Water Supply



Water Supply Vol 22 No 4, 4387 doi: 10.2166/ws.2022.057

# Research on partition strategy of an urban water supply network based on optimized hierarchical clustering algorithm

Wei Xia, Shi Wang*, Mingjun Shi, Qing Xia and Wenting Jin

School of Electronics and Information Engineering, Anhui Jianzhu University, Hefei, Anhui, China
*Corresponding author. E-mail: ws2233010@163.com

SW, 0000-0001-9891-6866

## ABSTRACT

The partitioning of the urban water supply network can significantly enhance water supply quality. Nonetheless, the bulk of the recently deployed partition approaches overlooked the question of whether the district's fluctuation regulation of flow data is consistent. When the district is modified, it most likely leads to an increase in pressure at a node. To tackle the problem, the flow data from a city's water supply network was evaluated in this article. The random forest approach was also used to extract time-domain characteristics from flow data, and the water supply network split was optimized using the random forest-hierarchical clustering (RF-HC) strategy. Finally, the results were examined and compared. The results suggest that the RF-HC-based water supply network partition technique can better meet the aim of consistent flow changes in the district, as well as offer a theoretical foundation and technological support for the optimal dispatch of press concerning the water supply network.

**Key words:** features of time-domain, partition strategy of water supply network, RF-HC, water supply network

## HIGHLIGHTS

- This paper realized the variables of the degree of flow change in single day.
- An optimized hierarchical clustering algorithm based on random forest is proposed.
- Take actual coordinate information as calculation parameters.
- The result could be helpful with the optimal dispatch of press.

## INTRODUCTION

District metering area (DMA) was founded on the partition method of an urban water supply pipe network system. Its primary application is to split the complete water supply pipeline network into various autonomous communities. It can detect and reduce leaks in urban water supply systems. Further study on the partition technique has also piqued the interest of numerous specialists and scholars in recent years.

An early DMA approach was the partition method based on experience. Geographical circumstances and regional functions are the main partitions in this strategy. The approach is straightforward, but it ignores the real hydraulic conditions (Wenqi *et al.* 2012; Bin *et al.* 2018).

The partition technique, which is based on the community structure approach, considers several comparable nodes in the pipe network as a group when calculating the optimal way to split. Diao K *et al.* advocated governing the water supply nodes as a community and calculating the border density inside the community as well as the boundary density between communities (Diao *et al.* 2013).

The partition processing of the water supply network, based on graph theory, may effectively portray the geographic information link among the nodes to a certain extent. Alvisi *et al.* proposed a method for obtaining DMA partitions of a specified size by using the depth-first search algorithm in graph theory (Alvisi and Franchini 2014); Liu proposed combining the shortest path method in graph theory with geographic information systems to calculate the optimal partition boundary (Jun &

Guoping 2012). Both of these approaches essentially address the growth of cross-regional partitioning. At the same time, it decreases the overall energy consumption of the pipeline network to some amount.

The partition approach, which is based on multi-objective optimization, can handle multi-objective issues that arise during the real operation of the water supply network. Paola *et al.* devised a partition technique that optimizes the uniformity of the water supply in the partition and the pressure differential between nodes in the partition (Paola *et al.* 2014).

The final optimization target for the majority of the partition methods listed above is static water flow and pressure homogeneity in the pipe network. This overlooked the variance in regulation between nodes in the same district. When the booster pump in this district was modified, this resulted in varied flow and pressure fluctuations.

The research proposes replacing traditional flow data with variation pattern and time-domain flow characteristics based on DMA partitioning, and then using random forest to improve hierarchical clustering (HC) to accomplish partition planning. Unlike the traditional flow data-based method, the patterns of change and time-domain features may guarantee that the flow rates of nodes within the same partition have comparable patterns of change as much as feasible, which can ultimately assist to the optimized scheduling of the pipe network

It should be mentioned that because the experimental environment of this work comes from a demonstration project in the city of HC. Because some of nodes have been abandoned in later projects, there are now a low number of nodes. This leads a to the unfavorable usage of random forests. However, even after the addition of additional nodes, this strategy is still relevant to the water supply network following node expansion.

## METHODS

The flow data of the urban water supply network expresses the water demand of consumers intuitively. The pipe network's districts might be split based on the characteristics of this data. The goal is to ensure that each node in the same district has a similarly changing flow data control. This research introduces the time-domain properties of the extracted flow data using the random forest technique, as well as the degree of change of a single node in a set period of time, as parameters to complete the partition of the water supply network, based on the HC algorithm. Figure 1 depicts the whole algorithm's flow.

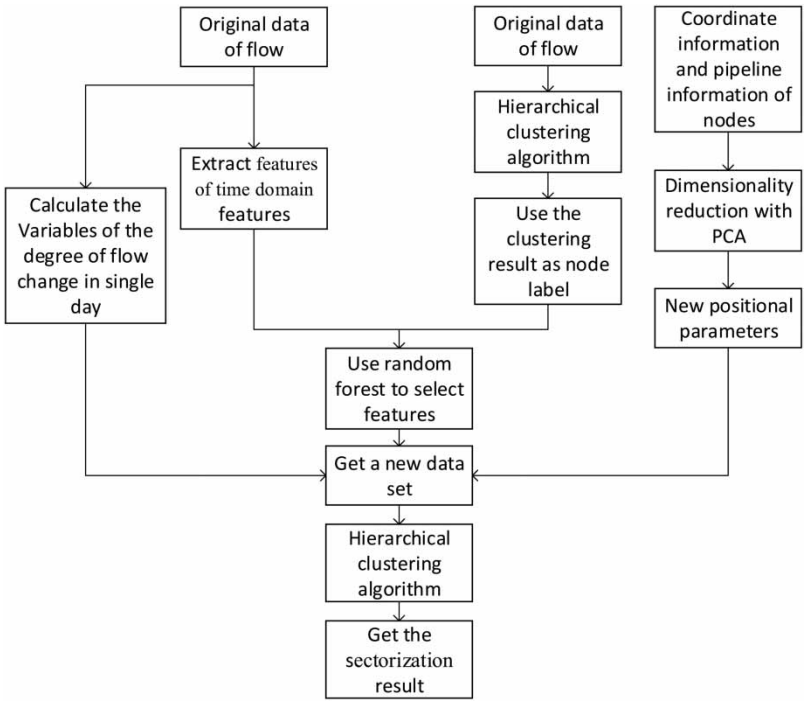


**Figure 1** | Overall algorithm flow chart.

The following are the procedures for partitioning the water supply network using the random forest-hierarchical clustering (RF-HC) method.

