A Multivariate Typhoon Time Series Similarity Measurement Method Postprint

Authors: Huang Dongmei, Zheng Xia, Zhao Danfeng, Wang Lilin

Date: 2018-05-18T00:00:00+00:00

Abstract

Research on typhoon similarity measurement methods holds significant importance for disaster prevention and mitigation, decision support, and related applications. Currently, studies on typhoon similarity predominantly concentrate on the similarity measurement of typhoon tracks. First, we systematically analyze multiple factors influencing typhoon similarity measurement and propose a multivariate time series-based description method for typhoon data; second, we propose assessment and repair methods for the integrity and consistency of typhoon time series; finally, to address the issue of unequal lengths in typhoon time series, we design a similarity measurement method based on principal component analysis and dynamic time warping distance. Experimental verification demonstrates that the proposed method can effectively measure typhoon similarity.

Full Text

Preamble

A Similarity Measure Method for Multivariate Typhoon Time Series

Huang Dongmei[1], Zheng Xia[1], Zhao Danfeng[1]†, Wang Lilin[2]
(1. College of Information, Shanghai Ocean University, Shanghai 201306, China; 2. East China Sea Forecast Center, State Oceanic Administration, Shanghai 200129, China)

Abstract: Research on typhoon similarity measurement is of great significance for disaster prevention and mitigation as well as auxiliary decision-making. Currently, most typhoon similarity studies focus on path similarity measurement. This paper first identifies multiple factors affecting typhoon similarity measurement and proposes a multivariate time series-based description method for typhoon data. Second, it presents methods for assessing and repairing the integrity and consistency of typhoon time series. Finally, to address the unequal length problem of typhoon time series, a similarity measurement method based on Principal Component Analysis (PCA) and Dynamic Time Warping (DTW) distance is designed. Experimental validation demonstrates that this method enables effective similarity measurement of typhoons.

Keywords: similarity measure; multivariate time series; integrity; consistency; weight calculation; dynamic time warping

0 Introduction

Typhoons represent one of the most significant marine disasters affecting China. In 2016 alone, typhoons caused 174 deaths, 24 missing persons, and direct economic losses of 76.65 billion yuan [1]. Therefore, typhoon research is crucial for disaster prevention, mitigation, and auxiliary decision-making. Typhoon similarity measurement serves as an important tool for path forecasting and disaster prediction, and its study can reduce casualties and economic losses. However, quality issues are prevalent in typhoon data, leading to biased analysis results that fail to provide accurate information and cause erroneous disaster forecasts, resulting in unnecessary property damage and casualties. Consequently, assessing typhoon data quality and performing repairs constitutes an essential component of similarity measurement.

Typhoon data quality primarily manifests in two aspects: integrity and consistency. Literature [2] reviews the research status of data integrity and consistency, but these methods mainly focus on the medical field with little involvement in the marine domain. Li Jianzhong' s team [3-6] has conducted extensive research on data integrity and consistency, but most of their work targets ordinary relational data. Since typhoon data is a special type of relational data with temporal attributes, these methods are not applicable. In the marine field, literature [7-9] uses sampling methods to test whether marine data quality is qualified but does not propose repair methods. Currently, research on marine data quality assessment and repair remains scarce, yet the quality of marine data directly affects related research such as marine forecasting. Thus, the widespread application of data urgently demands guarantees of data quality.

Researchers worldwide have conducted in-depth mining of time series and proposed various effective similarity measurement methods for different research focuses, which can be broadly divided into two categories. The first category suits equal-length time series, including pattern distance, cosine distance, and Euclidean distance. Pattern distance [10] represents the difference in trend between two sequences with clear physical concepts and reasonable division, but its representation is coarse and conclusions lack precision. Cosine distance based on the angle [11] is simple to compute but requires matching sequences to have identical lengths. Euclidean Distance (ED) proposed by Agrawal et al. [12] is computationally simple with linear time complexity and is widely applied, but it only applies to similarity measurement of equal-length time series.

The second category suits non-equal-length time series, with Dynamic Time Warping (DTW) being the most representative. DTW [13] is a similarity measurement method that warps the time axis to better match the morphology of time series. It was originally a mainstream method in speech recognition, and later Berndt et al. [14] applied it to time series similarity measurement. Literature [15] proposed a maximum-feature-first matching principle and adaptive constraint window method based on DTW for ECG signal similarity measurement. Literature [16] proposed a DTW clustering-based similarity mining method for hydrological time series, improving the efficiency of finding similar hydrological time series. Literature [17] proposed a novel DTW-based similarity measurement method for stock market time series, effectively addressing the price-volume relationship problem in stock market technical analysis. Literature [18] proposed a joint identification method that can effectively identify key joints through certain movements. Currently, using DTW for similar typhoon judgment remains rare. Literature [19] used Hausdorff distance to judge typhoon path similarity, while literature [20] generalized typhoon paths as curves on a plane and compared both numerical and shape similarity for path similarity judgment. Literature [21] used similarity divergence to search for typhoons with similar paths. These studies mostly focus on single-factor similarity of typhoons. This paper investigates multivariate typhoon time series with different lengths and varying importance of each factor in the similarity measurement process, thus selecting weighted DTW distance for similarity measurement research.

