Optimization of Rainfall Sensor Network Layout Based on the Correlation Coefficient Method

Hongling Zhao, Haibo Yang* and Zongmin Wang
School of water conservancy and environment engineering, Zhengzhou University, Zhengzhou 450001, China
*Corresponding author

Abstract—**The unreasonable layout of rain sensor networks results in various problems, such as the redundancy of observation nodes, large daily maintenance overhead, and the difficulty of solving faults in real time. A sensor network optimization algorithm based on the correlation coefficient method is proposed in this work. First, Spearman's rank correlation coefficient is adopted to calculate the correlation of each node, and computing results are stored in databases. Second, the nodes with a significant correlation with other nodes are selected, and historical data are used to verify the selected nodes. Finally, we improve the proposed method by introducing a sub-region calculated using the DEM and growing tree method. Experiment results show that the proposed method can achieve an accurate prediction of rainfall in an entire region with the optimization of the layout of observation nodes.**

Keywords-Rainfall sensor network; Correlation coefficient; Rainfall forecast; Spatial correlation style

I. INTRODUCTION

Given the limitation posed by geographical environment factors, rain sensors are typically installed in remote or inaccessible places in accordance with specific requirements [1]. However, maintaining rain sensors is relatively difficult. Moreover, limited power supply and the adverse effect of natural factors result in the relatively high failure rate of rain sensors [2-3]. The optimization of rainfall sensor networks facilitates the regional estimation of rainfall via the interpolation of optimized sites, which results in reduced costs and increased stability. Hence, the optimization of the rainfall sensor network layout is a hot research topic in the field.

The topology optimization problem of wireless sensor networks is studied, and energy balancing and the deployment of sensor network coverage are optimized by using game theory and fuzzy control theory [4]. An artificial immune system and the similar relations among wireless sensor networks are used to propose an optimization algorithm based on the immune stimulation mechanism of wireless sensor networks. This algorithm reduces the redundant nodes and reduces the network energy consumption. Meteorological hydrological monitoring sensors do not differ greatly from traditional wireless sensors in terms of their deployment schemes and transmission strategies [5]. A clustering analysis method is used to divide a certain area for hydrological regionalization and realize the regional hydrology and optimized deployment of sensor networks according to terrains, densities of sensor nodes, and other hydrological parameters [6]. The layout of sensor network deployment is optimized using the combination of spatial and regression Kriging algorithms and simulated annealing algorithms on the basis of the characteristics of sensor nodes within a certain area in a given space and time [7]. The acquisition and structure of hydrological information in wireless sensor protocols are investigated [8-9].

A certain area that receives precipitation with a space–time correlation [10] becomes a regional rainfall site that can be used for precipitation forecast. The text is based on the precipitation explained by the space–time correlation principle. Specifically, the correlation coefficient of sensor nodes between geographical and historical rainfall data is calculated, the calculated correlation coefficients are compared, and the relationship among other points is determined to be relatively strong. Through the correlation between precipitation sites, it can identify connections among points to determine representative points. The accurate monitoring of an entire area can only be realized by monitoring few representative points. Considering the complexity of the algorithm when applied in large areas, the present work aims to study the application of the hydrological partition algorithm in large areas on the basis of DEM data, combined with historical rainfall and rainfall distribution site data. By establishing certain rules, the study area is divided into several hydrologic regions to reduce the complexity of the algorithm and to provide hydrological basis for the optimization of sensor networks.

II. METHODOLOGY

A. Spatial Correlation

Spatial statistics is an important statistical method that has been globally used since the 1970s. This method is widely applied in various fields, such as remote sensing science, geography, biology, and epidemiology. The statistical analysis of location-related factors can determine the spatial dependence and spatial correlation of relevant factors. Analysis mainly involves two aspects: (1) the distribution rule and aggregation of points in a space and (2) the analysis of directional data.

According to American geographer Tobler, the first law of geography indicates that objects distributed in a geographic area do not exist in isolation. Studying the links between objects and their proximity is thus important in determining the distribution of objects in space. The relation between things can be accurately determined by calculating the spatial correlation. Rain is a natural phenomenon, and a spatial correlation necessarily exists between neighboring rainfall stations according to the first law of geography. At present, the analysis of rainfall in a region is mainly based on the spatial correlations of rainfall interpolation methods, such as the inverse distance weighted average method, Kriging method, and collaborative Kriging method.