(1) Partitioning preliminary. To analyze the original data, HC is employed, and the nodes with the least distance between classes are computed to accomplish clustering of nodes with the smallest difference in traffic between each other, resulting in the theoretically ideal partition state.
(2) Extraction of features. The major goal of this article is to use variation in flow data as the foundation for zoning calculations instead of flow data. The article provides a new parameter, $W_c$, to capture the degree of pressure change over the day as well as the time-domain properties of the pressure data for this purpose. Simultaneously, to ensure the partitioning's practicability, the location parameter is incorporated as a constraint, i.e. location parameter 1 is given a weight in the optimized hierarchical clustering process to magnify its importance in the clustering process.
(3) Partitioning schemes were created using RF-HC. The random forest technique is utilized to filter the time-domain features, with the theoretically best partition of each node serving as its node label. The filtered features are then time-domain random forested again to establish the significance parameters. The importance parameters are used as an optimized HC strategy to run clustering computations and give an optimized partitioning scheme.

The flow data used in the research was acquired by pressure and flow sensors installed in a city's water supply network monitoring system.

The topological diagram of the 17 monitoring nodes involved in this paper is shown in Figure 2.

## Partition of water supply network based on hierarchical clustering algorithm

The HC method creates a hierarchical clustering tree by assessing the similarity between data points belonging to distinct groups (Bishop and Tipping 1998; Tong 2006). Cluster trees are often created in two modes: top-down split and bottom-up merging. The former is uncommon in our cluster analysis.

The method is based on the actual data gathered by the urban water supply network information monitoring system. The flow data of the 17 nodes in the water supply network within a week has been gathered and clustered in this paper.

The clustering process in this article is as follows.

(1) Treat the flow data set of a single node as a cluster, and generate the initial 17 clusters.
(2) Calculate the various clusters, the distance matrix between two different clusters, and calculate the clusters $c_1$, $c_2$ that are closest to each other.
(3) Merge $c_1$ and $c_2$ to form a new cluster.

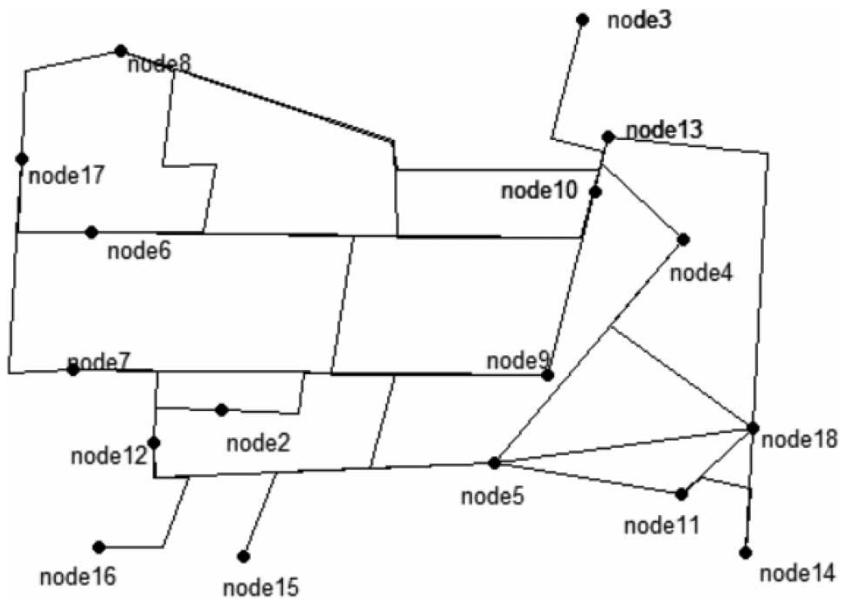


**Figure 2** | Topological structure diagram of monitoring node.

(4) Repeat step 2 and step 3 until a cluster containing all nodes is formed, and a clustering tree is generated at the same time.

The gathered flow data of 17 nodes in a week was used as the input of the HC method. The node data collection interval is one hour, so we got a 168*17 data set. The Ward's distance was used as the clustering algorithm for calculating the distance between clusters. For two different classes $(c_1, c_2)$:

$$D_1 = \sum_{x_i \epsilon c_1} (x_i - \bar{x}_1)^T (x_i - \bar{x}_1) \tag{1}$$

$$D_2 = \sum_{x_i \epsilon c_2} (x_j - \bar{x}_2)^T (x_j - \bar{x}_2) \tag{2}$$

$$D_{12} = \sum_{x_k \epsilon c_1 \cup c_2} (x_k - \bar{x})^T (x_k - \bar{x}) \tag{3}$$

In the formula, $D_1$, $D_2$ and $D_{12}$ are respectively the sum of squared deviations of $(c_1, c_2)$ and the union of $(c_1, c_2)$. $\bar{x}_1, \bar{x}_2, \bar{x}_{12}$ in the formula is the mean value of $c_1, c_2$ and the union of $c_1, c_2$.