This paper aims to design a similarity measurement method suitable for data with spatial, multi-factor, and unequal-length characteristics to achieve typhoon similarity determination and provide accurate and convenient auxiliary decision-making methods for relevant departments. First, we define multivariate typhoon time series and its description method. Then, we present quality assessment and repair methods for typhoon time series. On this basis, considering the unequal length of typhoon time series and the different importance of each factor in similarity measurement, we propose a multivariate typhoon time series similarity measurement method based on weighted dynamic time warping distance. Finally, we analyze the effectiveness of the proposed method through experiments.

---

1 Symbols and Definitions

Definition 1 (Multivariate Typhoon Time Series). A series of typhoon factor observation values arranged in chronological order is called a multivariate typhoon time series, denoted as:

$$MDMVPLoLaGVU_{12} = \{F_1, F_2, ..., F_m\}$$

where: - t: observation time point, $(t = 1, 2, ..., n)$ - j: number of observed variables, $(j = 1, ..., m)$ - MD: moving direction - MV: moving speed - P: pressure - La: latitude - Lo: longitude - G: grade - V: wind speed - U: typhoon factor set - μ: an abstract measurement function - L: typhoon time series length - A: integrity coefficient threshold - C: consistency coefficient threshold - Q_{max}: maximum wind speed near the center at near ground level for 1 minute average (unit: knots) - P_c: minimum pressure at typhoon center (unit: hPa)

Definition 2 (Factor Dependency). For any two factors F_i and F_j among all typhoon factors, if F_i is known, F_j can be calculated, then F_j is said to depend on F_i, denoted as $F_i \to F_j$. Typhoon grade G depends on wind speed V, denoted as $V \to G$.

---

2 Typhoon Time Series Quality Assessment and Repair Strategy

This chapter addresses the prevalent issues of data incompleteness and inconsistency in typhoon data, studies typhoon data integrity and consistency measurement problems, and designs repair methods based on factor dependency, neighbor values, and other typhoon data repair approaches.

2.1 Typhoon Time Series Quality Assessment Method

2.1.1 Typhoon Time Series Integrity Assessment Method This typhoon data integrity assessment method uses three concepts: factor integrity, point integrity, and sequence integrity.

a) **Factor Integrity**: Factor integrity refers to the completeness degree of a factor at time t. For any tuple S_i in S and factor F_j, the integrity of factor value $S_i(F_j)$ is denoted as $\mu_F(S_i(F_j))$, which can be expressed as:

$$\mu_F(S_i(F_j)) = \begin{cases} 1, & \text{if } S_i(F_j) \text{ is not null} \\ 0, & \text{otherwise} \end{cases}$$

Depending on the application, μ_F can have different forms. In this paper, the function μ_F is defined as above.

b) **Point Integrity**: Point integrity refers to the completeness degree of all typhoon factors at time t. For any data $S(t)$ at time t in S, the integrity of $S(t)$ is denoted as $\mu_{SF}(S(t))$. The integrity of $S(t)$ can be determined by the integrity of factor values in $S(t)$ and is defined as:

$$\mu_{SF}(S(t)) = \begin{cases} 0, & \text{if all main factors in } S(t) \text{ are null} \\ 0.2, & \text{if one main factor in } S(t) \text{ is non-null} \\ 0.4, & \text{if two main factors in } S(t) \text{ are non-null} \\ 0.6, & \text{if three main factors in } S(t) \text{ are non-null} \\ 0.8, & \text{if four main factors in } S(t) \text{ are non-null} \\ 1, & \text{if all main factors in } S(t) \text{ are non-null} \end{cases}$$

c) **Sequence Integrity**: Sequence integrity refers to the completeness degree of a typhoon data record. For any typhoon data S, its integrity is denoted as $\mu_{SF}(S)$. The integrity of S can be determined by the integrity of data $S(t)$ at time t in S and is defined as:

$$\mu_{SF}(S) = \frac{\sum_{i=1}^{n} \mu_{SF}(S(t_i))}{n}$$

where $\mu_{SF}(S(t_i))$ is the integrity of data $S(t_i)$ at time t_i, and n is the number of data points with integrity value. When $\mu_{SF}(S) \geq A$, the typhoon time series integrity is repairable.

2.1.2 Typhoon Time Series Consistency Assessment Method This typhoon data consistency measurement framework uses two concepts: point consistency and sequence consistency.

a) **Point Consistency**: Point consistency means that typhoon data at time t does not contain semantic errors or contradictory data. For any data $S(t)$ at time t in S, the consistency of $S(t)$ is denoted as $\mu_{SC}(S(t))$.

b) **Sequence Consistency**: Sequence consistency means that a typhoon data record does not contain semantic errors or contradictory data. The consistency of S can be determined by the consistency of data $S(t)$ at time t in S and is defined as:

$$\mu_{SC}(S) = \frac{N_L}{N}$$

where N_L is the number of data points satisfying consistency. When $\mu_{SC}(S) \geq C$, the typhoon time series consistency is repairable.

