Available online at www.sciencedirect.com
ScienceDirect
Procedia Engineering 186 (2017) 152 – 159
Procedia Engineering
www.elsevier.com/locate/procedia

XVIII International Conference on Water Distribution Systems Analysis, WDSA2016

Spectral Clustering and Genetic Algorithm for Design of District Metered Areas in Water Distribution Systems

Rui Han[a], Jun Liu[a]*

[a]*College of Civil Engineering and Mechanics, Yanshan University, Qinhuangdao 066004, China*

Abstract

District metered areas (DMAs) is an innovative layout and control technology to improve efficiency for water distribution systems (WDSs), there are many factors to be considered in DMAs design to reduce the adverse impact due to closure of valves. A new methodology for DMAs design is proposed based on complex network spectral clustering algorithm and graph theory. Water distribution network is regarded as undirected graph, which is mapped by weighted topology matrix. And normalized Laplacian matrix can be deduced to calculate the nontrivial eigenvectors. The *k*-means and genetic algorithm are used to determine clusters to minimize the sum of squared distance error between nodes in clusters and their centroids in Euclidean space. Eigenvector centrality is adopted for importance analysis of nodes in order to ascertain the location of meter of each DMA, which should be in the shortest path among the source and the central node in each DMA, and the other pipes between different DMAs will be installed with valves, at last the design of DMAs is achieved. And a real water distribution system is tested to ascertain the feasibility of the proposed method.

Peer-review under responsibility of the organizing committee of the XVIII International Conference on Water Distribution Systems

Keywords: Water distribution systems; District meter areas; Spectral clustering; Genetic algorithm

1. Introduction

District metered areas (DMAs) is an effective tool to manage water loss. After several years development, there are some empirical summary to guide DMAs design considering many criteria including size, pressure, leakage level,

* Corresponding author. Tel.: +86-18830454804
E-mail address: sdlj2008@163.com

Peer-review under responsibility of the organizing committee of the XVIII International Conference on Water Distribution Systems
doi:10.1016/j.proeng.2017.03.221

water quality, cost, reliability, and so on [1]. In other words, the adverse impact of DMA should be minimized while the benefit can be maximized. For DMAs design, the important issues are how to define DMAs boundary and where to locate meters. Due to the complexity, for a long time the DMAs design mainly rely on try-and-error method with hydraulic simulation to validate corresponding performance [2].

Recently researchers start to investigate the subject to automatically create DMAs. With regard to graph theory, Depth-first search (DFS), bread-first search (BFS) and spanning tree are the main methods to detect structure for DMAs design. In particular based on the principle of directly supplied by water source, the scope of different sources can be determined by DFS[3], and supposed the nodes at transmission mains are source nodes, water supply scope at scale of DMA size can be obtained using BFS [4]. Graph partitioning is an alternative to DMAs design, the aim is to obtain DMAs with equivalent sizes while to keep less boundary pipes between DMAs [5, 6]. As a metric to quantify the degree of uniformity of a particular partition, modularity-based DMAs design methods have been applied to network treated as a simple graph [7], or weighted graph [8].

Most approaches focus on obtaining DMAs with similar size, neglecting the topological structure. DMAs are maintained by closure of many boundary pipes, inappropriate pipe closure leads to low pressure, high water age. Therefore how to define the DMAs boundary with less pipe closure needs to be investigated deeply. The aim of this paper is to find the inherent clustering structure with less interaction for DMAs design to minimize the impact of pipe closure on water distribution performance.