In the present work, spatial correlation is employed to calculate the relevance of experimental stations in a specific region according to the longitude and latitude of precipitation stations and historical statistical rainfall data. Through this approach, the stations with the highest dependence can be identified as the main nodes according to specific regions, the current national guidelines on hydrologic network planning, the requirements for rainfall observation points, and the identified major rainfall monitoring stations in the region as a whole. By maintaining the rainfall sensors of main stations, the operation of rainfall sensor networks can be simplified.

B. Spearman's Rank Correlation Coefficient

Correlation coefficients are used to reflect the statistical indicators of the connection between two objects. They are also used as a clustering method for partitioning. By using statistical methods, the correlation coefficient of two objects in a set can be calculated according to given threshold values or confidence values using a corresponding algorithm for sorting objects. In practice, correlation coefficients come in three types: Pearson correlation coefficient, Spearman's rank correlation coefficient, and Kendall correlation coefficient.

The Pearson correlation coefficient is used to calculate the strength of the linear relationship between two objects. In the calculation, the variables are normal continuous variables, and a linear relationship exists between two objects. The Pearson correlation coefficient between two objects is the covariance and standard deviation plot of two variables x and y:

$$\rho_{xy} = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y} = \frac{E(X - \mu_x)(Y - \mu_y)}{\sigma_x \sigma_y} \tag{1}$$

Where σ_x and σ_y are the standard deviations of the parameters x and y, respectively; cov(x, y) is the covariance of x and y. To calculate the Pearson correlation coefficient, the variables should meet the above criteria, and the data should be evenly spaced across the logic. These requirements restrict the use of the Pearson correlation coefficient. Spearman's correlation coefficient is the most common non-parametric correlation coefficient analysis method. This method mainly analyzes the rank correlation between two objects, sorts the two objects within the data, and employs the sorted data in place of the actual data to achieve the order of precedence obtained by the correlation coefficient.

The Kendall correlation coefficient is named after Maurice Kendall and is often denoted as the Greek letter τ. The Kendall correlation coefficient tests the statistical dependence between variables by calculating the correlation between two random variables. This method is a type of hypothesis test of non-parameters, and the variables used in the operation satisfy the orderly classification. When collecting data distribution and when status is unknown, the Pearson or Spearman rank correlation coefficient should be adopted because the use of the Kendall correlation coefficient yields small results.

In practical applications, Spearman's rank correlation coefficient is generally considered after identifying the Pearson correlation coefficients of two objects. If the sample covariance and standard deviation of the rainfall calculation are used instead of the general covariance and standard deviation, then the sample correlation coefficient is

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{2}$$

All site statistics are based mainly on rainfall in 31 days. The correlation between the rainfall data in two sites is calculated because the rainfall of each site is random and the two sites have no linear relationship. In this study, Spearman's rank correlation coefficient is used to calculate the relationship between adjacent stations.

For a specific area for rainfall observations, the vantage point of rainfall sensors must be identified before using the permutation algorithm partitions, in which the total number of points is n, its combination number is ++...+=, and complexity is 0(n). An increase in combination increases the operation required. The total number of combinations equates to the exponential growth. Using the six-point operation speed for this experiment is acceptable on an ordinary PC. If the operation speed is extended to 24 points or more in the region, then the complexity of the algorithm exhibits an exponential growth. Furthermore, the computing could not be completed within the desired time using an ordinary computer. These conditions enhance the difficulty involved in the rainfall prediction of other points. If it was divided into m areas, complexity given by 0(m) is greatly reduced. By selecting the master node for monitoring precipitation in the region as a whole, relatively small amounts can be maintained to reduce the maintenance costs for the entire sensor network.