Calculate the distance between clusters as:

$$D(c_1, c_2) = D_{12} - D_1 - D_2 \tag{4}$$

The basis is that when the clustering effect is good enough, there must be a small value $D_1, D_2$ in the cluster, and the distance within the cluster is small enough. At the same time, the distance among the clusters must be large enough, that is, the size of $D_{12}$ is big enough. In this way, if the calculated value is large enough, a theoretically optimal global partitioning method based on flow data can be obtained.

The premise is that when the clustering effect is strong enough, there must be a small value $D_1, D_2$ in the cluster and a small distance inside the cluster. At the same time, the distance between the clusters must be great enough, that is, $D_{12}$ must be large enough. If the estimated number is big enough, a theoretically optimum global partitioning technique based on flow data can be derived in this manner.

### Reduces the dimension of coordinate information by principal component analysis

To solve the problem of the overlap of district, two parameters $X(x_1, x_2 \cdots x_i)$; $Y(y_1, y_2 \cdots y_i)$ were given to each node. Which describes the coordinate information of each node on the map, and define a new array $L(X; Y)$. To facilitate the processing of HC, the principal component analysis (PCA) method (Li *et al.* 1999) was used to reduce the two-dimensional location data to one-dimensional and generate a new location label $L_n(X_n)$.

In this paper, the PCA algorithm was implemented based on the eigenvalue decomposition covariance matrix. The main process is as follows:

(1) Remove the mean:

$$x_i^* = x_i - \bar{x}; y_i^* = y_i - \bar{y} \tag{5}$$

$x_i^*$; $y_i^*$ is the newly generated de-averaged coordinate information; $x_i, y_i$ is the original coordinate information; $\bar{x}; \bar{y}$ is the mean value of $X; Y$. Than get a new matrix $L_m(X_m; Y_m)$.

(2) Calculate the covariance matrix:

$$C = \frac{1}{n-1} L_m L_m^T \tag{6}$$

n is the number of samples, which means the number of nodes, and C is a symmetric matrix.

(3) Calculate the covariance eigenvalues and eigenvectors:

$$C = Q \Sigma Q^{-1} \tag{7}$$

Q is a matrix composed of eigenvectors of matrix C, and $\Sigma$ is a matrix whose diagonal elements are eigenvalues.

(4) Sort the eigenvalues in descending order. Select the larger eigenvalue and the corresponding eigenvector, and defined the corresponding eigenvector as $L_n(X_n)$.

### Time-domain features and the change degree of the daily flow data

Extract the features of the gathered flow data. Because the flow data was given in discrete time series, the following time-domain properties should be extracted first: mean, standard deviation, variance, effective value, peak value, crest factor, and kurtosis. Table 1 shows the feature parameters for each node.

A new data set can be formed by the integrated of mean, standard deviation, variance, root mean square, peak value, crest factor, kurtosis, and form factor.

Simultaneously, the single-day change degree $W_c$ of the flow data regarding water has been proposed, taking into consideration the problem of property in the everyday life of the selected district. It reflects the amount of variation in the daily flow of water. Meanwhile, defined $r$ as the threshold value of the ratio of flow rate data change one hour after a specific point to the flow rate at that point. It is assumed that the starting value of $W_c$ is 0. The calculating procedure for $W_c$ for each node's flow data is as follows.

$$W_c = \begin{cases} W_c - 1; \ \dfrac{x_{t+1} - x_t}{x_t} \le -r \\ W_c; \ -r < \dfrac{x_{t+1} - x_t}{x_t} < r \\ W_c + 1; \ \dfrac{x_{t+1} - x_t}{x_t} \ge r \end{cases} \tag{8}$$

Prior to this, some nodes had '0' traffic data when $x_t = 0$, or they had flow data that was excessively big. Because this condition occurs in real life, this data is saved, and the flow data is normalized.

$$x_i^* = \frac{x_i - \mu}{\sigma} \tag{9}$$

Here, $x_i$ is the original flow data, $\mu$ is the mean value of the raw flow data, $\sigma$ is the standard deviation, and $x_i^*$ is the normalized flow data.

Some of $W_c$, when r = 0.2 are shown in Table 2.

The size of the threshold $r$ has a distinct influence on the partition outcome. The value of $r$ directly influences the value of $W_c$, which represents the degree of flow change and hence influences the ultimate clustering outcome.

**Table 1** | Time-domain features of flow data

| | Node2 | Node3 | Node4 | ... | Node18 |
|---|---|---|---|---|---|
| Mean | 0.876 | 12.988 | 215.276 | ... | 504.069 |
| Standard deviation | 13.728 | 9.062 | 73.922 | ... | 129.773 |
| Variance | 188.444 | 82.123 | 5,464.422 | ... | 16,840.994 |
| Root mean square | 13.727 | 15.826 | 227.564 | ... | 520.438 |
| Peak value | 98.474 | 36.381 | 379.623 | ... | 636.057 |
| Crest factor | 7.174 | 2.299 | 1.668 | ... | 1.222 |
| Kurtosis | 5.722 | 2.519 | 2.771 | ... | 2.459 |
| Form factor | 1.602 | 1.219 | 1.057 | ... | 1.032 |

**Table 2** | $W_c$ of each node in a week

| | Node2 | Node3 | Node4 | Node5 | ... | Node17 | Node18 |
|---|---|---|---|---|---|---|---|
| Day1 | −9 | −8 | −6 | −2 | | −8 | −5 |
| Day2 | −5 | −5 | −4 | −1 | | −9 | −5 |
| ... | | | | | | | |
| Day7 | −6 | −4 | −0 | −5 | | −17 | −1 |

## Random forest algorithm to screen time-domain features

Because the Ward's distance algorithm handled all characteristics equally. This would imply that the distance between two clusters might be magnified. To increase clustering reliability, the random forest technique (Ho 1995, 1998; Fang *et al.* 2011) was employed to filter time-domain characteristics. The less significant features would be deleted, and various weights would be assigned to different features. The $W_c$ and $L_n(X_n)$ features, on the other hand, will not be checked; only the time-domain features have been examined. Because they were both more essential than the time-domain elements for the outcome.