2.2 Typhoon Time Series Repair Method

According to literature [22], repair results for integrity errors can cause changes in consistency, timeliness, and accuracy. Therefore, this paper repairs typhoon data in the order of integrity repair first, followed by consistency repair.

There exists a dependency relationship between typhoon grade and wind speed factors. This section considers factor dependency for integrity repair. For factors that cannot be repaired using factor dependency, this paper mainly considers four factors: wind speed, moving direction, moving speed, and pressure, using relationships between neighboring values for integrity repair. For pressure and wind speed that cannot be repaired using neighbor values, the wind-pressure relationship [23] is used for integrity repair. For moving direction and speed that cannot be repaired using neighbor values, repair is performed based on latitude, longitude, and observation time intervals. Since consistency errors in typhoon time series can only occur between grade and wind speed, we use factor dependency for consistency repair.

The specific algorithm is as follows:

```
Input: Typhoon time series S with integrity and consistency errors
Output: Repaired typhoon time series S'

for each S_i in S
    // Integrity repair based on factor dependency
    if G_i == null && V_i != null
        G_i = f(V_i) // G_i is the typhoon grade determined by wind speed V_i

    // Integrity repair based on neighbor values for wind speed, moving direction, moving sp
    if V_i == null
        V_i = (V_{i-1} + V_{i+1})/2, for i = 2,3,...,n-1
        V_i = (V_i + V_{i+1})/2, for i = 2,3,...,n-1, V_i < V_{i+1}
        V_i = V_{i-1}, for i = 1

    if MD_i == null
        MD_i = (MD_{i-1} + MD_{i+1})/2, for i = 2,3,...,n-1
        MD_i = MD_{i-1}, for i = 1

    if MV_i == null
        MV_i = (MV_{i-1} + MV_{i+1})/2, for i = 2,3,...,n-1
        MV_i = MV_{i-1}, for i = 1

    if P_i == null
        P_i = (P_{i-1} + P_{i+1})/2, for i = 2,3,...,n-1
        P_i = P_{i-1}, for i = 1

    // Integrity repair based on wind-pressure relationship
    if P_i == null && V_i != null
        P_i = 1010 - (V_i/6.7)^1.644

    if V_i == null && P_i != null
        V_i = 6.7 * (1010 - P_i)^0.644

    // Integrity repair for moving direction and speed based on latitude and longitude
    if MV_i == null && MD_i == null
```

```
        tan  = (Lo_2 - Lo_1)/(La_2 - La_1)
        // Calculate angle based on latitude and longitude
        if Lo > 0 && La < 0
            angle' = (90 - angle) + 90
        if Lo <= 0 && La < 0
            angle' = 180 + angle
        if Lo <= 0 && La < 0
            angle' = (90 - angle) + 270

    // Consistency repair based on factor dependency
    if G_i != V_i
        G_i = f(V_i) // G_i is the grade corresponding to wind speed V_i
end for
```

2.3 Case Analysis

This paper selects Typhoon "201525" data as reference data, numbered 1, and Typhoon "201526" and Typhoon "201527" data, numbered 2 and 3 respectively. Partial original data of Typhoon "201525" is shown in [Figure 1: see original paper].

[Figure 1: see original paper] "201525" Typhoon Partial Original Data

According to the method proposed in Section 2.2, the repair process is as follows (using Typhoon "201525" as an example):

Through examination, the factors requiring repair are moving direction and moving speed. The specific methods are:

a) **Moving Direction Repair**: The moving direction in the original typhoon data is recorded using 16 wind directions. For subsequent calculation convenience, numerical representation is needed: north is defined as 0, south as 8, clockwise from north to south as 1-7, and counterclockwise from north to south as 9-15.

The angle is calculated based on latitude and longitude, and the moving direction is supplemented by comparing with the 16-wind-direction diagram and numerical conversion relationship. For example, at time "2015-10-14T14:00:00" , the latitude and longitude are 156.9 and 15.0 respectively, while the next moment' s latitude and longitude are 154.4 and 15.8. Using the method in Section 2.2, the calculated angle is 288.3294336171956, which corresponds to west-northwest on the 16-wind-direction diagram, with a numerical result of 11.

b) **Moving Speed Repair**: Moving speed is calculated based on latitude, longitude, and time interval. For example, at time "2015-10-14T14:00:00" , the latitude and longitude are 156.9 and 15.0, while at the next moment "2015-10-15T02:00:00" , the latitude and longitude are 154.4 and 15.8. The calculated moving speed is rounded to 28.

The data after processing using the proposed repair method is shown in [Figure 2: see original paper].

[Figure 2: see original paper] Repaired Typhoon Data

In [Figure 2: see original paper], each element' s subscript indicates its position in the entire time series, with the last subscript value representing the length of the entire time series.