C. Theory of Hydrologic Regionalization

The optimization of rainfall sensor networks is aimed at optimizing network structure and rationally designing the network layout with minimal inputs to maximize the collection of rainfall information with the highest accuracy. The role of the network site is limited to the rainfall observations of the site that meet the demand for hydrological data from anywhere within the network at the time of site optimization to influence the precision of the precipitation interpolation of hydrological factors and the application that considers available data. Hydrologic regionalization from space reveals regional hydrological characteristics and provides hydrological basis for the optimization of sensor networks. It also serves as the basis of rainfall sensor network optimization. Moreover, hydrologic regionalization significantly reduces the complexity involved in maintaining master nodes and is thus conveniently applied in practice. This study uses the DEM of the study area, combined with the historical rainfall data and the distribution of rainfall sensor data extracted from the catchment and watershed in the study area. The tree method is adopted, and according to a certain rule, watershed hydrologic regionalization is merged with the final result.The following three main factors are considered in the establishment of rules:(1) The factors affecting the complexity of the extraction algorithm for the host node of the rainfall sensor: the number of rainfall sensor sites; (2) The factors affecting the accuracy of rainfall monitoring: equilibrium of distribution of rainfall sensor stations, regional shape coefficient, and average rainfall for a given number of years; and (3) The factors affecting the final application of available hydrological data: the addition of a regional basin.

D. *Algorithm design*

The Java development environment is used, the rainfall data of one area for the month of August are combined, and the methods of permutations and combinations are employed to select the points included in the operation. Let (v, p) be two sets, in which v is the measuring point and p is the point to be measured. The algorithm that combines permutation and combination methods is used to select the nodes in set v that are involved in the calculation. All the remaining nodes are input into set p. Then, the correlation coefficient of the nodes in sets v and p is calculated. If the combined number of nodes in v is greater than 1, then the related coefficient is averaged. The algorithm that combines permutation and combination methods is as follows:

- Select the combination of measuring points from all points. Let the set of all points be sum [N], and select a combination of c points from n. Step 1: initialize the array, mark c 1 in the array, and initialize the output calculation (). Step 2: set the end field: end is equal to false, and judge whether the algorithm is over or not. Step 3: find the first array element marked as 1 in a front-to-back order in the array (for loop). Step 4: find the first continuous “10”, exchange them, and label the switch symbol: swap is equal to true. Step 5: if the swap is equal to true, then exit this cycle. Step 6: determine whether all the last c digits are 1, and set the end as equal to true if they are. Step 7: call calculate (), output the result of combination, and enter step 2. The program chooses six points for testing. The program first introduces the points corresponding to the rainfall data in August into a two-dimensional array rain [N][M]. The points of this combination are then elected through permutation and combination algorithms. According to the array’s (sum[N]) element marked as 1 in calculate (), the point of set v and the point to be measured from set p are selected for the operation with the corresponding rain [N][M]. Two arrays (x[M],y[M]) are chosen for calculating the correlation coefficient. .

- The calculation of the correlation coefficient between rainfall stations. Step 1: input two arrays: x[M] and y[M] to the function Spearman(). Step 2: sort the elements in x[M] and y[M], set the serial number as the element value, and input them to x_Rank[M] and y_Rank[M]. Step 3: calculate the average value of x_Rank[M] and y_Rank[M]. Step 4: calculate the correlation coefficient of two points using Formula (3). Step 5: average the correlation coefficients, and enter them into the database.

E. *Experimental Results and Analysis*

Experimental data cover six rainfall stations in Zhengzhou, such as the stations in Zhengzhou City and Xingyang County. The data include the daily rainfall observation data for the period of 2012–2013, which spans seven to nine months. The rainfall data covering 65 days are randomly selected and used as the experimental data. The data covering the remaining 27 days are used for test verification. When we use the average constantly, the correlation coefficient decreases as points increase because the correlation coefficient of each observation point is known. The results are shown below.

TABLE I. CORRELATION COEFFICIENTS FROM THE TEST DATA FOR THE SELECTED RAINFALL STATIONS

Predictive points	Top three point combinations	Correlation coefficient
1	3	0.9994
	6	0.9994
	1	0.9969
2	1, 6	0.9982
	1, 3	0.9976
	3, 6	0.9963
3	1, 5, 6	0.9819
	1, 3, 5	0.9818
	3, 5, 6	0.9779
4	1, 2, 3, 5	0.9668
	1, 2, 5, 6	0.9666
	2, 3, 5, 6	0.9626
5	1, 2, 4, 5, 6	0.9527
	1, 2, 3, 4, 5	0.9526
	2, 3, 4, 5, 6	0.9493