Prior to calculation, the above data should be standardized:

$$t_i^* = \frac{t_i - \mu}{\sigma} \tag{10}$$

In the formula, $t_i$ is the time-domain feature parameters of each node, $\mu$ is the original mean value of each parameter of the node, $\sigma$ is the original standard deviation, and $t_i^*$ is the normalized parameter.

The classification label of each node is based on the previous global optimal partition ($A_{11}$, $A_{12}$, $A_{13}$, $A_{14}$), and the district is marked with 0, 1, 2, and 3 in turn.

The construction process of random forest in this article:

(1) From the training set $S(s_1, s_2 \cdots s_{12})$ composed of 12 nodes, there are 12 samples that are replaced to form a new training subset $S_{t1}(s_1, s_2 \cdots s_{12})$.
(2) For $S_{t1}(s_1, s_2 \cdots s_{12})$, three features are extracted from the eight-times domain features without replacement, and the selected features are used as the split features of the nodes on the decision tree.
(3) Repeat n times of step 1 and step 2 to form n subsets $S_{t1}, S_{t2} \cdots S_{tn}$, and form n decision trees, and combine these decision trees to form a random forest.
(4) Input the remaining 5 node data as the test set into the random forest, repeat the classification 5 times, and calculate the importance of each feature.

In this paper, the Gini index is used as the basis for splitting node in step 2. The calculation formula here is:

$$Gini = 1 - \sum_{0}^{3} p_i^2 \tag{11}$$

The parameter $p_i$ is the probability that the sample point belongs to the four categories 0,1,2,3.

The results obtained are shown in Table 3.

Parameters with a significance rate of less than 0.1 are eliminated. Peak value, variance, mean, form factor, and crest factor are more important in flow data grouping.

Further, Random Forest processing has been used again for peak value, variance, mean value, form factor, and crest factor. The weights $\omega_t(0.25, 0.25, 0.2, 0.16, 0.14)$ of the time-domain characteristics were used to determine the new significance rate of peak value, variance, mean, form factor, and crest factor.

**Table 3** | Importance rate of time-domain features of flow data

| Parameter | Importance rate |
|---|---|
| Peak value | 0.183467 |
| variance | 0.180575 |
| Mean | 0.146953 |
| Form factor | 0.123266 |
| Crest factor | 0.103743 |
| Root mean square | 0.093229 |
| Kurtosis | 0.084399 |
| Standard deviation | 0.084368 |

## RESULT AND DISCUSSION

### Partition results based on hierarchical clustering

The following is the outcome of a computation concerning the similarity of flow data collected by various nodes.

The flow data collected by the 17 nodes was analyzed using the HC technique. Figure 3 depicts the clustering findings achieved.

Based on the clustering findings, it is possible to deduce that the city's water supply region, which consists of 17 monitoring nodes, may be separated into four districts: $A_{11}, A_{12}, A_{13}, A_{14}$. When the distance is less than 3,200, the nodes in each district are as given in Table 4.

The flow curve diagram of node in each district is shown in Figure 4(a)–4(d).

Analyzing Figure 4, we can observe that the nodes in Figure 4(c) and 4(d) have relatively comparable alterations, although there are some inaccuracies in Figure (a) (b). To some extent, the partition approach based on flow data has achieved the aim of global optimization, however there are several nodes spread throughout a large geographic range in the same district.

### Partition results of the RF-HC method

The time-domain features of each node (peak value, variance, mean, form factor, crest factor), $W_c$, $L_n(X_n)$ have been integrated, and defined as $C(c_1, c_2 \cdots c_{17})$. Obviously, C is a 16*17 data set, $c_i(x_{i,1}, x_{i,2} \cdots x_{i,t} \cdots x_{i,16})^T$. HC was used again, where the importance rate was used as the weight of the time-domain features to calculate the distance between classes.

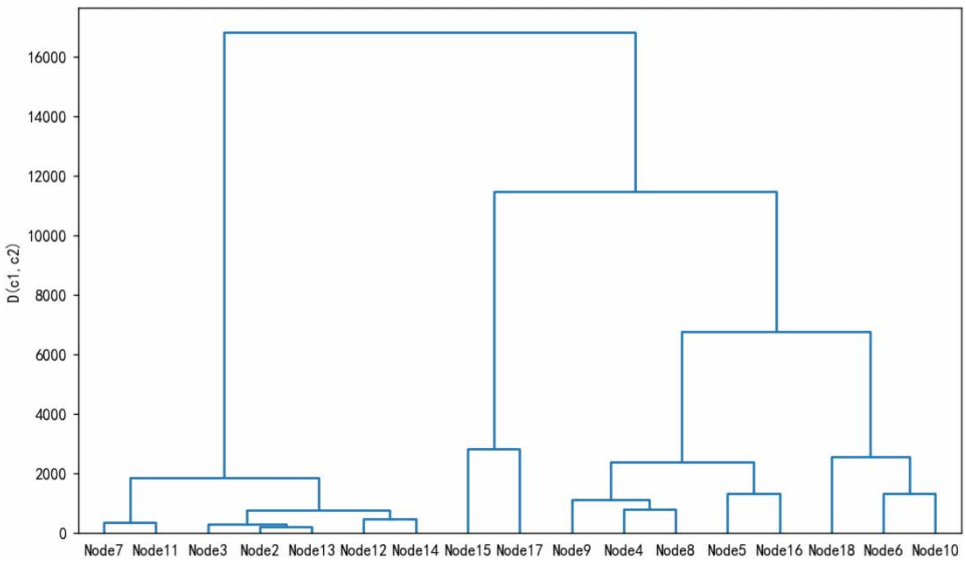


**Figure 3** | Clustering dendrogram based on flow data.

**Table 4** | Hierarchical clustering and partitioning results based on flow data

| District | Node |
|---|---|
| $A_{11}$ | Node2 Node13 Node3 Node14 Node12 Node11 Node7 |
| $A_{12}$ | Node15 Node17 |
| $A_{13}$ | Node4 Node8 Node9 Node5 Node16 |
| $A_{14}$ | Node6 Node10 Node18 |

