\left(d_{lon}\right)\right). \tag{2}$$

Where R is the radius of the earth, usually taken as 6371 kilometers. d_{lat} is the difference in longitude between these two points, d_{lon} is the difference in latitude between these two points, d is the straight-line distance between these two points, and $angle$ is the angle of the direction of movement of the typhoon.

Typhoon movement speed was calculated by dividing the straight-line distance between adjacent positions (derived from the semi-positive vector formula) by the time difference. To verify data accuracy, this study employed Pearson's correlation coefficient and SPSS regression analysis to assess relationships between estimated and actual values. Analysis revealed a Pearson's correlation coefficient of 0.71 for movement speed and 0.58 for movement direction, validating the supplementary data's reliability.

2.2. Data verification

This study employed the K-S test to assess data normality. The KS statistic (0-1) measures the maximum difference between sample and theoretical distributions, with larger values indicating greater differences. The p-value quantifies the probability of observing the KS statistic or more extreme values when the null hypothesis is true. Analysis of Table 1 revealed that typhoon intensity, class, wind speed, barometric pressure, air temperature, monsoon, movement direction, and movement speed all deviated from normal distribution.

Table 1. K-S test table

	KS statistic	p-value
dissociation	0.76	0.0
hierarchy	0.95	0.0
air velocity	0.95	0.0
pneumatic	1.0	0.0
temperatures	0.92	0.0
monsoon	0.91	0.0
direction of travel	0.82	0.0
travel speed	0.98	0.0

Since the data such as intensity, class, wind speed, barometric pressure, air temperature, monsoon, direction of movement, speed of movement, etc. do not conform to normal distribution, finally in this study the outliers are eliminated using box and line plot, which is a standardized method of displaying the distribution of data is used for detecting outliers, analyzing the shape of the data distribution and evaluating the degree of dispersion of the data. The results of the box-and-line plot calculation are shown in Table 2.

Table 2. Box-and-line diagram calculation results

	IQR	lower limit	limit
dissociation	2.0	-1.0	7.0
hierarchy	4.0	2.0	18.0
air velocity	17.0	-7.5	60.5
pneumatic	28.0	928	1040.0
temperatures	14.1	-7	48
monsoon	1.0	0.5	4.5
direction of travel	10.0	-12.0	28.0
travel speed	13.0	-7.5	44.5

3. Construction of Typhoon Classification Model

3.1. Correlation analysis between typhoon characteristic parameters and environmental factors

The Spearman correlation coefficient is a nonparametric statistic that measures the monotonic relationship between two variables [9]. Before analyzing the relationship between the characteristic parameters of typhoons (intensity, magnitude, wind speed, etc.) and air temperature, barometric pressure, and monsoon, it was found that there was no linear relationship between these data through a simple analysis of the relationship between these data, and therefore, Spearman's correlation coefficient was used in this study to perform the correlation analysis.

As shown in Fig. 1, there is a significant strong positive correlation between typhoon intensity and typhoon class and typhoon wind speed, and a significant negative correlation with air pressure, which means that the higher the intensity of a typhoon, the higher the typhoon class and wind speed, and the lower the central air pressure will be. In meteorology, a typhoon is a low-pressure system. When the typhoon intensity increases, the level increases and the wind speed becomes faster, the central air pressure of the typhoon will generally decrease, and low air pressure and high wind speed are also important characteristics of a typhoon.