As shown in the above results, points 1, 3, and 6 always achieve the largest correlation coefficients. Moreover, when point number reaches three to four, the correlation coefficient declines slowly. Accordingly, we can choose three points as rainfall observations, which can provide good observation basis for the other rainfall stations. In the one-point case, the correlation coefficients of points 1, 3, and 6 relative to the other points are high. According to Tables 1 and 2, the average correlation degree of point 3 in the one-point case is the highest. Thus, we recommend that point 3 be used to evaluate the total rainfall in this area. As shown in Table 1, in the two-point case, some combinations for points 1, 3, and 6 are (1,6), (1,3), and (3,6), respectively. When we verify the other months, point 4 appears in combination in Table 2. However, the correlation coefficient of the two-point combination relative to other points reaches 0.997 according to Table 1. If we add point 4, the correlation coefficient in Table 2 becomes 0.958. With system stability, (3, 6) is chosen as a node to evaluate rainfall in the area under the two-point case. Under the three-point case, point 5 joins the combination in Table 1. The top three combinations are (1, 5, 6), (1, 3, 5), and (3, 5, 6). Point 5 also joins the combination in Table 2; thus, (3, 5, 6) is the third combination. Therefore, we recommend (3, 5, 6) as the evaluated combination under the three-point case. Given the increasing number of combinations, the correlation of the combination relative to the other points is unstable, and the differences are highly significant. Thus, in the four-point case, we choose the combination (1, 3, 5, 6). In the five-point case, the correlation of the combination (1, 2, 4, 5, 6) relative to other points is higher than that of the other combination in the experiment and verification phases. Therefore, (1, 2, 4, 5, 6) is the best combination in the five-point case. Table 2 presents the recommended rainfall details.

TABLE II. NUMBER OF COMBINATION POINTS AND CORRESPONDING RAINFALL INFORMATION

Top three point combinations	**Correlation coefficient**
1	3
2	3, 6
3	3, 5, 6
4	1, 3, 5, 6
5	1, 2, 4, 5, 6

Given that the experiment uses permutation and combination algorithms and that the correlation coefficient with only six points is calculated, the computational complexity is relatively low. In practice, one city often has more than 20 stations. If we arrange and combine the data, then the operation becomes time consuming. We cannot realize the real-time rainfall observations of other points. This study thus partitions these rainfall stations in a large area using the zoning method.

III. RESULTS

When dividing the area, we should consider various factors, such as climate, hydrological characteristics, and geographical conditions. We should also consider the integrity of the river system. Partitions should match the analysis of the station network. The 174 stations in the study area are mainly located in the same small area (the same urban district). On the basis of the 30 m resolution DEM of Zhengzhou City downloaded from a geo-spatial data sharing platform, we extract the small watershed using the hydrologic analysis tools in ArcGIS and extract the basin river system in Zhengzhou. After the registration of the scanning electronic edition of the 1:50,000 paper map for Zhengzhou, we can correct the water system from the DEM and obtain the final field of the water system based on the extracted river network. Zhengzhou City is then divided into five basins: Ying basin, Jialu River basin, Yellow River basin (Yiluo River–Sishui River basin), White Drop River basin, and Double JI River basin. Meanwhile, the basin codification of each watershed is added into the watershed attribute table vector file. According to the historical rainfall observed by rainfall stations in Zhengzhou City, we can obtain the raster image of the historical rainfall in different periods by performing an interpolation for every period. Interpolation grids are overlaid to derive the figure of the average rainfall of every grid for a period of many years. Using the zonal statistics tool in ArcGIS and the figure of grid rainfall and watershed vector, the average rainfall for a period of many years is calculated. The area and perimeter of each watershed are calculated through a self-defined VBA code in the field calculation tools of ArcGIS. The barycentric coordinate of the polygon of each watershed is calculated using the field calculation tools in ArcGIS. The watershed vector and figure of the rainfall station are then overlaid. The number of rainfall stations in each watershed is added into the watershed attribute table using the spatial join tool in ArcGIS.

Through the above steps, we can finally obtain the vector watershed file. The growing tree method is used to merge the watershed on the basis of the discriminant Formula. According to the results of the partition, the nodes in the partition region are fewer than those in the previous large region. Combining and optimizing algorithms are obviously convenient. Meanwhile, the nodes after the partition can efficiently report real-time data to meteorological stations via ad hoc networks.