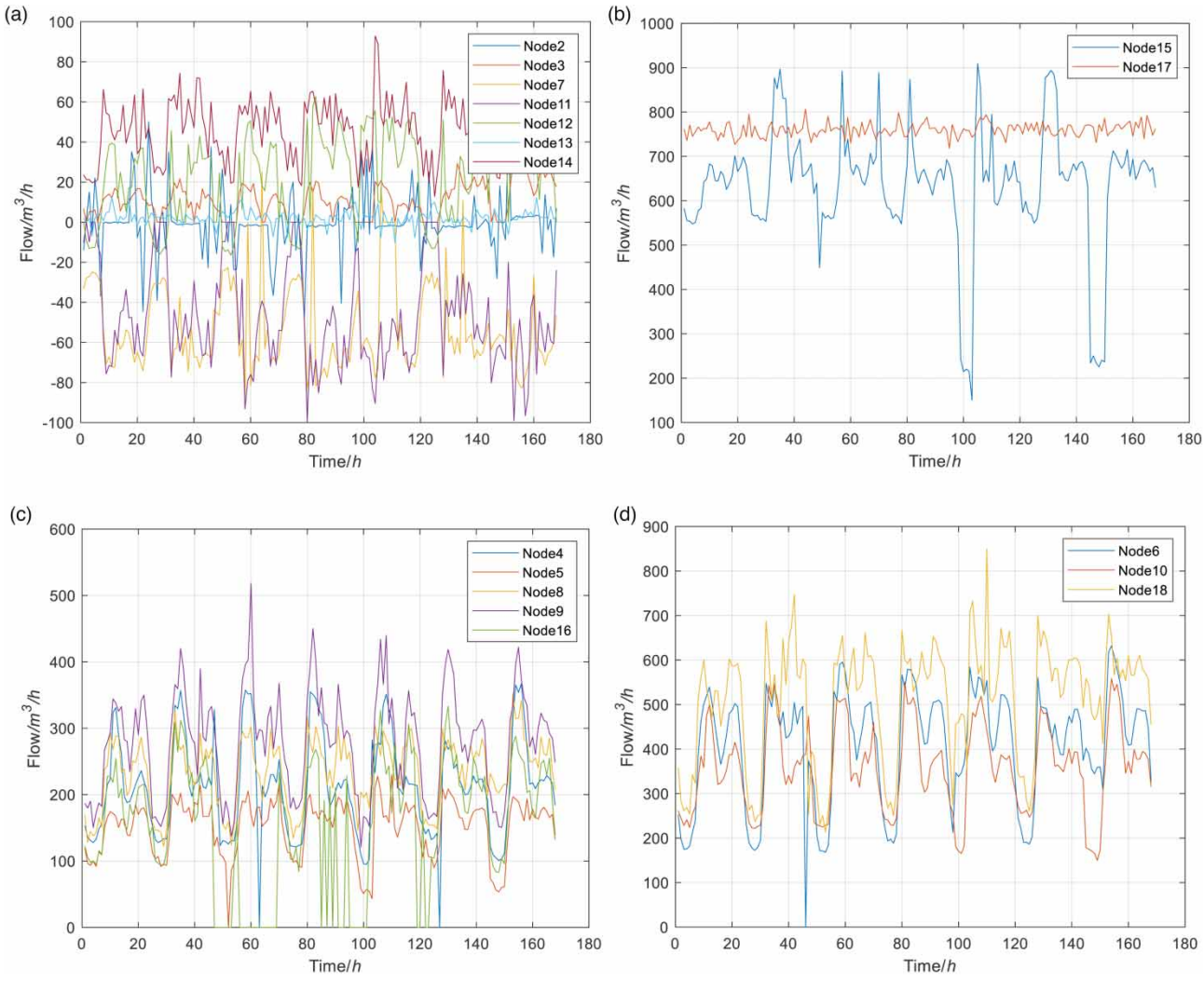


**Figure 4** | Flow curve diagram of each district (a) Flow curve diagram of district $A_{11}$ (b) Flow curve diagram of district $A_{12}$ (c) Flow curve diagram of district $A_{13}$ (d) Flow curve diagram of district $A_{14}$.

And the other parameters were given a weight value of 1 to keep it unchanged.

$$\omega_{i,t} = \begin{cases} 0.25; t = 1 \\ 0.25; t = 2 \\ 0.2; t = 3 \\ 0.16; t = 4 \\ 0.14; t = 5 \\ 1; else \end{cases} \tag{12}$$

where $\omega_{i,t}$ represents the weight parameter; defined the weight matrix as $\omega(\omega_{i,1}, \omega_{i,2} \cdots \omega_{i,t})$ then the new distance could be calculated as:

$$D_1 = \sum\nolimits_{x_i \epsilon c_1} (x_i - \bar{x}_1)^T (x_i - \bar{x}_1) \tag{13}$$

$$D_2 = \sum\nolimits_{x_i \epsilon c_2} (x_j - \bar{x}_2)^T (x_j - \bar{x}_2) \tag{14}$$

$$D_{n12} = \sum_{x_k \epsilon c_1 \cup c_2} ((x_k - \bar{x})\omega)^T ((x_k - \bar{x})\omega) \tag{15}$$

$$D_n(c_1, c_2) = (D_{n12} - D_1 - D_2) \tag{16}$$

When r = 0.1; r = 0.2; when r = 0.5, the weighted Ward's method was used in the clustering method. The result is shown in Figure 5(a)–5(c).

