The Rainfall and Meteorological Data Mining Model Based on Multi-Dimensional Precipitation Time Series

Sizhe Ding*

1 School of Mathematics and Statistics, Wuhan University, Wuhan City, Hubei Province, 430072, China

*Corresponding author’s e-mail: dingsizhe@whu.edu.cn

Abstract. In order to study the relationship between rainfall and related meteorological elements in rainy weather, find out the changes of related meteorological elements before and after rainfall, a multi-dimensional time series data mining model is proposed. The model first performs dimension selection pre-processing on the time series of meteorological elements to remove irrelevant and redundant dimensions, then uses the proposed extreme slope piecewise linear fitting method to segment the time series, data compression and eigenvalue extraction, and finally uses k-means clustering algorithm to symbolize the processed multi-dimensional sequence, and uses rules to extract the rainfall weather model. Experimental results show that the model has good practical value.

1. Introduction

With the rapid development of global meteorological science, more and more attention has been paid to the accurate and interpretable intelligent weather services. Among them, as an important decision support service, meteorological forecast service can provide annual, monthly, daily and even hourly precipitation forecast to water resources management departments, agricultural departments and other decision-making departments by way of forecasting the basic precipitation in the future. It is an indispensable and important part of national economic and social development[1].

Weather forecasting is based on meteorological observation data, applying the principles and methods of meteorology, dynamic meteorology, statistics, etc., and referring to the climate background and weather evolution rules of a certain area, to make a qualitative analysis of the weather conditions in the future for a certain period or quantitative predictions. At present, the weather forecast mainly uses four forecast processing methods: the first is the empirical forecast method, based on the weather map situation forecast, based on the future position and intensity of the weather system, to predict the future weather distribution; the second is statistical forecasting methods, by counting the probability of a phenomenon appearing under specific environmental conditions in history, to speculate the possibility of occurrence in a similar environment in the future; the third is the numerical forecasting method, using the atmospheric motion equations in a certain integrate system of equations under the conditions of the initial and boundary values to predict the future weather; and there is also an integrated forecasting method, that is, a combination of multiple forecasting results of different forecasting methods on the same element, resulting in a superior Forecast results of a single forecast method.

At present, scholars from all walks of life have made some achievements in the prediction of annual precipitation, and there are many existing methods. Among them, the method based on mathematical statistics is still the most widely used, that is, by assuming the distribution law of
precipitation in each year, then fitting the data of the later years, and finally obtaining the basic model equation, and forecasting the precipitation in the future years. Among them, in addition to directly using the distribution function for fitting, some scholars have proposed the Markov chain prediction method based on the distribution probability in recent years. For example, Wang Wensheng et al. [2] proposed to use Markov chain to predict annual precipitation on the basis of one-step transfer probability; Xu Xiaoyu et al. [3] further proposed weighted Markov Chain considering the dependence of multi-year precipitation series. The chain prediction method has successfully predicted the annual precipitation in Beijing and Lujiang County.

Another direct prediction method based on astronomical period is different from the mathematical statistics method. Because this method does not depend on the specific precipitation distribution, it can eliminate the interference of many precipitation factors, and only uses historical data to predict and evaluate. At present, it has attracted extensive attention in academic circles. This method is feasible in principle by using the basic atmospheric circulation model and the change period of meteorological elements in a wide area to correspond with the astronomical period, and to infer that the surface precipitation is directly or indirectly affected by it. For example, The Comprehensive Research Group of Guangshen Tiandi Sheng et al. [4] used the weather cycle as the time curve of meteorological elements of a single station in Guangdong Province, and successfully predicted the annual precipitation of the region according to the relationship between weather system and weather process.

However, the above two methods have irreparable defects. First of all, the distribution function is regarded as the most important dependent information based on mathematical statistics, and precipitation in most areas has certain chaotic characteristics, so it is impossible to fit some sudden change periods or potential dynamic factors only by using distribution function. Therefore, the prediction based on mathematical statistics usually has a good effect in the year when the precipitation changes gently, but the fitting error will increase rapidly in other years. Pessimistically, even if we introduce Markov model to randomize it, the dynamic change of climate is likely to have continuous effect, which makes the Markov model not have robust analysis property, and can only ensure that the prediction accuracy can be controlled in a short time, and its time granularity cannot be further reduced to arbitrary accuracy.

Although the direct prediction method based on astronomical period overcomes the dependence on distribution function, its dependence on astronomical period makes it fall into another dilemma. The prediction based on astronomical period needs to assume that the climate changes in an intrinsic astronomical period, which cannot overcome the influence of extreme conditions and outliers. In addition, in recent years, abnormal climate patterns occur frequently, which will further challenge the constraints of traditional astronomical cycles on climate. In addition, the precipitation model based on long-term precipitation is only suitable for the long-term prediction of astronomical period.