Nomenclature

$\mathbf{W}$	adjacency weighted matrix of graph
$\mathbf{D}$	degree matrix of graph
$\mathbf{L}$	Laplacian matrix of graph
$\mathbf{L}_{sym}$	normalized Laplacian matrix of graph
$\mathbf{E}$	corresponding matrix of $\mathbf{L}_{sym}$
C_i	ith nodes cluster
c_i	centroid of C_i
SSE	sum of squared error
$\mathbf{Y}$	nontrivial eigenvectors selected
y_i	row vector of $\mathbf{Y}$
λ	largest eigenvalue of matrix

2. Methodology

The proposed method consists of four main steps including:

- Data input: Determine the number of DMAs to be created, select pipe weight to differentiate importance from hydraulic simulation, represent topology of graph.
- Matrix calculation: Construct Laplacian matrix, calculate eigenvectors.
- Spectral clustering: *k*-means and GA to find the best node clusters, which correspond to DMAs.
- Meter location: According to engineering experience, it is reasonable to install meter at supply point of each DMAs to measure flow.

2.1. Topology representation of water distribution network

Water distribution network (WDN) can be represented by a graph whose edges and vertices are the pipes and nodes, respectively. Given an undirected graph with node set , we assume that the graph is weighted, its adjacency matrix is defined to be constant matrix， whose ijth entry is if nodes and are connected, and is 0 otherwise. As G is undirected graph, adjacency matrix is always symmetrical. The sum of the ith row (or ith column) equals the degree of the ith node, k_i. The degree matrix **D** is the diagonal matrix with the node degrees along its diagonal. The graph Laplacian matrix is defined as follows.

$$\mathbf{L} = \mathbf{D} - \mathbf{W}$$

The element of the Laplacian matrix are

$$L_{ij} = \begin{cases} k_i & if \quad i = j \\ -w_{ij} & if \;\; i \neq j, \; v_i \; adjacent \; v_j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

In order to produce DMA within certain size, minimize weights between communities and avoid small communities having few nodes simultaneously, normalized cut method is used based on normalized graph Laplacian which is defined as follows:

$$\mathbf{L}_{sym} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I}\text{-}\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{E} \tag{2}$$

According to matrix properties, $\mathbf{L}_{sym}$ and **E** have the same eigenvectors, sum of both eigenvalues equals one. Solution to smallest eigenvalue of $\mathbf{L}_{sym}$ is converted to solve largest one of **E**, because it is more efficient to calculate the largest eigenvalue of a matrix and the corresponding eigenvector. Those eigenvalues of **E** less than one are called nontrivial eigenvalues. And nontrivial eigenvectors corresponding to the nontrivial eigenvalues have the partition information. When a graph has a clear partition, only one nontrivial eigenvector is enough due to different characteristic structure. However most of WDN has no clear cluster partitioning, a few eigenvectors are needed for clusters [9]. And *k*-way partition via *k*-means can be used to generate clusters, which is introduced below in detail.

2.2 Spectral clustering

Table 1 DMA-oriented normalized spectral clustering algorithm

Number m of clusters to construct.

1: Construct a weighted adjacency matrix $\mathbf{W} \in R^{n \times n}$ of WDN graph.

2: Compute the normalized Laplacian **E**.

3: Compute the eigenvectors of **E**, and select the first r nontrivial eigenvectors.

4: Let $\mathbf{X} \in R^{n \times r}$ be the matrix containing clusters information.

5: Form the matrix $\mathbf{Y} \in R^{n \times r}$ from *X* by normalizing the rows to norm 1, that is $Y_{ij} = X_{ij} \Big/ \sqrt{\sum_j X_{ij}^2}$.

6: For $i = 1, \cdots, n$, let $y_i \in R^r$ be the vector corresponding to the *i*th row of **Y**.

7: Cluster the points (y_i) $i = 1, \cdots, n$ with the *k*-means algorithm into clusters $C_1, \cdots, C_m$

- a: arbitrarily choose m points from Y as the initial cluster centers;
- b: repreat
- c: (re)assign each point to the cluster to which the point is the most similar, based on the mean value of the points in the cluster;
- d: update the cluster means, that is, calculate the mean value of the points for each cluster;
- e: until no change