After verification, the Root Mean Square Error (RMSE) value for moving direction is 0.187867287325544, and the RMSE value for moving speed is 0.264797369811, indicating high accuracy of the repair method.

---

3 Similarity Measurement of Typhoon Time Series Based on Weighted DTW Distance

Typhoons exhibit spatial, seasonal, and multi-factor characteristics. This paper comprehensively considers five factors: moving direction, moving speed, pressure, latitude, and longitude. To address the different importance of each factor and the unequal length characteristics of typhoon time series, weighted DTW distance is adopted for similarity measurement.

Similarity measurement mainly includes the following steps: seasonal similarity judgment, typhoon factor weight design, and weighted DTW distance calculation. Seasonal similarity is easy to implement and will not be elaborated here. This chapter focuses on typhoon factor weight calculation and weighted DTW distance calculation.

3.1 Typhoon Factor Weight Design

This paper uses Principal Component Analysis (PCA) to determine five factors describing typhoons: moving direction, moving speed, pressure, latitude, and longitude. Their weights are denoted as W_{MD}, W_{MV}, W_P, W_{La}, and W_{Lo} respectively. The weight calculation method uses PCA to compute the correlation coefficient matrix and its eigenvectors and eigenvalues. Based on a set threshold of principal component contribution rate, the number of principal components is determined, and the weight of each typhoon factor is calculated according to the coefficients of each factor in the principal components. The calculation process is as follows:

a) Standardize the original typhoon data

b) Calculate the correlation coefficient matrix

c) Calculate eigenvectors and eigenvalues of the correlation coefficient matrix

The eigenvector matrix B can be expressed as:

$$B = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix}$$

d) Select principal components and calculate their contribution rates

The principal components are arranged in descending order of eigenvalues: $\lambda_1, \lambda_2, ..., \lambda_n$. The principal components are expressed as:

$$\begin{cases} C_1 = a_{11}MD + a_{12}MV + a_{13}Lo + a_{14}La + a_{15}P \\ C_2 = a_{21}MD + a_{22}MV + a_{23}Lo + a_{24}La + a_{25}P \\ ... \\ C_n = a_{n1}MD + a_{n2}MV + a_{n3}Lo + a_{n4}La + a_{n5}P \end{cases}$$

where $n \leq 5$.

e) Calculate the weight of each typhoon factor

Assuming there are n principal components ($n \leq 5$), the weights of moving direction, moving speed, pressure, latitude, and longitude are calculated as:

$$\begin{cases} W_{MD} = \frac{C_1}{\lambda_1+\lambda_2+...+\lambda_n} \times a_{11} + \frac{C_2}{\lambda_1+\lambda_2+...+\lambda_n} \times a_{21} + ... + \frac{C_n}{\lambda_1+\lambda_2+...+\lambda_n} \times a_{n1} \\ W_{MV} = \frac{C_1}{\lambda_1+\lambda_2+...+\lambda_n} \times a_{12} + \frac{C_2}{\lambda_1+\lambda_2+...+\lambda_n} \times a_{22} + ... + \frac{C_n}{\lambda_1+\lambda_2+...+\lambda_n} \times a_{n2} \\ ... \\ W_{Lo} = \frac{C_1}{\lambda_1+\lambda_2+...+\lambda_n} \times a_{15} + \frac{C_2}{\lambda_1+\lambda_2+...+\lambda_n} \times a_{25} + ... + \frac{C_n}{\lambda_1+\lambda_2+...+\lambda_n} \times a_{n5} \end{cases}$$

3.2 Weighted DTW Distance Calculation

For different typhoons, each factor has a different impact on similarity measurement. This paper uses weighted DTW distance for similarity measurement. The specific definition is as follows:

Let time series $X = \{X_1, X_2, ..., X_m\}$ and $Y = \{Y_1, Y_2, ..., Y_n\}$, where $X_i = \{S_{MD}(i), S_{MV}(i), S_P(i), S_{La}(i), S_{Lo}(i)\}$ and $Y_j = \{S'_{MD}(j), S'_{MV}(j), S'_P(j), S'_{La}(j), S'_{Lo}(j)\}$. The weighted DTW distance between X and Y is defined in Equation (1):

$$M(i,j) = \begin{cases} 0, & i = 0, j = 0 \\ \infty, & i = 0, j \neq 0 \text{ or } i \neq 0, j = 0 \\ \min\{M(i-1,j-1), M(i-1,j), M(i,j-1)\} + d(x_i, y_j), & \text{otherwise} \end{cases} \tag{1}$$

where $d(x_i, y_j)$ represents the base distance between x_i and y_j, which can be calculated using different distance metrics depending on the situation. This paper uses the following formula:

$$d(x_i, y_j) = W_{MD} \times |MD_i - MD_j| + W_{MV} \times |MV_i - MV_j| + W_P \times |P_i - P_j| + W_{La} \times |La_i - La_j| + W_{Lo} \times |Lo_i - Lo_j|$$

