Pressure transfer modeling for an urban water supply system based on Pearson correlation analysis

Bin Zhu and Jingqi Yuan

ABSTRACT

This paper presents an approach to modeling the water pressure transfer among nodes in an urban water supply network for the purpose of pressure control. The network is divided into different sub-networks based on the Pearson correlation analysis of the nodal pressure measurements. The Pearson correlation analysis is performed to find out the set of nodes, whose water pressures are highly correlated, and thus a corresponding sub-network is formulated. As a case study, 47 sub-networks are recognized for a region with an area of 250 km^2 and 77 nodes in total. For each sub-network, a linear model is constructed to quantify the pressure transfer. The output of the model is the pressure estimate for the node of our interest which is called the *center node*. The rest of the nodes in the sub-network are called the *correlated nodes* of the center node, and the pressure measurements at the correlated nodes constitute the input to the model. The average relative error of the model is less than 3%. A pressure regulating method based on the model is proposed and tested numerically.

Key words | network division, Pearson correlation coefficient, pressure transfer model, sensor network, water supply system

Bin Zhu
Jingqi Yuan (corresponding author)
Department of Automation,
Shanghai Jiao Tong University,
800 Dongchuan Lu,
200240 Shanghai,
China
and
Key Laboratory of System Control and Information Processing,
Ministry of Education of China,
800 Dongchuan Road,
200240 Shanghai,
China
E-mail: *jqyuan@sjtu.edu.cn*

INTRODUCTION

Modeling for an urban water supply system is the basis of water pressure control, the establishment of emergency (e.g., pipe bursts) response mechanism, and the optimal scheduling of pump stations. The hydraulic model has different formulations which include flow equations at nodes in the network, the head loss–flow relationships for individual pipes and head losses around closed loops or for paths between fixed head nodes in the network (Simpson & Elhay 2009). Based on the hydraulic model, pump scheduling models are formulated and different pressure control methods are proposed (Germanopoulos & Jowitt 1989; Jowitt & Germanopoulos 1992; Ormsbee & Lansey 1994; Yu *et al.* 1994). There exist quite a few commercial or open source softwares such as Infoworks WS, Mike Urban, Pipe2000, EPANET, etc. for hydraulic simulations. However, the hydraulic model requires a detailed system layout with exact parameters for every network component (Kritpiphat *et al.* 1998; Rossman 2000). On the shortage of detailed system information, such a model is very hard to construct. An alternative is to install sensors on the pipe lines to form a sensor network. With data in hand, data mining methods are applied to monitor the system. Mounce *et al.* (2003) presented an artificial neural network (ANN)-based method for burst detection and location. On the same topic, methods applying Discrete Fourier Transform and Discrete Wavelet Transform have also been proposed (Lee *et al.* 2012; Srirangarajan *et al.* 2013). Babovic *et al.* (2002) reported on the use of data mining methods to evaluate the risks of pipe bursts with the data of historical burst events. In the present work, the water supply system under consideration is in a Chinese city where wireless water pressure sensors are already placed; however, the exact layout of the pipe lines is not available. One goal of the research is to maintain the water pressure in a specified interval because excess pressure increases the energy consumption and the risks of leak and burst while low

doi: 10.2166/hydro.2014.037

pressure leads to poor service. When the pressure at a certain location is too high or too low, the means of adjustment is to regulate the outlet pressure of the pump station(s) nearby. Then, a practical problem is how to find out the pump station(s) that has the immediate impact on the pressure at a given location and how to operate the pump(s). More precisely, the correlation between the pump discharge and the local pressure at some network nodes must be identified in order to define a pumping strategy to reach a fixed pressure target at the nodes.

PROBLEM DESCRIPTION

As shown in Figure 1, wireless water pressure sensors (solid squares) are installed in the water supply pipe network of our concern. The water pressure data are time series measured at 77 locations in the network. In this paper, we focus on the 'sensor node' instead of the node in the context of the hydraulic model although the two notions may overlap. Thus for convenience, the word 'node' refers to the sensor node in the following part of the paper. Nodes of our interest are called center nodes. The nodes whose pressures vary synchronously with that of a center node are called correlated nodes. We call the synchronization relationship 'pressure transfer'. The modeling task of the pressure transfer is divided into two parts. The first part is how to choose the center nodes and determine the corresponding correlated nodes. The second is how to model the pressure transfer between a center node and its correlated nodes. In this paper, the Pearson correlation analysis is used to identify the sub-networks, each of which contains one center node and several correlated nodes, while a linear model is constructed to describe the pressure transfer within the sub-network.