8: Clusters with

Partition information is embedded in the eigenvectors, components corresponding to nodes belonging to the same communities are still strongly correlated, in each eigenvector, there are similar values among themselves. *k*-means method is used to identify the communities. Suppose data set, *Y*, contains N nodes information. Partitioning method distributes the nodes in *Y* into m clusters, , that is, and . Of all the nodes in , centroid

c_i is used to represent that cluster, the centroid can be defined by the mean of nodes assigned to the cluster.

$$c_i = \frac{1}{m_i} \sum_{y \subset C_i} y \tag{3}$$

The quality of cluster C_i can be measured by the sum of squared error (*SSE*) between all the nodes in C_i and the centroid c_i, defined as

$$SSE = \sum_{i=1}^{m} \sum_{y \subset C_i} dist\left(y, c_i\right)^2 \tag{4}$$

where SSE is the sum of squared error for all nodes in the data set; y is the point in space representing a give node; and c_i is the centroid of cluster C_i (both y and c_i are multidimensional) . Normalized *k*-way spectral clustering algorithm is provided below. Trivial eigenvector must be excluded because there is no information for cluster.

2.3 Genetic Algorithm

k-means method is a local search algorithm to converge to a local optimum, and the results may depend on the initial random selection of cluster centers. To obtain good clusters, it is common to run the *k*-means algorithm multiple times with different initial cluster centers, or to generate cluster centers with uniform distribution. In this paper global search method Genetic Algorithm (GA) is combined with *k*-means to converge to global optimum.

As discussed above, centroid is the representation of cluster, the solution to clusters is to find each centroid corresponding to a cluster. For GA in a generation, k centroids is composed of an individual. The objective of GA can be formalized as follows

$$\min \ OF = SSE \tag{5}$$

The scope of real number in eigenvector is between -1 and 1, and binary-coded GA is used. Rank-based fitness assignment is used, which assumes a selective pressure of 2 and linear ranking, giving the best individual a fitness value of 2 and the least fit individual a fitness value of 0. During evolution, some centroids in the individual may have no node, which means there is null clusters corresponding to the centroids, they will be kept unchanged. All the centroids of real clusters will be updated by *k*-means before next evolution.

2.4 Eigenvector centrality for meter location

Obviously the nodes in WDN have different importance, those nodes in truck mains or connected to much more nodes than others have more importance. In complex network theory degree centrality is a measure for node importance, the more adjacent nodes it has, the more important it is. But for WDN, nodes have different functions due to their location in water distribution system, for example, the node in truck main may be more important than the node in distribution pipe, that is to say a vertex's importance in a network is increased by having connections to other vertices that are themselves important. This is the concept behind eigenvector centrality, which satisfies the following requirement:

(6)

x is the leading eigenvector of adjacency matrix .This is the eigenvector centrality[10].

After clusters and centrality are determined, the boundary and important nodes of DMAs are defined. From the point of management, there are only a few inlets with meters, other boundary pipes will be cut by isolated valves.

Then the main task is to ascertain the location of meters form the boundary pipes. Actually there are main supply path from water source to DMA, it is highly possible the path flows be through the most important nodes in DMA, therefore the meter location can be identified by finding the boundary pipe through the shortest path among water source and most important node. The Dijkstra algorithm [11] has been chosen in this paper. If there are low-pressure nodes, another meter location should be selected to improve pressure to the most extent.

3. Case study

In order to prove the effectiveness of the proposed method, the approach was applied on a real case study, the Colorado Spring WDS. The network is skeletonized by deleting those pipe with diameter no more than 6 in. The main source is Reservoir 3001, therefore pump station is also deleted. The skeletonized network is illustrated in Fig. 1, and the basic characteristics are reported in Table 2. Transmission mains are composed of pipe diameters with 16 in and 24 in. and distribution mains are the pipe with 8 in and 12 in.