The algorithm for multivariate typhoon time series similarity measurement is as follows:

```
Input: Historical typhoon data arr2 and reference typhoon data arr1
Output: Distance d between historical typhoon and reference typhoon

def Distance(i, j, arr1, arr2, W):
    if i == 0 and j == 0:
        return 0
    elif i == 0 or j == 0:
        return float("inf")
    else:
        data1 = arr1[i]
        data2 = arr2[j]
        result = 0
        for i in range(len(arr1[0])):
            d = d + pow(data1[i] - data2[i], 2) * W[i]
        result = pow(result, 0.5)
        return result + min(Distance(i-1, j-1, arr1, arr2, W),
                            Distance(i-1, j, arr1, arr2, W),
                            Distance(i, j-1, arr1, arr2, W))
```

Algorithm Analysis: Let the length of reference typhoon data be M and the length of historical typhoon data be N. The time complexity of this algorithm is $O(MN)$.

3.3 Case Analysis

This section uses the data processed in Section 2.3. The specific calculation process is as follows (using data numbered 1 and 2 as examples):

First, calculate the weight of each factor for Typhoon 1. This paper uses the PCA method, and the results are shown in .

Principal Components, Contribution Rates, and Weight Results

Then calculate the distance between Typhoon 1 and Typhoon 2. Substituting the result into Equation (2) yields 578.4718754376307. Similarly, the distance between Typhoon 1 and Typhoon 3 is calculated as 707.0062487978438. Since the distance between Typhoon 1 and Typhoon 2 is smaller, these two typhoons are more similar.

4 Experiments

4.1 Experimental Environment and Data

Experimental Environment: MATLAB R2013a, IntelliJ IDEA 2016.3.4, Windows 7 SP1, 1 TB hard disk, 8 GB RAM, Intel(R) CoreTM i7-3770.

Experimental Data: Typhoon data from 1945-2016 is selected as the research object. The typhoon data includes values for factors such as Chinese and English typhoon names, occurrence time, end time, latitude and longitude, intensity, wind speed, moving direction, moving speed, and pressure.

4.2 Experimental Results and Analysis

Similar typhoons refer to two typhoons with similar seasons that satisfy the weighted DTW distance requirement. A typhoon data record is randomly selected as reference data, numbered 1, and collected historical typhoon data are numbered 2, 3, ···, N (N represents the number of historical typhoon data records). If records numbered 2, 3, ···, N all satisfy seasonal similarity, their distances to the reference data are calculated and denoted as $d_{1,2}, d_{1,3}, ..., d_{1,N}$. The smaller the distance, the more similar the two typhoons. This paper designs three experiments: Experiment 1 on typhoon time series quality assessment and repair; Experiment 2 on effectiveness validation of the multivariate typhoon time series similarity measurement method; and Experiment 3 on practicality proof of the multivariate typhoon time series similarity measurement method.

Experiment 1: Typhoon Time Series Quality Assessment and Repair
This experiment consists of two groups. The first group assesses and repairs the integrity of experimental data, and the second group validates the effectiveness of the proposed typhoon time series repair method.

1) Typhoon Time Series Integrity Assessment and Repair

(1) Integrity Threshold Determination

Experimental data from Typhoons "195526", "200008", "200917", and "201117"are used. Using the proposed method for integrity assessment, Typhoon "200008" (No. 3) has an integrity of 0.6, and Typhoon "200917" (No. 4) has an integrity of approximately 0.8. Typhoon "195526" (No. 2) data is modified to have an integrity of 0.4. The DTW distances between these three typhoons and Typhoon "201117" (No. 1) are calculated, as shown in .

Distance Considering Path Similarity ("201117")

Typhoon	Weighted DTW Distance
195526	

Typhoon	Weighted DTW Distance
200008	
200917	

The results show that the similarity ranking with Typhoon "201117" from high to low is "195526", "200008", and "200917". However, it is known that Typhoon "200008" is more similar to Typhoon "201117" , indicating an error. That is, when integrity is below 0.6 during data supplementation, the repaired data has a high error rate. Therefore, in this paper, only when integrity is greater than or equal to 0.6 can repaired data be used for similarity measurement with higher accuracy.

(2) Typhoon Time Series Integrity Assessment

This experiment selects data needed for the following two experiments: Typhoons "201117" , "200008" , "195526" , "199327" , "200917" , "201323" , "197010" , "201312" , "200713" , "200414" , "195615" , "199112" , "195316" , "198510" , and "199111" (15 typhoon data records). These data mainly have integrity issues. The integrity comparison before and after repair is shown in [Figure 3: see original paper].

[Figure 3: see original paper] Integrity Comparison of Typhoon Time Series Before and After Repair

The figure shows that after repair, the integrity of typhoon time series reaches 100%, facilitating subsequent typhoon similarity measurement and improving the accuracy of similarity measurement.

2) Validation of Typhoon Time Series Repair Method Accuracy

Experimental data from Typhoons "200713" , "200817" , "200917" , "201011" , "201117" , "201215" , "201312" , "201323" , "201416" , and "201521" (10 typhoon data records) are used.