The figure shows the nodes of a district where four branch nodes appear, when the distance is 4.5 with r takes different sizes. And the result is shown in Table 5.

Each branch node is clearly split into related sections, either by resemblance or by proximity. When r = 0.5, this partition and the partition in which node 10 is located have some overlap. The findings are satisfactory for r = 0.1 and r = 0.2 for the partitions containing nodes 14, 15, and 16, and there are no issues with overlapping pipe networks. However, with r = 0.1, node 2 causes partition $A_{22}$ (including nodes 6, 8, 15, and 16) to have a pipe network overlap, indicating that the zoning is unclear. It is obvious that using $0.1 < r < 0.5$, r ≈ 0.2 when the partitioning impact is better, take r = 0.2 here.

After adjusting for the effect of the difference in $W_c$, the new dataset was determined to be just 13*17. In comparison, the same dataset was partitioned using a Pearson correlation analysis-based partitioning algorithm for any two nodes $X$, $Y$ with their correlation coefficients $r_{X,Y}$.

$$r_{X,Y} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}} \tag{17}$$

where $\bar{X}$, $\bar{Y}$ is the sample mean of $X$, $Y$.

The calculated correlation matrix of size 17*17 was converted to a grey-scale graphical representation, as shown in Figure 6 as a grey-scale schematic of the correlation coefficients for nodes 2–18.

When $r_{X,Y}$ is between 0.8 and 1, the nodes are strongly positively correlated; when $r_{X,Y}$ is between 0.5 and 0.8, the nodes are moderately favorably correlated. To fulfill real-world control needs, only positive correlation is employed as the foundation for partitioning here, and the two nodes with the highest $r_{X,Y}$ values between them are joined to establish a new partition. Because of the phenomena of overlapping pipe networks across partitions, the partitioning of these 17 nodes is accomplished by removing them based on the connection of the pipe network. Table 6 shows the state of each partition node before and after subjective rejection.

As it can be observed, the correlation for node 8 is too low to fit into the other partitions, resulting in a substantial overlap of partitions $A_{31}$,$A_{32}$. Figure 7 shows the flow curve diagrams of nodes in each district ($A_{21}$, $A_{22}$, $A_{23}$, $A_{24}$) when the distance is around 5 when r = 0.2:

Figure 7(a) and 7(b) shows that, when compared to the correlation analysis-based technique, Node 2, although being physically adjacent to Nodes 6, 12, and 15, obviously has a distinct traffic pattern than the other nodes, where RH-partitioning HC's results are clearly superior. The two approaches provide substantially the same partitioning results in Figure 7(c)–7(e). Analysis of the data from node 11, where the flow fluctuation is not identical to that of many other nodes, indicates that there may be an issue with the data collection equipment, resulting in what seems to be a lag in the data from node 11. Both approaches divide nodes 7 and 17 into the same partition. Nodes 2, 6, and 8 have similar trends in Figure 7(g), and nodes 3 and 13 have similar trends to the other nodes in most of their trends, for example: from 22 to 30, there is a rising and then falling trend in this partition, and from 33 to 45, there is also a falling and then rising trend, but nodes 3