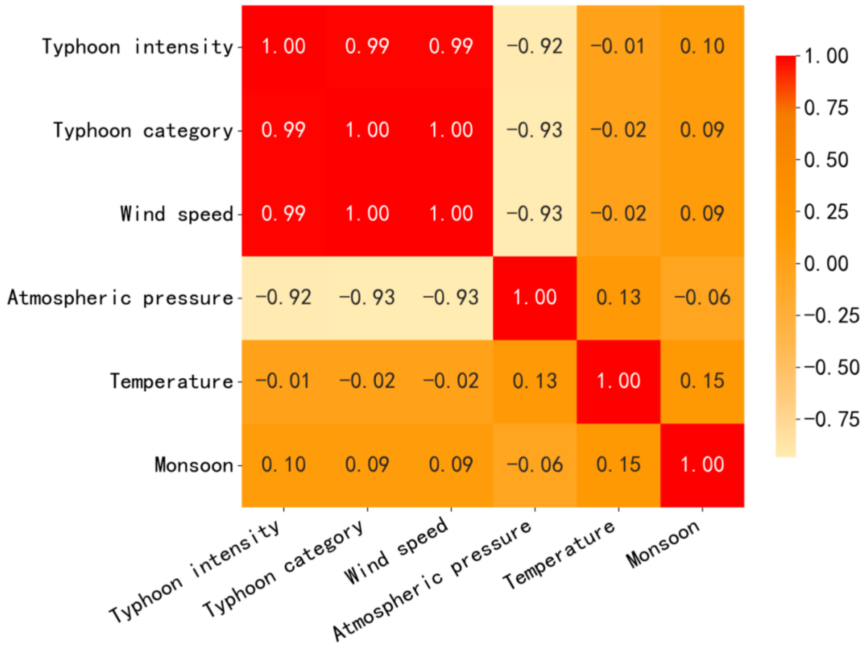

Figure 1. Heat map of correlation

The weak correlation between environmental factors such as temperature and monsoon and the intensity, wind speed, magnitude, and barometric pressure of typhoons implies that the relationship between environmental factors such as temperature and barometric pressure and the intensity, wind speed, magnitude, and barometric pressure of typhoons is weak and less influential.

3.2. Typhoon cluster analysis

K-means is a commonly used unsupervised learning algorithm for datasets without category labels [10]. Given that the given dataset is not accompanied by label information, the K-means clustering algorithm is used in this study to categorize typhoons.

The previous heat map of the correlation between typhoon characteristic parameters and environmental factors shows that there is a strong correlation between typhoon intensity and typhoon class, wind speed, and air pressure, so the intensity, wind speed, class, and air pressure of typhoon are chosen as the characteristics of typhoon intensity clustering in this study.

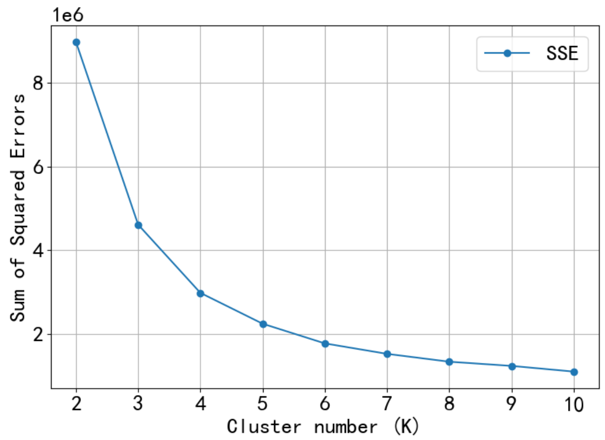

Figure 2. Strength elbow method results in

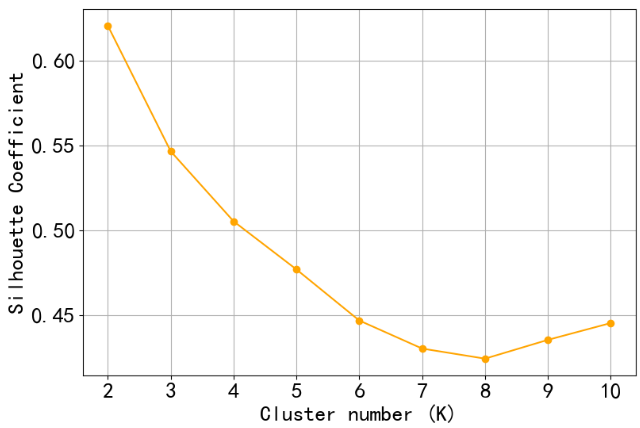

Figure 3. Strength profile coefficients.