(1) Accuracy Validation of Integrity Repair Method

According to literature [26], repair accuracy can be measured using:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(e_i - \hat{e}_i)^2}$$

where e_i is the original true value of the typhoon factor, and $\hat{e}_i$ is the repaired value. A smaller $RMSE$ indicates higher accuracy.

10% of the originally complete data is randomly selected and replaced with null values. The accuracy of the neighbor value repair method is shown in [Figure 4: see original paper].

[Figure 4: see original paper] Accuracy of Neighbor Value Repair Method

10% of the originally complete data is randomly selected, with at least two adjacent null values in each group. Pressure and wind speed in the selected data are replaced with null values and repaired using the wind-pressure relationship method. The repair accuracy is shown in [Figure 5: see original paper].

[Figure 5: see original paper] Accuracy of Wind-Pressure Repair Method

10% of the originally complete data is randomly selected, with at least two adjacent null values in each group. Moving direction and speed in the selected data are replaced with null values and repaired using the latitude-longitude calculation method. The repair accuracy is shown in [Figure 6: see original paper].

[Figure 6: see original paper] Accuracy of Latitude-Longitude Calculation Repair Method

(3) Accuracy Validation of Consistency Repair Method

10% of the originally correct data is randomly selected, and the grade data is modified to be inconsistent with wind speed. The accuracy of the consistency repair method is shown in [Figure 7: see original paper].

[Figure 7: see original paper] Accuracy of Consistency Repair Method

[Figure 4: see original paper] shows that the RMSE value for each factor is below 0.8. [Figure 5: see original paper] shows that the RMSE values for wind speed and pressure are both less than 4. [Figure 6: see original paper] shows that the RMSE values for moving direction and speed are both less than 0.7. Therefore, the proposed integrity repair method demonstrates high repair accuracy.

Experiment 2: Effectiveness Validation of Multivariate Typhoon Time Series Similarity Measurement Method This experiment consists of two groups. The first group uses Typhoon "201117" data and its similar path typhoon data, and the second group uses Typhoon "201323" data and its similar path typhoon data. Using the proposed method and considering only path similarity (moving direction and speed factors), the weighted DTW distance is calculated.

1) Typhoons with Similar Path to "201117"

This experiment sets the contribution rate threshold at 90%. The weight calculation results are shown in the first two columns of .

Principal Components, Contribution Rates, and Weight Results for "201117" and "201323"

Principal Component	Weight (“201117”)	Contribution Rate (%)	Weight (“201323”)	Contribution Rate (%)

According to literature [26], typhoons similar to “201117” in descending order of similarity are “200008” , “195526” , “199327” , and “200917” , numbered 2, 3, 4, and 5 respectively. Considering only path similarity, shows that the weights for moving direction and speed are 0.0099 and 0.9901 respectively. The distances between Typhoon 2 and Typhoon 1, Typhoon 3 and Typhoon 1, Typhoon 4 and Typhoon 1, and Typhoon 5 and Typhoon 1 are calculated, with results shown in .

Distance Considering Path Similarity (“201117”)

Typhoon Pair	Weighted DTW Distance
1-2	
1-3	
1-4	
1-5	

The calculation results show that the distances sorted from low to high are 1-2, 1-3, 1-4, and 1-5. That is, typhoons similar to “201117” in descending order of similarity are “200008” , “195526” , “199327” , and “200917” , matching the results given in literature [26].

2) Typhoons with Similar Path to “201323”

Typhoon “201323” data is selected as reference data, numbered 1. According to literature [20], typhoons similar to “201323” in descending order of similarity are “197010” , “201312” , and “200713” , numbered 2, 3, and 4 respectively. Considering only path similarity, the last two columns of show that the weights for moving direction and speed are 0.1881 and 0.8119 respectively. The weighted DTW distance calculation results are shown in .

Distance Considering Path Similarity (“201323”)

Typhoon Pair	Weighted DTW Distance
1-2	
1-3	
1-4	

The calculation results show that the distances sorted from low to high are 1-2, 1-3, and 1-4. That is, typhoons similar to “201323” in descending order of similarity are “197010” , “201312” , and “200713” , matching the results given in literature [20].

The above two experiments demonstrate that this method can effectively judge typhoon path similarity.

Experiment 3: Practicality Validation of Multivariate Typhoon Time Series Similarity Measurement Method Typhoon “200414” data is selected as reference data, numbered 1. This experiment sets the contribution rate threshold at 90%. Using the weight calculation method in Section 2.1, the weights for moving direction and speed are calculated as 0.0315 and 0.9685 respectively. If only path similarity is considered, the 5 similar typhoons obtained in descending order of similarity are “195615” , “199112” , “195316” , “198510” , and “199111” , numbered 2, 3, 4, 5, and 6 respectively. The weighted DTW distances are shown in .

