Research on Construction Method of Urban Simulated Rainfall Station Based on Sina Weibo Geographical Location Data

Pengl LI, Zening WU, Huiliang WANG*

College of Water Resources Science and Engineering, Zhengzhou University, Zhengzhou 450001, China

*Corresponding author’s e-mail: wanghuiliang@126.com

Abstract: In recent years, with the continuous improvement of the level of urbanization, the phenomenon of urban infighting has intensified. At present, urban rainfall data is mainly measured by rainfall stations, while the number of traditional rainfall stations and their uneven distribution result in the inability to obtain high-precision surface rainfall data. With the advent of the era of big data, more and more experts and scholars have applied big data to the research of natural disasters. Therefore, this article uses web crawler technology to obtain Sina Weibo data with geographic location information. By analyzing the correlation between the number of micro-blogs related to rainfall and the rainfall of the field, establishing the relationship between the two functions, and constructing the simulated rainfall station in the urban area of Zhengzhou City, more refined surface rainfall data can be obtained by interpolation. The experimental results show that the method of construction of simulated rainfall station can effectively improve the accuracy of interpolation through traditional rainfall stations.

1. Introduction

With the continuous improvement of China's urbanization level, the phenomenon of urban waterlogging has also attracted more and more attention from the public. In recent years, large and medium-sized cities such as Beijing, Shanghai, Guangzhou, Nanjing, Wuhan, and Zhengzhou have successively suffered from urban waterlogging disasters, causing huge losses to the local society and economy, and showing an increasing trend year by year. Rainfall data from urban rainfall stations is a necessary factor for studying prevention of urban waterlogging disasters. Wang Qingjing et al. [1] researched: ① The more uniform the spatial distribution of the rainfall station network, the smaller the rainfall interpolation error. ② In the case of uniform layout of rainfall station networks, the difference of interpolation results of each spatial interpolation algorithm is small; when the layout of rainfall station network is uneven, the number of stations is smaller, and the difference of interpolation results of each spatial interpolation algorithm is larger. Dai Peipei et al. [2] found that in the interpolation analysis of urban areas, the number of stations and the density of station networks have a greater influence on the interpolation accuracy, while the influence of station distribution is smaller.

With the continuous development of Internet technology, people are more and more willing to publish their own life trends on social platforms. Under the background of big data, potential values may be hidden. Wang Yandong [3] used social media data to simulate the urban air quality trend surface, qualitatively showed the relative air quality of different regions of the city, and also quantitatively and finely displayed the city air quality. Chu Junxian et al. [4] explored a method of capturing and displaying micro-blogs for the location of earthquake disasters, and realized the extraction of disasters and the graphical display of the severity. Amir Hassan Zadeh et al. [5] used big data analysis technology to
obtain the spatial and temporal correlations between the two flu trends and explain behaviors, helping health authorities designate more effective interventions during the outbreak to reduce transmission and impact.

This paper presents a method for constructing urban simulated rainfall stations using Sina Weibo geographic location data. By analyzing the correlation between Sina Weibo data and rainfall data, fitting the function relationship between the two, using this function correlation, screening and establishing a simulated rainfall station in Zhengzhou urban area. Obtaining simulated rainfall corresponding to the simulated rainfall station from the number of Weibo, and Kriging interpolation is carried out by using GIS software to get a more precise rainfall distribution map of Zhengzhou city. This is the first experiment to obtain refined urban rainfall data using Sina Weibo data with geographic location information.

2. Construction method of urban simulated rainfall station
The construction of urban simulated rainfall station and the acquisition of refined data mainly include the following 4 parts: rainfall data acquisition and Sina Weibo data collection and preprocessing, correlation analysis of the number of Sina Weibo and rainfall, function fitting and accuracy verification, urban simulated rainfall station construction.

2.1. Data collection and preprocessing
The research area of this paper is the urban area of Zhengzhou City, Henan Province, China. The data required for the research is from 2015 to 2018. Data collection and preprocessing are divided into the following 3 parts:
(1) 15 rainfall data provided by 12 rainfall monitoring stations in Zhengzhou from 2015 to 2018, 12 of which are sample data and 3 are verification data
(2) Use web crawlers to obtain Sina Weibo data. The web crawler is an important tool for obtaining network data[6]. It uses the Python requests module to simulate landing on the online version of Sina Weibo[7]. Analyze the source code corresponding to the advanced search keywords, types, inclusion, time, and location tags of Weibo, and use the bs4 module to extract the corresponding data and store it[8]. Based on the theme, the crawler program sets the parameters as "Keyword": "Rain", "Type": "Original", "Include": "All", "Location": "Zhengzhou, Henan", and "Time" according to "Year-Month-Day-Hour "format sets the start and end time of 15 rainfalls.