Table 2 Characteristics of water distribution network

Attribute	Value
Number of nodes	1100
Number of links	1302
Number of valves	3
Number of reservoirs	1
Number of pumps	1
Pipe diameters(in)	8;12;16;24
Peak demand(GPM)	6987.83

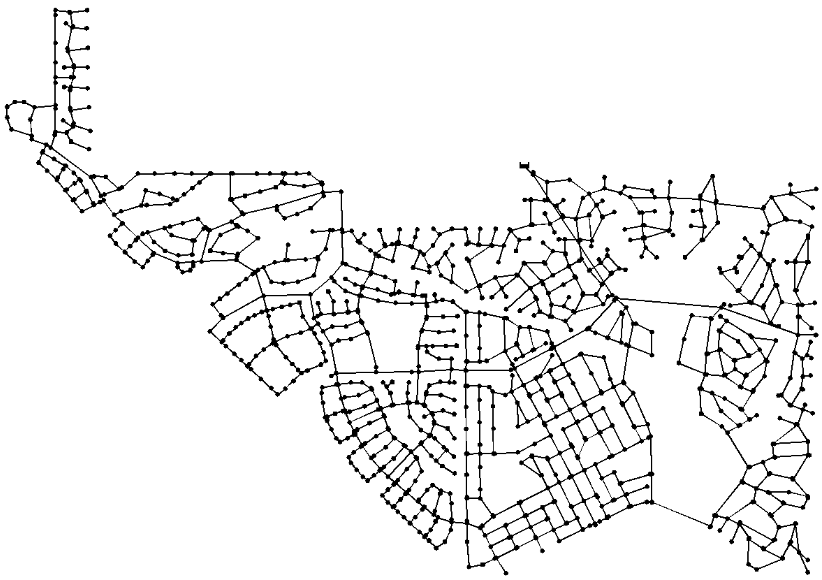

Fig.1 Case study network of Colorado Spring WDS

The hydraulic simulation was performed using EPANET2. For the sake of simplicity, we carried out steady-state simulation, and pipe flows, q_{ij}, were selected as weights of the adjacency matrix. In the proposed method, pipe weight reflects the connection strength between adjacent nodes, therefore flow direction is less important, pipe weight for the undirected graph matrix is $w_{ij} = q_{ij}$.

Based on the adjacency matrix, normalized graph Laplacian matrix were formed, then its nontrivial eigenvalues and corresponding eigenvectors were obtained by matrix calculation. Three nontrivial eigenvectors corresponding first three largest nontrivial eigenvalues are displayed in Fig. 2. It can be seen that there is a gathering center containing most nodes information, obviously it is impossible to partition them easily, while those circumjacent nodes who have larger or smaller values can be clustered simply.

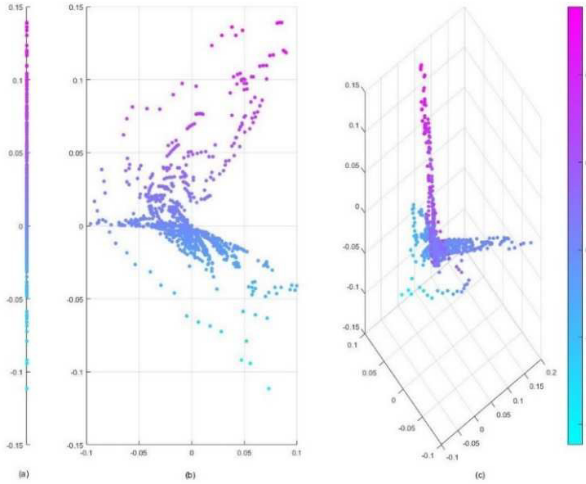

Fig. 2 Nontrivial eigenvectors of Colorado Spring WDS: (a) node value distribution in one nontrivial eigenvector, (b) node values distribution in two nontrivial eigenvectors, and (c) node values distribution in three nontrivial eigenvectors

According to the results from Scibetta, et al [7], the number of DMAs is 12 because of its higher modularity reflecting better aggregations. Due to the difficulty in clustering, *k*-means and GA were used to explore the clusters. A fifty generations were carried out with a population composed of fifty individuals, the crossover was 0.7, and mutation was set to be the ratio of 0.7 and chromosome length. The effectiveness of the proposed approach is shown in Fig.3, in which the best and mean value of objective function for each generation are displayed with snowflake dot and line, respectively. It is possible to observe that after 25-30 generations, the proposed method is able to rapidly find a good solution.