In Summary, meteorological data has the characteristics of spatio-temporal attributes, multi-dimensional, multi-scale, non-stationary, uncertain, strong periodicity, high attribute correlation, etc. It is difficult to analyse and process meteorological data using traditional methods. Applying data mining methods to the analysis and processing of data in the meteorological field, exploring various meteorological elements and their internal connections with weather phenomena, looking for various potential laws to reveal unknown meteorological theories It is important and can have a positive and important impact in enriching weather forecasting methods and improving weather forecasting. At present, more commonly used forecasting techniques based on data mining methods include artificial neural networks, genetic algorithms, support vector machines, Bayes, decision trees, and association rule mining [5]. If we want to focus on overcoming the shortcomings of traditional methods, we must start to develop a comprehensive prediction scheme with independent mode and data driven. The multi-dimensional precipitation time series mining model proposed in this paper follows this idea.

In view of the above problems, this paper proposes a new multi-dimensional time series data mining model. This model is based on the study of a large number of meteorological data, from the perspective of data analysis to mine the relationship between various meteorological elements and
rainfall and establish a rainfall weather model. The analysis proves that the model has scientific significance, and a large number of experimental results prove that the model can provide a valuable method for meteorological data research, and can promote the further development of time series data mining research [6].

The purpose of this paper is to reveal the unique analysis method and idea of this method in the independent prediction of precipitation model analysis, and does not deliberately pursue the correction of prediction accuracy. Of course, the accuracy correction based on this method can be supported by any data-driven time series mining method precision correction technology, and these methods have been well known by the academic community. In conclusion, this paper focuses on the model analysis and prediction of annual precipitation, which provides the most important macro decision support.

2. Time Series Prediction Model

Through the analysis of the elements in the mean-generation function matrix, it can be seen that the mean-generation function matrix also contains all the possibilities that there may be at least two cycles in the known sequence, and the average value of the elements in the same position in each cycle, To eliminate the accidental nature of individual data. This is the meaning of averaging in the average generating function. After obtaining the mean generating function matrix, in order to predict the known sequence, the period extension matrix $F$ needs to be generated from the mean generating function matrix. Assuming that the known sequence length is $N$ and the required prediction time point is $q$, then $F$ is expressed as follows:

$$F = \begin{bmatrix} F_{[N/2]} \\ F_q \end{bmatrix}$$  \hspace{1cm} (1)$$

and among

$$F_{[N/2]} = \begin{bmatrix} \bar{x}_1 (1) & \bar{x}_2 (1) & \ldots & \bar{x}_{[N/2]} (1) \\ \bar{x}_1 (1) & \bar{x}_2 (2) & \ldots & \bar{x}_{[N/2]} (2) \\ \bar{x}_1 (1) & \bar{x}_2 (1) & \ldots & \bar{x}_{[N/2]} (3) \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1 (1) & \bar{x}_2 (i_1) & \ldots & \bar{x}_{[N/2]} \left( \frac{N}{2} \right) \\ \end{bmatrix}$$  \hspace{1cm} (2)$$

and

$$F_q = \begin{bmatrix} \bar{x}_1 (1) & \bar{x}_2 (i_1) & \ldots & \bar{x}_{[N/2]} (1) \\ \bar{x}_1 (1) & \bar{x}_2 (i_2) & \ldots & \bar{x}_{[N/2]} (2) \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1 (1) & \bar{x}_2 (i) & \ldots & \bar{x}_{[N/2]} (q) \\ \end{bmatrix}.$$  \hspace{1cm} (3)$$

$\bar{x}_2 (i)$ means taking one of $\bar{x}_2 (1)$ and $\bar{x}_2 (2)$, and $\bar{x}_2 (i)$ means the other. It can be seen that the periodic epitaxial matrix is essentially to expand the mean generating function matrix according to the length of each column vector period to make periodic expansion to fill zero elements. There are two reasons for this: on the one hand, the upper limit of the number of rows of the mean generating function matrix is extended, so that the matrix $F$ has the function of fitting and prediction; on the other hand, the sparse matrix $H$ is transformed into a full matrix $F$, and then the sequence simulation Each period law (that is, the matrix contains at least two periods to no period law) can be reflected in the sum prediction. $F_{[N/2]}$ and $F_q$ in the periodic epitaxial matrix $F$ are used for fitting and
prediction, respectively. The orthogonalized periodic extension matrix $F^*$ is used for fitting and prediction [7]. The fitting and prediction formulas are as follows,