The results of crawling data are automatically saved and a csv table is generated according to the format of "user ID-post time-geographic location-Longitude-Latitude-Weibo content". The amount of raw data corresponding to each rainfall extracted is more than 3,000, including detailed geographic location data accounting for about 10%. The non-zhengzhou urban area and non-rainfall related microblogs are removed through data preprocessing operations such as deduplication and noise reduction, and the topic correlation rate is increased to more than 97%. The amount of microblog data in this part is represented by $N_1$.

(3) Standardize the Weibo data. Due to the influence of uneven population distribution in the urban area, it is not possible to directly use the rainfall theme Weibo number and rainfall to establish a functional relationship, that is, to eliminate the impact of the relatively large number of corresponding Weibo numbers in areas with high population density. Microblogs that do not contain keywords and have geographic location tags can reflect the urban population distribution to a certain extent. This article uses this part of data to standardize the number of microblogs to eliminate the impact of uneven population distribution. This part of the data is represented by $N_2$. The number of microblogs after standardization is denoted by $N$, and the calculation formula (1) is as follows:

$$N = \frac{N_1}{n} \sum_{i=1}^{n} N_2$$

In the formula, $N$ is the number of standardized microblogs, $n$ is the number of time slots in different periods of the year, $N_1$ is the number of topical microblogs in the buffer, and $N_2$ is the number of all
microblogs in the corresponding period in the buffer.

2.2. Correlation analysis

From the beginning of the rain to the end of the rain, a large number of Weibo users will publish rainfall-related blog posts in real time during the rain period, describing their status at the time, and the data containing potential value needs to be urgently scientifically excavated. Figure 1 shows the process line of rainfall-related microblogs and rainfall over time. The number of microblogs shows a significant correlation with the change of rainfall intensity. Therefore, it is inferred that there is also a certain correlation between the total rainfall of the event and the number of Weibo related to the rainfall. In this paper, the rainfall data of 12 rainfall monitoring stations in Zhengzhou City and Sina Weibo data are taken as samples to analyze the correlation between the number $N$ of microblogs around the rainfall monitoring station site and the rainfall of each session.

The correlation analysis process is as follows: First, use ArcGIS to visualize the Sina Weibo tag data with latitude and longitude obtained by the web crawler. Then, draw a buffer with a radius of 1.5km, 2.0km, 2.5km, 3.0km, and 3.5km centered on the rainfall monitoring station area, count the number of $N_1$ and $N_2$ in the buffer of each site, and substitute the formula (1) to obtain the number $N$ of microblogs after standardization. Statistics show that the amount of $N_2$ in the buffer zone of the rainfall station in the urban fringe area of Zhengzhou is very small or even 0, resulting in drifting during correlation analysis. Therefore, this article only conducts rainfall and Weibo number $N$ at five rainfall stations (North China University of Water Conservancy and Hydroelectric Power, Zhengzhou Municipal Water Resources Department, Henan Provincial Party Committee, Zhengzhou Railway Bureau, Zhongyuan District Zhengzhou Hydrological Bureau Office). The correlation analysis of Table 1 is the correlation coefficient of the number $N$ of microblogs in each buffer radius and the rainfall Pearson. It can be found that when the buffer radius is 2.0km, the Pearson correlation coefficient is at most 0.812, and the significance level is below 0.01, so it shows that there is a significant correlation between the number of microblogs $N$ and rainfall.

![Fig. 1 Rainfall related Weibo quantity and rainfall process line](image1)

Table 1 Correlation between the number $N$ of microblogs in each buffer radius and rainfall

| Buffer radius/km | Correlation coefficient | Significance |
|------------------|-------------------------|--------------|
| 1.5              | 0.692**                 | 0.000        |
| 2                | 0.812**                 | 0.000        |
| 2.5              | 0.647**                 | 0.000        |
| 3                | 0.648**                 | 0.000        |
| 3.5              | 0.693**                 | 0.000        |

Note: ** indicates significant correlation at the .01 level (both sides).