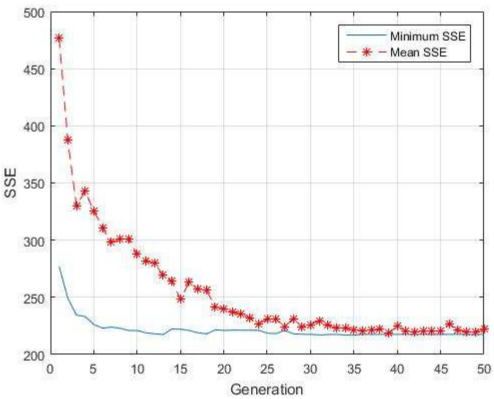

Fig.3 SSE and number of GA generations for 12 DMAS

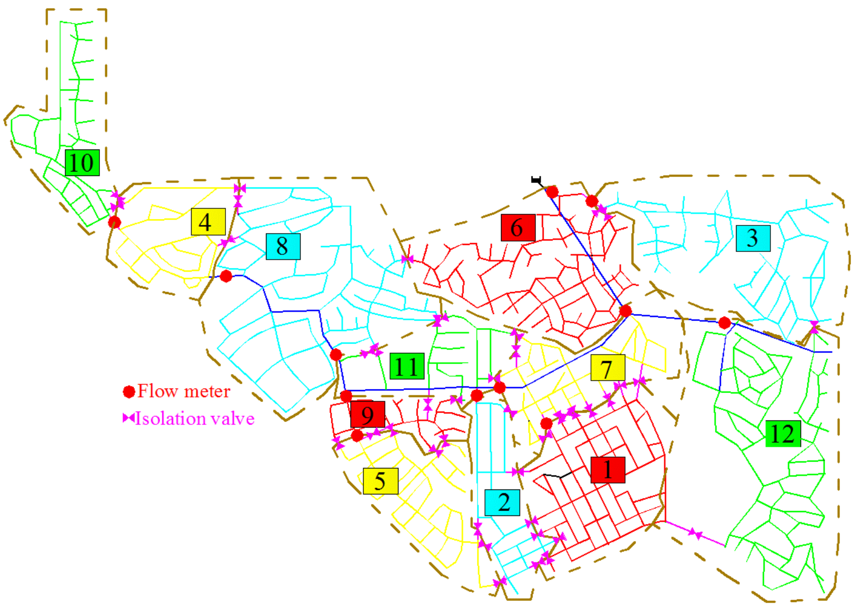

Fig.4 Division of network into 12 DMAs

Table 3 Size of DMAs and number of pipe cuts to form DMA boundary

DMA index	Number of nodes	Demand (GPM)	Number of pipe cuts
1	102	860.25	16
2	43	362.66	11
3	92	759.06	4
4	60	237.30	7
5	104	431.92	9
6	135	1015.83	5
7	68	556.65	16
8	179	711.89	10
9	52	209.14	9
10	82	462.53	3
11	49	241.98	13
12	135	1138.60	3

On the basis of spectral clustering steps described above, the network was partitioned into 12 DMAs. By setting $w_{ij} = \Delta H_{ij} / Q_{ij}$, supply point (inlet) of each DMA was determined. The obtained DMAs with boundary valves connecting different DMAs and supply point (inlet) of each DMA with meter are shown in Fig.4 and Table 3. All DMAs, excluding 1, 3, 5 and 10, which are located further from the reservoir, have the direct connection to the transmission mains. Those areas with highly loop structure are more difficult to create DMAs, hence must close much more boundary valves, as shown in 1, 2, 7, and 11.