By analyzing the results of the elbow method of intensity clustering in Fig. 2 and the contour coefficients of intensity clustering in Fig. 3, the following conclusions are drawn: in the graph of the results of the elbow method of intensity clustering of the typhoon when $K = 3, 4, 5, 6\cdots$, there is a point of inflection, in which the curve of the change of the SSE with the value of K is $K = 3$ curved when $K = 3$; in the graph of the contour coefficients of intensity clustering of the typhoon the average contour coefficient in the point of inflection is the largest at $K = 3$, that is to say, the most optimal effect of the clustering $K = 3$ and most stable, the value at this time is optimal on the number of clusters.

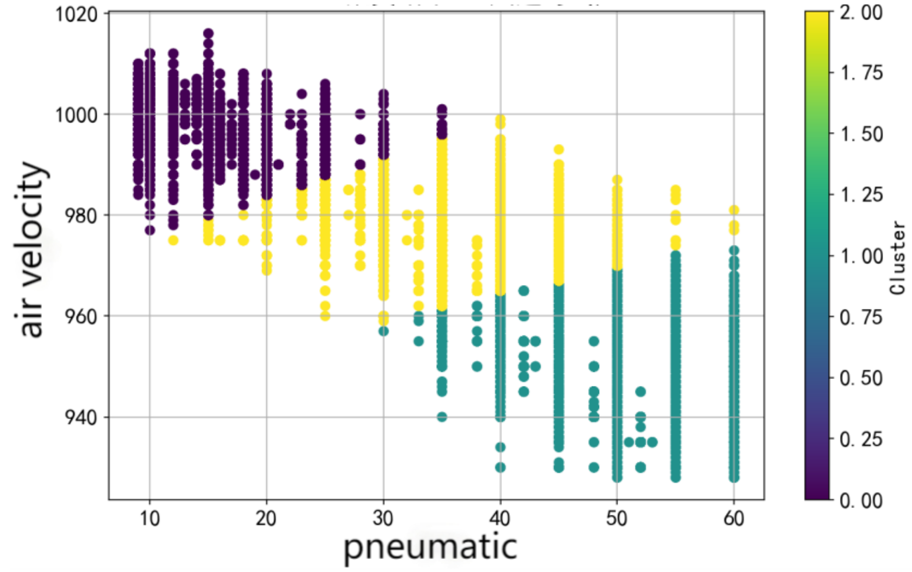

Figure 4. Intensity clustering results

Analysis of intensity clustering results in Fig. 4 reveals three typhoon categories:

Strong typhoons feature wind speeds exceeding 30m/s and lower air pressure ranging from 930-970hpa. These powerful storms cause significant coastal damage and are categorized as highly destructive.

Medium typhoons exhibit wind speeds between 20-35m/s and air pressure from 960-1,000hpa. With moderate intensity, they cause partial coastal impacts.

Weak typhoons show wind speeds between 10-20m/s and higher air pressure ranging from 980-1,020hpa. Their limited intensity results in minimal coastal impact.

While the dataset provides latitude and longitude information, using only these coordinates as clustering features proves insufficient. Therefore, this study enhances path clustering by constructing five additional parameters: longitudinal center of gravity, latitudinal center of gravity, longitudinal variance, latitudinal variance, and covariance for each typhoon. The computational formulas are as follows:

$$\bar{X} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i x_{i,} \tag{3}$$

$$\bar{Y} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i y_{i,} \tag{4}$$

$$Var(x) = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \left(x_i - \bar{X} \right), \tag{5}$$

Where $\bar{X}, \bar{Y}$ is the latitude and longitude center of mass of each typhoon, x_i, y_i is the latitude and longitude of the typhoon at moment i, n is the number of times the typhoon path is localized, and w_i is the intensity weight of the typhoon path at moment i. (Here in this study, we chose the square root of the wind speed as a proxy to avoid too large a difference in the weights of the typhoon data.)