2.3. Function fitting and accuracy verification

Count the 10 rainfall events measured in five rainfall stations in the central urban area of Zhengzhou in 2015-2018 and the number of microblogs in the buffer zone corresponding to the 2km buffer radius of each rainfall station $N$.

![Curve fitting of the rainfall of the event and the number $N$ of Weibo](image2)

Figure 2 shows the number $N$
of microblogs and the scatter plot and fitting curve of the rainfall amount. There is a clear linear relationship between the two. Use function drawing software Origin to fit the curve, the function equation obtained is:

\[ Y = 135.45X + 10.39 \]  

(2)

In the formula, \( X \) is the number of standardized microblogs \( N \), and \( Y \) is the rainfall of each session. The goodness of fit \( (R^2) \) of the equation is 65.9\%, which indicates that the regression line fits the observed value to 65.9\%. The \( F \) test statistic is 90.84, which is significant at the level of 0.01, indicating that the number of Weibo \( N \) can explain the rainfall of the event well. The above \( R^2 \) and \( F \) test statistics show that the regression line equation fits well. At the same time, it can be seen from Figure 2 that the rainfall of a session increases linearly with the increase of the number \( N \) of microblogs.

![Fig. 2 Weibo number N and rainfall scatter plot and fitting curve](image1)

In addition, the accuracy of the fitting function was verified by selecting 5 rainfall events from 2015 to 2018. The number of microblogs at the site was counted and the number of measured field rainfalls. Substitute \( N \) into the fitting equation formula (2). The average relative error between the simulated and measured values of each field is shown in Figure 3. The average relative error between the measured and simulated values of each field is between 15\% and 25\%. As shown in Figure 4, the two rainfall scales of Zhengzhou City on June 5, 2016 and September 18, 2018 are not the same, and the measured rainfall is compared with the simulated rainfall. It can be found that using the fitting equation (2) The simulated value of is close to the measured value of the rainfall station, and the function fitting effect is better. The above results all indicate that it is feasible to simulate the rainfall of the number of times using Weibo number \( N \).

![Fig. 3 Average relative error between measured and simulated rainfall values](image2)
2.4. Construction of simulated rainfall station

In the above study, it is found that the error between the measured rainfall at the rainfall station and the corresponding simulated rainfall basically meets the requirements, so it is assumed that other areas outside the urban area outside the rainfall station also meet this correlation. Based on this assumption, this paper builds a city simulated rainfall station based on the N correlation between rainfall and Weibo number and validates the analysis. The construction of simulated rainfall stations can be divided into three parts: selection of simulated rainfall stations, GIS cross-validation, and refined rainfall data acquisition:

(1) Screening of simulated rainfall stations. Due to statistical data analysis, it is found that the N2 variability of the number of microblogs in each buffer zone of non-urban rainfall stations in Zhengzhou is relatively large, and it is impossible to construct a functional relationship between rainfall and the number of microblogs. The central urban area is the research area (the area surrounded by the north, west and south ring roads of Zhengzhou City and Zhongzhou Avenue). Combined with 5 traditional rainfall stations in this area, the simulated rainfall stations were evenly arranged. According to the number of microblogs in the multi-field rainfall buffer, it was found that there are also no keyword microblogs in the buffer zone of the city. For places with low population density, such as parks, these simulated rainfall sites change the number of simulated rainfall sites with the number of theme microblogs N1 in the simulated rainfall sites to be too high to simulate rainfall effectively, so the simulated rainfall in these areas is eliminated. At the site, 12 simulated rainfall stations are retained as shown in Figure 5:

![Fig. 5 Distribution of simulated rainfall stations in urban areas](image)

(2) Cross-validation, its essence is to take a part of each time as a prediction point, these points do not participate in the interpolation process, but as a verification point to verify the predicted value of this point, so that you can get the actual value of all points deviation from the predicted value To a certain extent, the deviation of is reasonable for us to provide interpolation methods. In this paper, the common Kriging interpolation cross-validation method is used to verify the accuracy of the interpolation based on simulated rainfall sites[9]. The accuracy of the model that meets the following standards is optimal: The standard average is closest to 0, the root mean square error is closest to the average standard error, and the standard root mean square error is closest to 1. It means that the variability in prediction
is correctly estimated. Randomly choose 5 rainfalls, count the number N of microblogs in the 2km buffer area of each simulated rainfall site and the corresponding rainfall of traditional rainfall sites, and substitute the number N of microblogs of each simulated site into formula (2) to get the simulated rainfall. According to the simulated rainfall and the rainfall data measured by the traditional rainfall station, the ordinary Kriging interpolation of five rainfalls is conducted and cross-validated. The prediction errors corresponding to the rainfall interpolation of each field are shown in Table 2. The standard average error is between -0.0159 and 0.0241, and the standard root mean square error is between 0.9307 and 1.0499. The error results show that the accuracy of the model is better, that is, the construction of the simulated rainfall station meets the accuracy requirements.