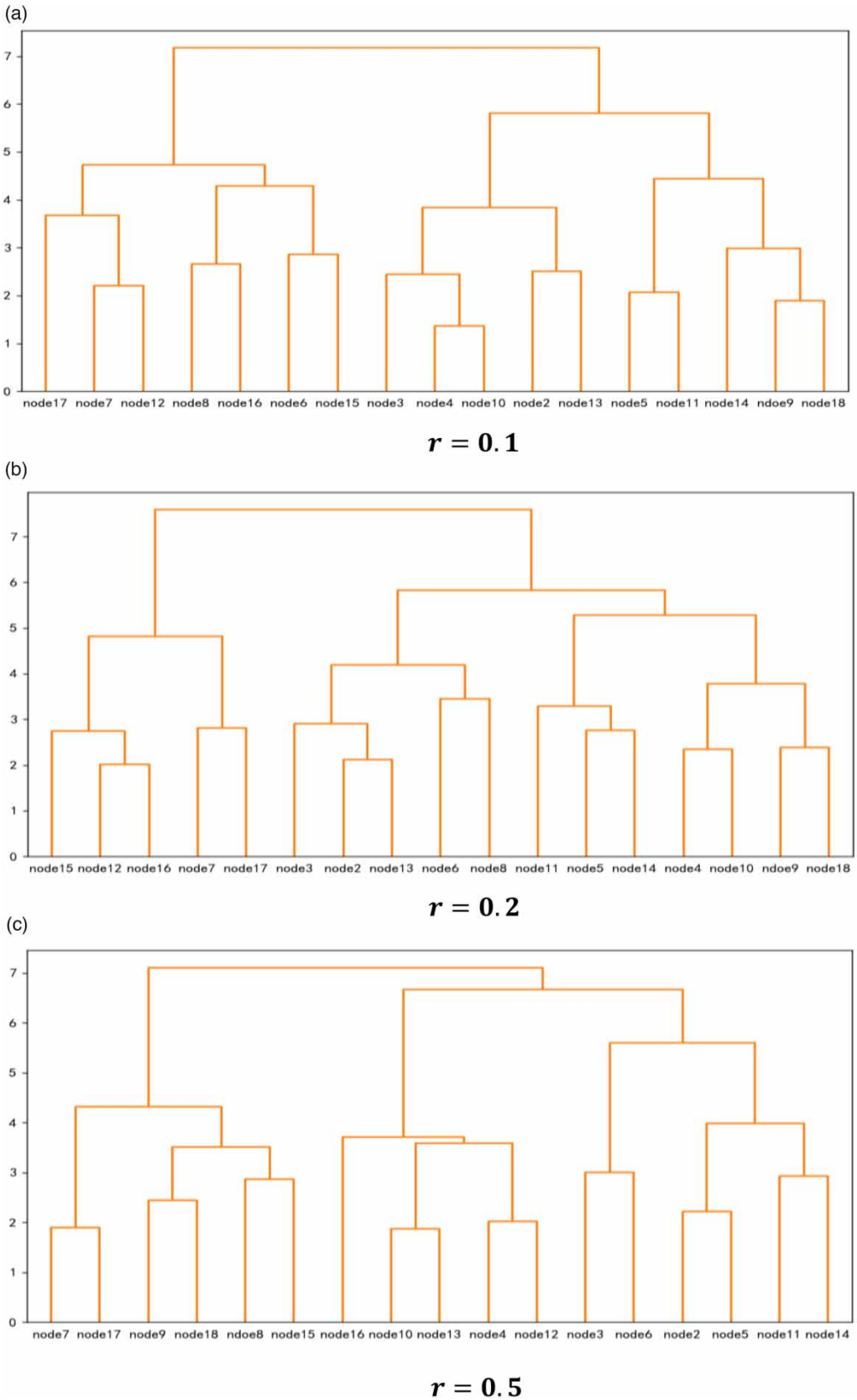


**Figure 5** | Clustering results of weighted Ward's method (a) $r = 0.1$ (b) $r = 0.2$ (c) $r = 0.5$.

**Table 5** | The partition node where the branch node is located when r takes different sizes

| | $r = 0.1$ | $r = 0.2$ | $r = 0.5$ |
|---|---|---|---|
| District with Node3 | Node2 Node3 Node4 Node10 Node13 | Node2 Node3 Node6 Node8 Node13 | Node3 Node6 |
| District with Node14 | Node5 Node11 Node14 | Node5 Node11 Node14 | Node5 Node11 Node14 |
| District with Node15 | Node12 Node15 Node16 | Node12 Node15 Node16 | Node7 Node8 Node15 Node9 Node17 Node18 |
| District with Node1 | Node15 Node12 Node16 | Node15 Node12 Node16 | Node4 Node10 Node12 Node13 Node16 |

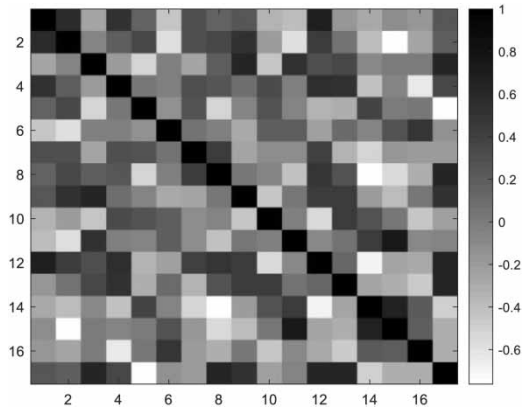


**Figure 6** | Grayscale diagram of Pearson correlation coefficients for each node.

and 13 can be separated from partition A 23. Of course, nodes 3 and 13 can be split from partition A 23. Node 10 in Figure 7(h) has most of the temporal similarities with 3 and 13, but it cannot be separated again.

Finally, nodes 12, 9, 5, 7, and 6 on the main pipe of water supply network were chosen as regulating nodes for the partition. They can aid in real pressure regulation.

It is obvious that with limited data samples, the partitioning method based on RF-HC gives more accurate results than the method based on Pearson correlation. It also decreases model complexity by avoiding the need to re-plan partitioning due to overlapping pipe networks.

## CONCLUSION

In comparison to the original HC approach based on flow data and the partitioning findings based on Pearson correlation analysis, the RF-HC approach gives better partitioning results and the model is less complex. It ensures the similarity of

**Table 6** | Partitioning results based on Pearson correlation analysis

| Partition | Each partition node before elimination | Each partition node after elimination |
|---|---|---|
| $A_{31}$ | Node6 Node16 Node12 Node15 | Node6 Node16 Node12 Node15 Node2 |
| $A_{32}$ | Node8 Node9 Node11 Node18 | Node4 Node9 Node18 Node14 Node5 Node11 Node8 |
| $A_{33}$ | Node2 Node3 Node5 Node13 | Node13 Node3 Node10 |
| $A_{34}$ | Node7 Node17 | Node7 Node17 |
| $A_{44}$ | Node4 Node10 | – |