$Var(x)$, $Var(y)$ and $Var(xy)$ are the variance of the latitude and longitude and the covariance, respectively.

After obtaining the feature parameters for path clustering, the same elbow hair rule and contour coefficients are used on them to determine their optimal number of clusters, and the results are shown in Fig. 5 and Fig. 6, and then the clustering is performed using k-means.

The optimal number of clusters for path clustering is selected in the same way as for intensity clustering, and the optimal number of clusters for paths is obtained to be 4, and k-means clustering is performed on them according to.

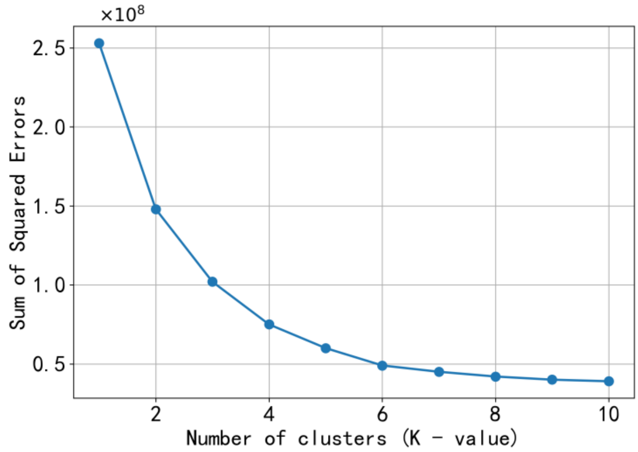

Figure 5. Path clustering elbow rule results in

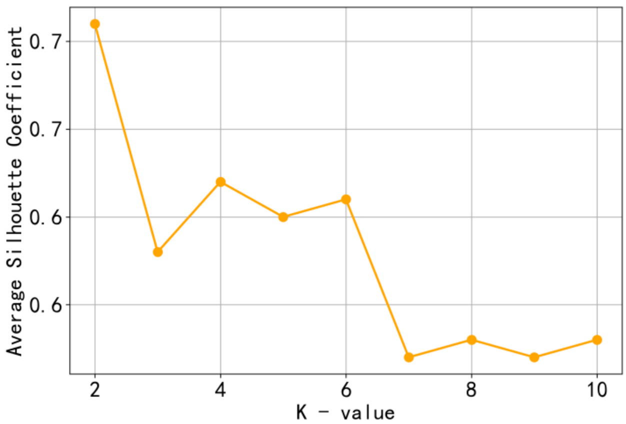

Figure 6. Path clustering profile coefficients.