Distance Considering Path Similarity (“200414”)

Typhoon Pair	Weighted DTW Distance
1-2	
1-3	
1-4	
1-5	
1-6	

In addition to path similarity and considering typhoon spatial characteristics and disaster impact, this paper also comprehensively considers pressure, latitude, and longitude to enable more comprehensive similarity measurement, thereby effectively predicting disaster severity and preparing protective measures in relevant areas in advance. The weights for moving direction, speed, pressure, latitude, and longitude are 0.0008, 0.0052, 0.9926, 0.0012, and 0.0002 respectively. The weighted DTW distance results are shown in .

Distance Considering Multiple Factors

Typhoon Pair	Weighted DTW Distance
1-2	
1-3	
1-4	
1-5	
1-6	

The above table shows that the 5 typhoons similar to “200414” in descending order of similarity are “198510” , “199111” , “199112” , “195316” , and “195615” .

The most path-similar typhoon to "200414" is "195615" . However, as shown in the typhoon paths and intensity information from the China Typhoon Network ([Figure 8: see original paper]), the intensity change of Typhoon "195615" is least similar to "200414" , while the intensity change of Typhoon "198510" is most similar to "200414" .

[Figure 8: see original paper] Typhoon Information

According to literature [27], the damage caused by Typhoon "200414" (Yunna) in Zhejiang mainly manifested in three aspects: (a) Strong winds caused a large number of houses to collapse, accounting for 66.5% of total deaths; people blown down accounted for 5.5%; and deaths from falling power poles or electrocution accounted for 3%. (b) Due to strong winds and low pressure, seawater was strongly piled up toward the coast, causing a sudden rise in tide levels and damaging 4,059 coastal embankments (563 km) with 1,222 breaches (88 km). (c) Heavy precipitation from the typhoon trapped 440,000 people in floods and triggered strong geological disasters. Literature [28] points out that Typhoon "198510" caused continuous heavy rain in Guangxi for a week. Literature [29] indicates that Typhoon "198510" triggered mudslides that damaged water conservancy projects. Evidently, the disaster severity brought by Typhoon "198510" is similar to that of Typhoon "200414" .

4.3 Performance Comparison

4.3.1 Time Performance Comparison Ten typhoon data records ("194930" , "195820" , "196911" , "197613" , "198510" , "199215" , "200414" , "201117" , "201323", and "201521") are selected for runtime comparison. Using the proposed similarity measurement method, the time required for similarity measurement is compared between considering only moving direction and speed (2 factors) and additionally considering pressure, latitude, and longitude (5 factors). The results are shown in [Figure 9: see original paper].

[Figure 9: see original paper] Time Performance Comparison

[Figure 9: see original paper] shows that after adding pressure, latitude, and longitude, the runtime increases, but the increase is very small. The experiment demonstrates that the proposed method considers more comprehensive factors while maintaining almost unchanged efficiency.

4.3.2 Parameter Performance Comparison This paper compares several similar typhoon judgment methods in terms of considered factors, correctness, and efficiency. The results are shown in , demonstrating that the proposed method is superior to other methods in all aspects.

Parameter Performance Comparison

Method	Considered Factors	Correctness	Efficiency
WebGIS-based analysis system [19]			
Similarity divergence principle [21]			
ArcGIS-based screening method [24]			

5 Conclusion

This paper first combines typhoon characteristics to formally define typhoon data using multivariate time series, achieving quantitative description of data with spatial, seasonal, and multi-factor characteristics. It then assesses and repairs integrity and consistency issues in typhoon data. Based on this foundation, a multivariate typhoon time series similarity measurement method based on weighted DTW distance is proposed. Experimental results show that the proposed typhoon data quality assessment method can improve the accuracy of similarity measurement and disaster prediction. The similarity measurement method can set different weights according to different reference typhoons to suit various similar typhoon judgments and can effectively determine typhoon-induced disasters.

Future work will focus on real-time typhoon trend and intensity prediction based on historical typhoon data to provide disaster prevention, mitigation, and auxiliary decision-making support for relevant departments.

References

[1] China Meteorological Administration. 2016 China Climate Bulletin. [EB/OL]. [2017-09-04]. http://m.cma.gov.cn/root7/auto13139/201705/t20170525_415102.html.

[2] Li Jianzhong, Wang Hongzhi, Gao Hong. Research Progress on Big Data Availability [J]. Journal of Software, 2016, 27(7): 1605-1625.

[3] Liu Yongnan, Zou Zhaonian, Li Jianzhong, et al. Data Integrity Assessment Method [J]. Journal of Computer Research and Development, 2013, 50(S1): 230-238.

[4] Miao Dongjing. Research on Computational Complexity Theory and Algorithms for Data Consistency [D]. Harbin: Harbin Institute of Technology, 2016.

[5] Miao Dongjing, Liu Xianmin, Li Jianzhong. Approximate Functional Dependency Mining Algorithm in Probabilistic Databases [J]. Journal of Computer Research and Development, 2015, 52(12): 2857-2865.

[6] Liu Xianmin, Li Jianzhong. An Extended Conditional Functional Dependency Discovery Algorithm [J]. Journal of Computer Research and Development, 2015, 52(1): 130-140.