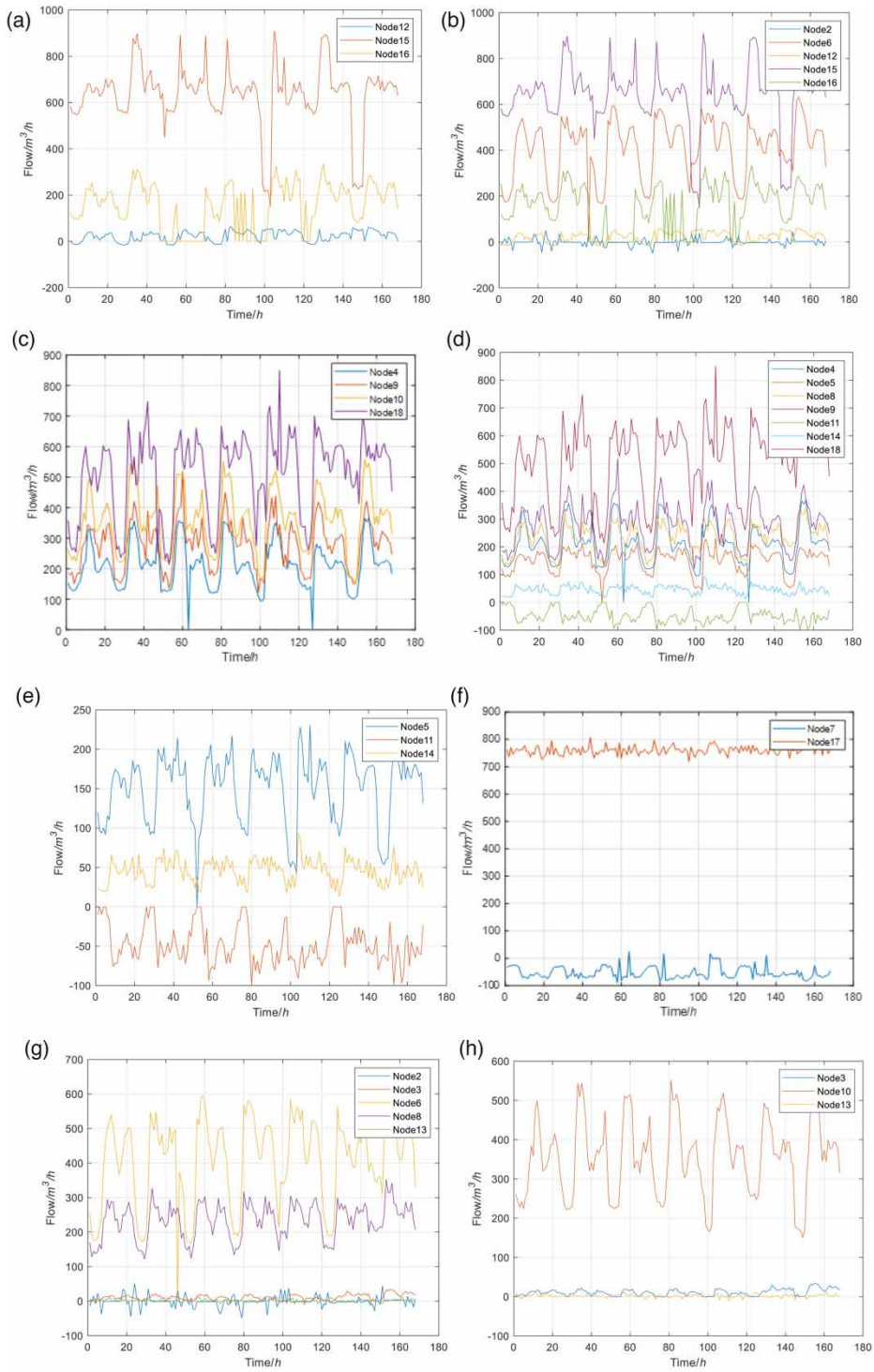


**Figure 7** | New flow curve diagram of each partition (a) Flow curve diagram of district $A_{21}$ (b) Flow curve diagram of district $A_{31}$ (c) Flow curve diagram of district $A_{22}$ (d) Flow curve diagram of district $A_{32}$ (e) Flow curve diagram of district $A_{24}$ (f) Flow curve diagram of district $A_{25}/A_{34}$ (g) Flow curve diagram of district $A_{23}$ (h) Flow curve diagram of district $A_{33}$.

the flow variation pattern of the nodes in the partition. Also, it can effectively help the actual regulating treatment of the problem of inconsistent flow and pressure variation in the demand about each node. Surely, the value $w_c$ for the degree of change in the flow on a single day, as well as the judgement threshold r, may be optimized further.

It should be noted that this paper proposes a water supply network partitioning strategy with a small amount of data, which is convenient and accurate. However, the paper has a limitation in that it is not as efficient in solving for very large pipe network models as the new machine learning technique.

## ACKNOWLEDGEMENTS

We would like to give thanks for the funding of the project: Demonstration project of water supply safety guarantee and optimized operation of Chaohu pipe network, and the teachers and students at Anhui Jianzhu University, who helped to completed this project.

## DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

## REFERENCES

Alvisi, S. & Franchini, M. 2014 A heuristic procedure for the automatic creation of district metered areas in water distribution systems. *Urban Water Journal* **11** (2), 137–159.

Bin, L., Guo-li, Z., Jin-xu, N. & Hao, J. 2018 An overview of optimized DMA zonal methods in water distribution system. *Journal of Guangdong University of Technology* **35** (2), 19–27.

Bishop, C. M. & Tipping, M. E. 1998 A hierarchical latent variable model for data visualisation. *IEEE TPAMI* **20** (3), 281–293.

Diao, K., Zhou, Y. & Rauch, W. 2013 Automated creation of district metered area boundaries in water distribution systems. *Journal of Water Resources Planning &Management* **139** (2), 184–190.

Fang, K., Wu, J., Zhu, J. & Xie, B. 2011 A review of technologies on random forests. *Journal of Statistics and Information* **26** (03), 32–38.

Ho, T. K. 1995 Random decision forest. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, Canada, Vol. 8, pp. 278–282.

Ho, T. K. 1998 The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (8), 832–844.

Jun, L. & Guoping, Y. 2012 Integration of graph theory and GIS for district metered areas of water distribution systems. *Journal of Tongji University(Natural Science)* **40** (12), 1863–1869.

Li, Y., Zeng, Z., Zhang, M. & Yu, S. 1999 Application of primary component analysis in the methods of comprehensive evaluation for many indexes. *Journal of Hebei University of Technology* **01**, 96–99.

Paola, F. D., Fontana, N., Galdiero, E., Giugni, M., Uberti, G. S. D. & Vitaletti, M. 2014 Optimal design of district metered areas in water distribution networks. *Procedia Engineering* **70** (70), 449–457.

Tong, Z. 2006 Research of clustering algorithm based on hierarchical and partitioning method. *Computer Engineering and Applications* **08**, 178–180.

Wenqi, Y., Xiaoming, Z., Zhaokai, X. & Qing, Y. 2012 Research on regional water distribution network planning with district meter area method. *Water & Wastewater Engineering* **48** (07), 98–102.

First received 24 August 2021; accepted in revised form 4 February 2022. Available online 18 February 2022