$$x(t) = \sum_{i=2}^{\lfloor N/2 \rfloor} \phi_i f_i(t), t, N \quad (4)$$

$$x(N+q) = \sum_{i=2}^{\lfloor N/2 \rfloor} \phi_i f_i(N+q) \quad (5)$$

According to Equation (4), the vector $i (i=2, \ldots, \lfloor N/2 \rfloor)$ is obtained by the least square method,

$$\Phi = \left(F^T F^*\right)^{-1} F^T X' \quad (6)$$

Among them, $X'$ is a new sequence after normalizing the known data sequence $X$:

$$x'(t) = \frac{x(t) - \bar{x}}{\sigma} \quad (7)$$

In the formula, $\bar{x}$ and $\sigma$ are the average and standard deviation of the known sequence $X$, respectively. The reason for standardizing the known data series here is to make the known series fluctuate around the zero value and be at the same numerical level.

Finally, let's explain the essential principle of periodic epitaxial matrix fitting and prediction. It can be seen from Equations (4) and (5) that the use of periodic epitaxy matrix fitting and prediction actually separates any periodicity that may exist in the known sequence, and then respectively gives corresponding weights. For example, the vector in the first column of the periodic extension matrix indicates that each element in the known data sequence has an independent period. For the predicted value of the unknown sequence, the contribution of this period is the average value of all elements in the known data sequence multiplied by this periodic weight; and the last column of vectors has the largest period, and the vector contains only two periods. When predicting the nth value, the vector's contribution is each element in the first few periods of the same period as the predicted value. average value. The meaning of the remaining column vectors can be deduced by analogy. Therefore, it can be seen that the method of predicting with the mean-generation function model includes the possibility of a variety of periodic changes in the known sequence, and then the coefficient vector is used as a trade-off between the possibility of different periods. This method is reasonable and feasible [8].

3. Examples of Rainfall Prediction

3.1. Rainfall Data

The simulation data in this paper comes from the measured rainfall from 1970 to 1989 in a certain area. The data is shown in Table 1. The rainfall data is divided into two parts. The rainfall from 1970 to 1989 is used as a sample of mean generating function for simulation fitting, and the rainfall from 2001 to 2019 is predicted. According to the prediction method based on the average generating function described above, the above algorithm is implemented in MATLAB language [9].

| Year | 1970   | 1971   | 1972   | 1973   | 1974   | 1975   | 1976   | 1977   | 1978   |
|------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| V01301| 57251  | 57251  | 57251  | 57251  | 57251  | 57251  | 57251  | 57251  | 57251  |
| V04001| 919.9  | 566.4  | 783.5  | 1037.5 | 842.2  | 1070.9 | 772.8  | 713.8  | 788.3  |
| V13305| 130    | 106    | 105    | 127    | 116    | 101    | 108    | 110    | 114    |
| V13353| 90.9   | 63.8   | 75     | 63.8   | 73     | 112.9  | 71.6   | 45.5   | 44.4   |
| V13052| 57251  | 57251  | 57251  | 57251  | 57251  | 57251  | 57251  | 57251  | 57251  |
3.2. Results verification

The extreme value slope piecewise linear fitting algorithm is used to realize sequence segmentation, data compression and extract the meteorological feature values. The line segments and line segment feature values are stored in the database table of Table 2. The partial serial graphic display of the air pressure, air temperature, and relative humidity in the original time series is shown in Figure 1. The comparison between the sub-time series and the line segment sequence after linearization is shown in Figure 2.

As can be seen from Figure 2, the amount of time series data is significantly reduced after piecewise linearization, and the main features of meteorological elements (slope, interval length, etc.) are extracted [10].

| Years | True value (m) | Calculated value (m) | Absolute error | Relative error (%) |
|-------|----------------|----------------------|----------------|--------------------|
| 2001  | 1.2057         | 1.1765               | 0.0292         | 2.4218             |
| 2002  | 0.7385         | 0.7803               | -0.0418        | -5.6601            |
| 2003  | 1.3454         | 1.3464               | -0.0010        | -0.0743            |
| 2004  | 1.2279         | 1.2622               | -0.0343        | -2.7934            |
| 2005  | 0.7833         | 0.7046               | 0.0787         | 10.0472            |
| 2006  | 1.2281         | 1.1742               | 0.0539         | 4.3889             |
| 2007  | 0.1936         | 0.3019               | -0.1083        | -55.9401           |
| 2008  | 1.2095         | 1.1783               | 0.0312         | 2.5796             |
| 2009  | 1.2827         | 1.2250               | 0.0577         | 4.4983             |
| 2010  | 1.2495         | 1.1165               | 0.1330         | 10.6443            |
| 2011  | 0.1730         | 0.2689               | -0.0959        | -55.4335           |
| 2012  | 1.3152         | 1.1097               | 0.2055         | 15.625             |
| 2013  | 1.1339         | 1.1102               | 0.0237         | 2.0901             |
| 2014  | 1.4558         | 1.4469               | 0.0089         | 0.6113             |
| 2015  | 0.6176         | 0.5637               | 0.0539         | 8.7273             |
| 2016  | 1.1175         | 1.1608               | -0.0433        | -3.8747            |
| 2017  | 1.2014         | 1.1477               | 0.0537         | 4.4698             |
| 2018  | 0.7867         | 0.8883               | -0.1016        | -12.9147           |
| 2019  | 1.0272         | 1.1882               | -0.1610        | -15.6737           |
The interannual distribution characteristics of summer precipitation days in this area are complex, and it is difficult to find an accurate pattern intuitively. The precipitation days have a declining development trend;