We can simplify the nodes in the area by using the correlation coefficient method discussed. Some regions only have four rainfall observation points. Hence, we should use one to three points with the highest correlation degree to measure the correlation degree relative to the other points. The groups with the highest correlation degree on the map are shown in Figure 1. In this figure, the first candidate points show the situation of the points with the highest correlation degree in the entire area when using only one point in a small area. The second candidate points show the situation of the rainfall station when using two points. They include the first candidate points. The third candidate points show the distribution of the entire rainfall station when three points are chosen. They include the first and second candidate points. In practical applications, observers can evaluate total rainfall by observing the specific points according to a given accuracy.

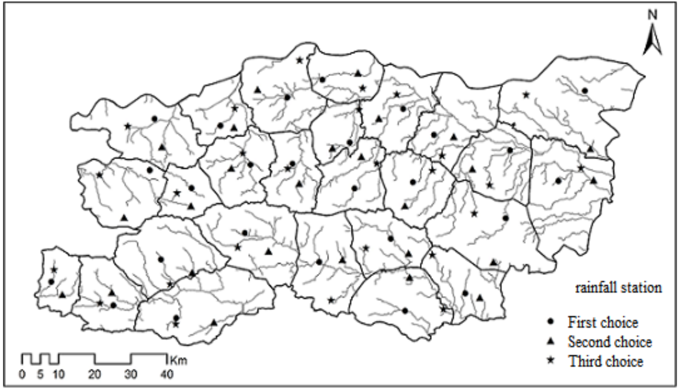

FIGURE I. EXAMPLE OF A TWO-COLUMN FIGURE CAPTION: (A) THIS IS THE FORMAT FOR REFERENCING PARTS OF A FIGURE.

IV. CONCLUSION

Considering the problems brought by the large number of stations and the difficulty in maintaining them, we propose a correlation coefficient method for identifying points with higher correlation coefficients than the other points in a specific area. Then, we can simplify the maintenance of all nodes via the maintenance and prediction of key points. The main algorithms include the permutation and combination algorithms, as well as Spearman's rank correlation coefficient. In the algorithm that combines the permutation and combinations, we mainly add a flag bit and constantly move the flag bit to determine different combinations. We then calculate the correlation coefficients of target points according to these different combinations. The point with the highest correlation among the remaining points is determined as the key point by analyzing the results. Spatial clustering analysis allows the entire region to be divided into small areas to calculate rainfall in real time and to manage rainfall stations easily. The experimental results show that we can evaluate the rainfall of an entire area using only a few nodes by using this method.

ACKNOWLEDGMENT

This work was financially supported by Science and technology project of Henan Province (142102210069).

REFERENCES

[1] ZhaoliangPeng, Ziru Wang, Guoli Wang, etc.Correction and integration multi-mode of China quarterly precipitation forecast[J]. Advances in water science, 2014, 25 (001): 1-9.

[2] XueWang ,YangChen,and WenhuiYuan.The causes and solutions of automatic weather stations rainfall sensor error [J]. Meteorological hydrological and marine instrument, 2013, 30 (1): 114-115.

[3] CuihuaZhang, TaoBian, lirongWang. Analysis of factors influencing quality of AWS precipitation data [c][J]. the 27th China Meteorological Institute of atmospheric physics and atmospheric environment parallel sessions symposium 2010.

[4] GengzhongZheng. Topology control in wireless sensor networks and optimization [D]. Xidian University, 2012.

[5] YongjunChen. Wireless sensor network optimization and fault diagnosis research based on artificial immune system[D]. Nanjing Aeronautics and Astronautics University, 2011.

[6] Qili BI. Optimization study of hydrologicalsensor network deployment [D]. Zhengzhou University 2010.

[7] YongGe, JianghaoWang,JinfengWang, etc. Optimization in eco-hydrological wireless sensor networksbased on regression kriging [J]. Advances in Earth science, 2012, 27 (9): 1006-1013.

[8] Lei Lu, Xinghui Yin, Dadong Zhang. Application of WSNs on water information collection [J]. Sensors and Microsystems, 2011, 30 (001): 134-136.

[9] Yun Bai. The precise monitoring and early warning of urban flood control based on rain sensor -Xicheng District in Beijing as an example [J]. China information, 2012 (17): 61-63.

[10] Lu G Y, Wong D W. An adaptive inverse-distance weighting spatial interpolation technique[J]. Computers & Geosciences, 2008, 34(9): 1044-1055.