| Error / rainfall | 2018/5/15 | 2018/8/18 | 2018/9/18 | 2017/8/18 | 2017/7/29 |
|------------------|-----------|-----------|-----------|-----------|-----------|
| standard average | 0.0241    | -0.0634   | -0.0159   | 0.0225    | 0.0214    |
| root mean square | 19.6750   | 5.5990    | 11.9526   | 21.0555   | 21.1195   |
| average standard | 22.0310   | 5.4340    | 12.6126   | 21.5029   | 19.7621   |
| standard root    | 0.9307    | 1.0499    | 0.9533    | 0.9900    | 1.0487    |

(3) Obtain the rainfall trend graph. From the above research content, it can be obtained that using Weibo number to simulate rainfall is feasible, and the accuracy meets the requirements. Using GIS mapping, the number of microblogs after normalization N is substituted into the optimal function equation to simulate the rainfall corresponding to the rainfall site, and the rainfall trend graph is drawn in conjunction with the rainfall data of the traditional rainfall site. And it can make up for the defect of low interpolation accuracy due to the small number of rainfall stations and uneven distribution.

3. Conclusion
This article uses web crawler technology to obtain Sina Weibo data with geographic location information, and analyzes the correlation between the number of rainfall-themed Weibo and the number of rainfall events. The study found that when the buffer radius is 2km, the Pearson correlation coefficient is 0.812 at the highest, and a functional relationship between the two is established. 12 simulated rainfall stations within the downtown area of Zhengzhou are constructed. It makes up for the shortcomings of the city's small rainfall stations, uneven distribution, and the construction of actual rainfall stations that cost a lot of manpower, material, and financial resources, and provides a new data source for the acquisition of traditional urban hydrological rainfall data. After verification, the research method is considered feasible and credible.

The rise of social media big data provides a new platform for interdisciplinary research. Facing these types of data with relatively low value density, fast processing speed, and high timeliness requirements. How to dig out potential value from it will be a process of continuous in-depth research.

References
[1] Wang Qingjing, Yang Xinyue, Chen Hua, et al. Influence of Rainfall Network Distribution on rainfall interpolation algorithm and runoff response[J]. J Yangtze River Sci Res Inst, 2019, 36(4):19-26
[2] Dai Peipei, Jing Changfeng, Du Mingyi. Uncertainty analysis of urban rainfall interpolation based on rain gauge[J]. Yellow River, 2015, 37(4):38-42
[3] Yandong W, Tong J, Wei J, et al. Modeling urban air quality trend surface using social media data[J]. Geomatics and Information Science of Wuhan University, 2017
[4] Xu Jinghai, Zhu Junxiu, Nie Gaozhong, et al. Earthquake disaster extraction based on location microblog[J]. J Nat Disaster 2015(5):12-18
[5] Hassan Zadeh A, Zolbanin H M, Sharda R, et al. Social media for nowcasting flu activity: spatio-temporal Big Data Analysis[J]. Information Systems Frontiers, 2019
[6] Lu S, Ma H, Gao Y. Designed a web crawler which oriented network public opinion data acquisition[C]// Sixth International Conference on Electronics and Information Engineering. International Society for Optics and Photonics, 2015
[7] Sun Qingyun, Wang Junfeng, Zhao Zongqu, et al. Microblog data collection scheme based on analog login[J]. Computer Technology and Development, 2014(3):6-10

[8] Kumar M, Bindal A, Gautam R, et al. Keyword query based focused Web crawler[J]. Procedia Computer Science, 2018, 125:584-590

[9] Liu Shiwei, Wu Jinkui, Zhang Wenchun, et al. Comparative analysis of sampling methods for estimating regional precipitation based on kriging Interpolation[J]. Journal of Glaciology and Geocryology, 2015, 37(3):650-657