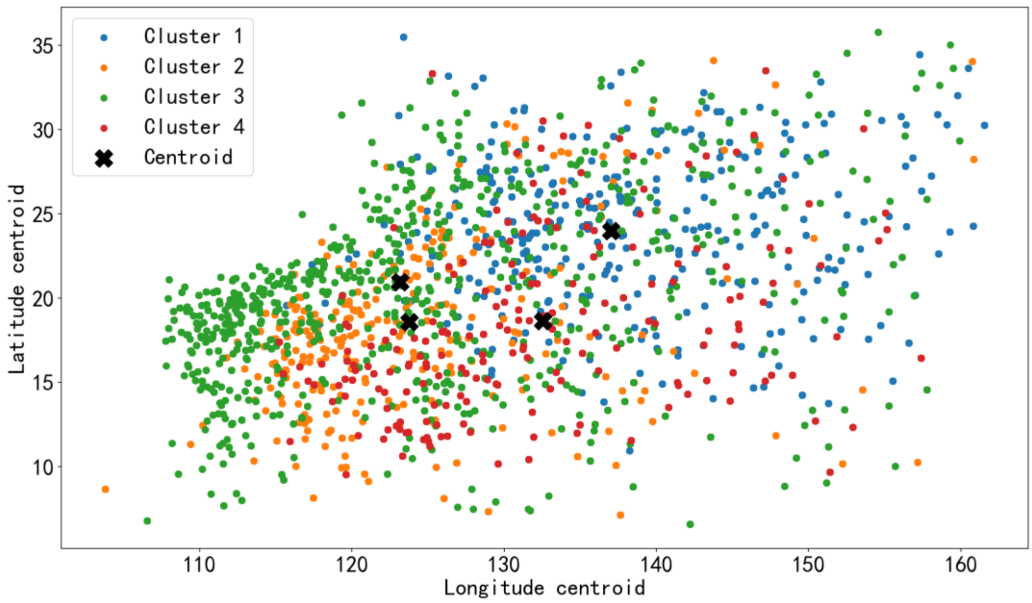

Figure 7. Path clustering results

Based on path clustering results in Fig. 7, typhoons are classified into four categories with distinct movement patterns.

Category 1 typhoons move westward, generated in northwestern Pacific tropical waters east of Philippines and in the South China Sea. They follow stable westward paths crossing the Philippines into the South China Sea, affecting Southeast Asian countries like Vietnam, Laos, and Cambodia, with limited impact on eastern China. These typhoons are primarily active between 5°-25°N and 105°-135°E.

Category 2 typhoons follow northwest-northeast trajectories, originating in similar regions as Category 1. They initially move northwestward, then turn northeastward near the Philippines or southeastern China due to westerly systems. Their influence extends to Japan and potentially Alaska, with activity mainly between 5°-40°N and 110°-140°E.

Category 3 typhoons take southwest-northeast paths, beginning in northwestern Pacific tropical waters. They initially move westward or southwestward before turning northeast near the Philippines or southern China. These affect a wide area including the Philippines, southeastern China coastal regions, the island region east of Fujian province, Japan, and Northeast Asia, typically active between 15°-40°N and 120°-160°E.

Category 4 typhoons travel northbound, generated in similar regions but moving directly north or northeast. They affect eastern China, South Korea, Japan, and Russia's far east coast, with primary activity between 10°-30°N and 110°-135°E.

3.3. Integrated Typhoon Classification

In order to verify the accuracy of the typhoon classification model established in this study, this study obtains the indicators of the 2024 Chinese typhoons from the China Meteorological Network, and analyzes the 2024 typhoons according to the established classification model, so as to provide a more accurate typhoon early warning and disaster prevention and mitigation support, and the results of the integrated consideration of the intensity characteristic category and the path characteristic category of the 2024 Chinese typhoons are shown in Table 3.

Table 3. Typhoon categories and provinces passing through China in July and September 2024

months	Typhoon category	by way of provinces
July Summer Typhoon	Northwest-Northeast Weak Typhoons	Fujian, Jiangxi, Hubei
	Northwest-Northeast Strong Typhoon	Fujian
	Northwest-Northeast Central Typhoon	not have
	Westbound Weak Typhoon	Hainan
September Autumn Typhoon	Westbound Weak Typhoon	Hainan
	Westbound Strong Typhoon	not have
	Westbound Central Typhoon	Hainan, Guangdong
	Northwest-Northeast Weak Typhoons	Jiangsu, Anhui, Henan
	Northwest-Northeast Strong Typhoon	Shanghai, Jiangsu, Zhejiang
	Northbound Weak Typhoons	Zhejiang, Shanghai, Jiangsu