[7] Huang Dongmei, Chen Kuo, Wang Zhenhua, et al. Residual Optimization Selection Algorithm in Marine Data Quality Inspection Scheme [J]. Computer and Digital Engineering, 2014, 42(10): 1752-1757.

[8] Wang Ruihan, Huang Dongmei, Wang Zhenhua, et al. A Spatial Sampling Method for Marine Data [J]. Computer Applications and Software, 2015, 32(5): 228-230+245.

[9] Wang Zhenhua, Zhou Xuenan, Huang Dongmei. Research on Quality Sampling Inspection Model for Uncertain Marine Data [J]. Computer Science, 2015, 42(2): 182-184+190.

[10] Zhu Tian, Bai Sixue. Time Series Similarity Search Based on Pattern Distance Measurement [J]. Microcomputer Information, 2007, 23(30): 216-217.

[11] Tan P N, Steinbach M, Kumar V, et al. Wikipedia [EB/OL]. [2017-09-04]. http://zh.wikipedia.org/wiki, 2013.

[12] Agrawal R, Faloutsos C, Swami A. Efficient Similarity Search in Sequence Databases [C]//Proc of International Conference on Foundations of Data Organization and Algorithms. Berlin: Springer, 1993: 69-84.

[13] Keogh E, Pazzani M. Derivative Dynamic Time Warping [C]//Proc of the 1st SIAM International Conference on Data Mining: ResearchGate. 2001: 1-11.

[14] Berndt D J, Clifford J. Using Dynamic Time Warping to Find Patterns in Time Series [C]//Proc of International Conference on Knowledge Discovery and Data Mining. [S.l.]: AAAI Press, 1994: 359-370.

[15] Tu Hui, Liu Li, Zhang Zhengjin. ECG Signal Similarity Measurement Using Improved DTW Algorithm [J]. Computer Engineering and Applications, 2015, 51(16): 215-218.

[16] Yang Yanlin, Ye Feng, Lü Xin, et al. A DTW Clustering-Based Similarity Mining Method for Hydrological Time Series [J]. Computer Science, 2016, 43(2): 245-249.

[17] Feng Jun, Chen Huanlin, Tang Zhixian, et al. A Novel DTW-Based Similarity Measurement Method for Stock Market Time Series [J]. Data Acquisition and Processing, 2015, 30(1): 99-105.

[18] Reyes M, Dominguez G, Escalera S. Feature Weighting in Dynamic Time Warping for Gesture Recognition in Depth Data [C]//Proc of IEEE International Conference on Computer Vision. 2012: 1182-1188.

[19] Kong Lingna. Research on Typhoon Path Similarity Prediction Based on GIS [D]. Changsha: Central South University, 2012.

[20] She Liangliang, He Liang, Miao Nengbin. WebGIS-Based Analysis of Historical Typhoon Similar Paths in Ningbo [J]. Zhejiang Water Conservancy Science and Technology, 2014, 42(1): 24-26.

[21] Liu Yong, Wu Biwen, Wang Dongyong. Research on a Typhoon Path Similarity Retrieval Algorithm [J]. Meteorology, 2006, 32(7): 18-24.

[22] Ding Xiaoou, Wang Hongzhi, Zhang Xiaoying, et al. Research on Correlation Among Multiple Data Quality Properties [J]. Journal of Software, 2016, 27(7): 1626-1644.

[23] Zou Yan, Zhao Ping, Qiao Lin. Reconstruction of Typhoon Wind-Pressure Formula Based on Typhoon Yearbook Data [J]. Journal of Tropical Meteorology, 2009, 25(2): 163-168.

[24] Wang Xina, Huang Huabing, Ban Ya, et al. Typhoon Similar Path Screening and Prediction Using GIS Spatial Analysis [J]. Bulletin of Surveying and Mapping, 2014(5): 115-118.

[25] Zhang Shichao, Jin Zhi, Zhu Xiaofeng. Missing Data Imputation by Utilizing Information Within Incomplete Instances [J]. Journal of Systems & Software, 2011, 84(3): 452-459.

[26] Zhang Guofeng, Zhang Jinghong, Tian Guanghui, et al. Similar Typhoon Screening in Typhoon Disaster Assessment [J]. Hubei Agricultural Sciences, 2012, 51(7): 1334-1337.

[27] Xue Genyuan, Yu Shanxian, He Fengpian, et al. Characteristics of Yunna Typhoon Disaster and Preliminary Study on Typhoon Disasters in Zhejiang Province [J]. Journal of Natural Disasters, 2006, 15(4): 39-47.

[28] Wu Xingguo. Characteristics of Typhoons in Guangxi Since the 1980s [J]. Guangxi Meteorology, 1988, 9(3): 6-10.

[29] Liang Biqi, Liang Jingping. Characteristics of Typhoon Disasters in Guangdong and Their Impact on Economic Development [J]. China Disaster Reduction, 1993, 3(3): 35-37, 34.

Note: Figure translations are in progress. See original paper for figures.