The size of DMAs has a large variable ranges of demand, but they are within the desired demand range. The reason is in that the proposed method tends to create DMAs with less cuts with lower impact to the performance, therefore resulted demand per DMA is different. However the number of DMAs to be obtained is based on the same demand per DMA predetermined.

Fig.5 indicates the pressure distribution in each DMA. The approach indentifies DMAs with different pressure scope with corresponding mean and median pressure. 2, 8 and 12 have larger pressure scope, it may be due to the higher elevation variation. The only required information in this paper are network topology and flows, the terrain elevation variation is not considered.

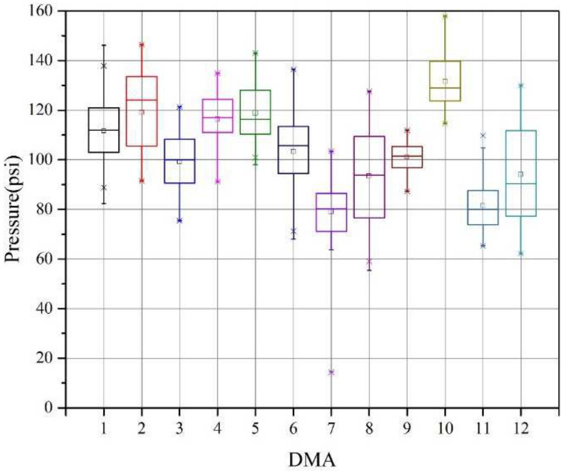

Fig.5 Pressure distribution in each DMA. The box edges are the 25th and 75th percentiles.

4. Conclusions

This paper describes a spectral clustering method for developing DMAs configuration. The approach combines graph theory and cluster algorithm and is formulated as optimization problem to obtain the optimal centroid corresponding to a cluster(DMA), eigenvector centrality and Dijkstra algorithm are used to identify supply point of each DMA, and meter location. The method is tested for large-scale network, which shows its efficiency.

References

[1] Morrison, J., Tooms, S., and Rogers, D. DMA management guidance notes. IWA Publication, London, 2007.
[2] Grayman, W. M., Murray, R. and Savic, D. A. Effects of redesign of water systems for security and water quality factors. Proc. World Environmental and Water Resources Congress, ASCE, Reston, VA, 2009, pp. 504-114.
[3] Di Nardo, A., Di Natale, M., Santonastaso, G. F., Tzatchkov, V. G. and Alcocer-Yamanaka, V. H. Water network sectorization based on graph theory and energy performance indices. J. Water. Resour. Plann. Manage., 2014, 140(5):620-629.
[4] Ferrari, G., Savic, D., Becciu, G. Graph-theoretic approach and sound engineering principles for design of districted metered areas. J. Water. Resour. Plann. Manage., 2014, 140(12): 04014036.
[5] Di Nardo, A., Di Natale, M., Santonastaso, G. F. and Venticinque, S. An automated tool for smart water network partitioning. Water Resour. Manage., 2013, 27(13): 4493-4508.
[6] Perelman, L. S., Allen, M., Preis, A., Iqbal, M., and Whittle, A. Flexible reconfiguration of existing urban water infrastructure systems. Environmental Science & Technology, 2015, 49(22): 13378−13384.
[7] Scibetta, M., Boano, F., Revelli, R. and Ridolfi, L. Community detection as a tool for complex pipe network clustering. EPL, 2013, 103(48001): 1-6.
[8] Diao, K., Zhou, Y. and Rauch, W. Automated creation of district metered area boundaries in water distribution systems. J. Water. Resour. Plann. Manage., 2013, 139(2): 184-190.
[9] Capocci, A., Servedio, V.D.P., Caldarelli, G. and Colaiori, F. Detecting communities in large networks. Physica A, 2005, 352: 669-676.
[10] Bonacich, P. F. Power and centrality: A family of measures. Am. J. Sociol., 1987, 92: 1170-1182.
[11] Dijkstra, E. M. A note on two problems in connections with graph. Numerische Mathematik, 1959, 1: 269-271.