4. Conclusion

In this study, firstly, the scatterplot between typhoon characteristic parameters and environmental factors was plotted using SPSS, and the results showed that the typhoon characteristics did not conform to a linear relationship with the environmental factors. For this reason, the Spearman correlation coefficient was used to analyze the correlation between typhoon intensity, class, wind speed, and environmental factors such as temperature, air pressure, and monsoon, and it was found that the typhoon intensity was significantly and positively correlated with the class and wind speed, while it was negatively correlated with the air pressure. On this basis, the paths of typhoons of different intensities are visualized and analyzed, revealing the seasonal and geographical differences in typhoon paths. Given that typhoon types are affected by a combination of factors, this study chooses typhoon intensity and moving path as the main features and uses the K-means clustering algorithm to classify them separately, classifying typhoons into three categories of strong, moderate, and weak according to intensity by the elbow rule and the profile coefficient, and four categories of westward, northwest-northeast, southwest-northeast, and northward according to the moving trajectory, and considering the results of the two kinds of clusters together to Typhoon classification is carried out. Finally, based on the established classification model, the study classifies and predicts typhoons in 2024. However, there are still some shortcomings in this study, such as the K-means clustering algorithm used in the study lacks robustness to outliers and may be affected by extreme typhoon events, which affects the stability and reliability of the clustering results. In the future, more robust clustering methods such as DBSCAN or GMM can be tried to improve the stability of classification. In addition, more environmental variables, such as SST, humidity, and climate index (ENSO), can be introduced to further enhance the science and accuracy of typhoon classification.

This study presents an effective framework for typhoon characterization and prediction, proving its viability in meteorological data classification. By integrating typhoon intensity and path characteristics through Spearman correlation analysis and K-means clustering, the model enables scientific typhoon classification and prediction. This approach offers both a reference for future typhoon prediction models and methodological support for other meteorological studies, revealing hidden patterns in complex data and enhancing prediction accuracy and reliability.

References

[1] Zhou Junhua, Shi Peijun, Chen Xuewen. Spatiotemporal differentiation of tropical cyclone activity in the Northwest Pacific from 1949 to 1999 [J]. Journal of Natural Disasters, 2002, (03): 44 - 49.

[2] Qin Dahuo, Sun Honglie, Sun Shu, Liu Yingjin. Development strategy for China's meteorological services from 2005 to 2020 [J]. Advances in Earth Science, 2005, (03): 268 - 274.

[3] Ma Suohong, Shen Xueshun, Gong Jiandong, et al. Development of the National Meteorological Center's tropical cyclone numerical forecasting system[J]. Meteorological Monthly, 2021, 47 (6): 685 - 692.

[4] Liao Fei, Li Wenting, Zhang Ziran, et al. Statistical characteristics of rapid intensity and path changes of tropical cyclones in the South China Sea from 1949 to 2017 [J]. Acta Oceanological Sinica, 2019, 41 (09): 126 - 135.

[5] Qian Qifeng, Wang Chuan, Xu Yajing, et al. A typhoon intensity estimation technique based on deep learning[J]. Meteorological Monthly, 2021, 47 (5): 601 - 608.

[6] Zhou Guanbo, Qian Qifeng, Xu Yinglong. Application of artificial intelligence-based typhoon intensity sudden change discrimination technology[J]. Journal of Natural Disasters, 2023, 32 (06): 96 - 103.

[7] Zhao Danfeng, Xu Qiang, Huang Zhou, et al. A typhoon track classification method based on similarity threshold learning [J]. Marine Environmental Science, 2022, 41 (4): 644 - 652.

[8] Zhou Xiaotian, Zhang Feng, Du Zhenhong, et al. Optimization of typhoon track forecasting based on neural network ensemble forecasting [J]. Journal of Zhejiang University (Science Edition), 2020, 47 (02): 196 - 202.

[9] Mary Balogun, Obeng-Gyasi Emmanuel. Association of Combined PFOA, PFOS, Metals and Allostatic Load on Hepatic Disease Risk [J]. J Xenobiot, 2024, 14 (2): 516 - 536.

[10] Kristina-P Sinaga, Yang Miin-Shen. Unsupervised K-Means Clustering Algorithm [J]. IEEE Access, 2020, 880716 - 80727.