The summer precipitation days in this area only have a significant correlation with the Eurasian meridional circulation index 3) Randomly divide the 62a data into a training set sample (52a) and a test set sample (the remaining 10a) to establish the model's reliability, and the summer precipitation days and various climate factors in the region through the CART algorithm The classification and prediction model of whether the precipitation days are too much and whether they are too little is jointly established. The training accuracy rate of the model with whether the precipitation days are too many is 90.38%, and the training accuracy of the model with whether it is too few is 86.54%. The verification of the two models is accurate the rate is 80%, which achieves a good classification and prediction effect [11].
4. Conclusion
By applying multi-dimensional time series data mining technology to the field of meteorology, a rainfall weather model based on multi-dimensional time series data mining is established. The model initially shows the relationship between rainfall and temperature, air pressure, and relative humidity, which explains the weather phenomenon of rainfall well.

However, the model also has deficiencies, such as insufficient rainfall prediction accuracy, and fuzzy rules for moderate and light rain. If this model can be further improved and applied to the prediction of the near-rainfall weather, great achievements will be made. If we further promote the model and apply the idea of establishing the model to the analysis and prediction of other weather phenomena, this will surely achieve new results and will play a positive role in promoting the development of the meteorological caus.

References
[1] Li Wen Hui, Chen Yan, Liang Jiawen, et al. Application of time series mean generating function model in precipitation prediction. Guangdong Meteorological journal, 41 (4) (2019) 11-14.
[2] Wang Wensheng, Ding Jing, Jin Juliang. Random Hydrology. Beijing: China water resources and Hydropower Press, (2018).
[3] Xu Xiaoyu. Application of weighted Markov chain to annual precipitation prediction in Beijing. Forum of Science and Technology Association (Second half of the month), 25 (1) (2010) 111-112.
[4] Guangshen Tiandi Sheng Comprehensive Research Group on Cosmogenesis. Astronomical cycle and weather process forecast in Guangdong. Guangdong Meteorological journal, 15 (3) (1993): 23-25.
[5] García-Marín, A.P., Estévez, J., Medina-Cobo, M. T. & Ayuso-Muñoz, J. L. Delimiting homogeneous regions using the multifractal properties of validated rainfall data series. Journal of Hydrology, 529(1) (2015) 106-119.
[6] Muthuwatta, L. Perera, H. P. T. W., Eriyagama, N., Surangika, K. B. N. U., & Premachandra, W. W. Trend and variability of rainfall in two river basins in sri lanka: an analysis of meteorological data and farmers' perceptions. Water International, 42(8) (2017) 981-999.
[7] Bringi, V. N. Tolstoy, L., Thurai, M., & Petersen, W. A. Estimation of spatial correlation of drop size distribution parameters and rain rate using nasa's s-band polarimetric radar and 2d video disdrometer network: two case studies from mc3e. Journal of Hydrometeorology, 16(3) (2015) 1207-1221.
[8] Tang, J. Riley, W. J. & Niu, J. Incorporating root hydraulic redistribution in clm4.5: effects on predicted site and global evapotranspiration, soil moisture, and water storage. Journal of Advances in Modeling Earth Systems, 7(4) (2015) 1828-1848.
[9] Akcay, H. & Filik, T. Short-term wind speed forecasting by spectral analysis from long-term observations with missing values. Applied Energy, 191(4) (2017) 653-662.
[10] Yin, J. C. Wang, N. N., & Hu, J. Q. A hybrid real-time tidal prediction mechanism based on harmonic method and variable structure neural network. Engineering Applications of Artificial Intelligence, 41(5) (2015) 223-231.
[11] Suzuki, Y. Kaneda, Y. & Mineno, H. Analysis of support vector regression model for micrometeorological data prediction. Gastroenterology, 3(2) (2015) 37-48.