NETWORK DIVISION

Pearson correlation coefficient (PCC)

The PCC between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. In statistics, PCC is a measure of the linear

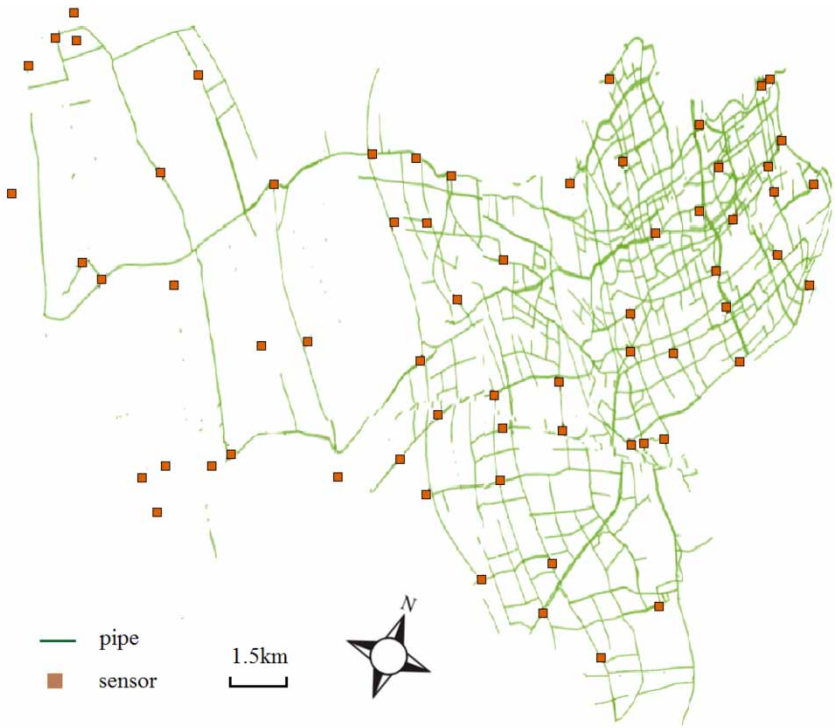

Figure 1 | The layout of main water supply pipes (diameter ≥800 mm) and pressure sensors.

dependence between two variables, giving a value between −1 and 1, where 1 means totally positive correlation, 0 no correlation, and −1 totally negative correlation.

PCC is also referred to as the population correlation coefficient when applied to populations and is represented by the Greek letter ρ. For two random variables X and Y, the formula for ρ is given as

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

In real applications, however, the covariance and the standard deviations can only be estimated with a limited number of samples. Therefore, the corresponding estimation for PCC, also known as the sample correlation coefficient, is presented as

$$r_{X,Y} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2} \sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{s_X} \right) \left(\frac{Y_i - \overline{Y}}{s_Y} \right) \tag{2}$$

where, $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, and $s_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2}$

As guaranteed when the law of large numbers can be applied, the sample correlation coefficient is a consistent estimate of the population correlation coefficient as long as the sample means, variances, and covariance are consistent. It is therefore reasonable to use the sample correlation coefficient to characterize the dependence (correlation) between the pressures of two nodes although we do not actually know the probability distribution of the nodal pressure.

Network division based on the PCC

Among the total 77 nodes in our sensor network, 29 are at pump stations, three are at water treatment plants, and the remaining 45 measure the water pressures at different locations. At each node, pressure measurements of 31 days are collected with a sampling interval of 10 minutes, that is, 4,465 data per node. Since the water supply network covers ca. 250 km^2, it is advantageous to divide the 77-node sensor network so that the divisions can be managed separately.

To quantify the notion of 'highly correlated', a threshold *r_thres* is set for the correlation coefficient and two nodes are claimed to be highly correlated if the PCC between their pressures are greater than *r_thres*. A node is chosen as a center node if it is highly correlated with at least one other node. A center node together with its (highly) correlated nodes is called a sub-network. Clearly, one node may be chosen as a correlated node for several center nodes, i.e., there may be overlaps among the sub-networks.

However, it is found that some highly correlated nodes are located far away from each other, which affects the interpretability of the division. To improve the network division based on the Pearson correlation analysis, field engineers are consulted for the correlation between nodes. In fact, a network division scheme based on their experience is provided, where factors such as pipe connection between nodes and flow direction are taken into account. With such knowledge as reference, correlated nodes respecting both divisions (PCC based and empirical) are retained. As a result, the correlated nodes are reasonably close to the center node in a sub-network.

The 77×77 PCC matrix obtained from the pressure data is shown in Figure 2, in which the numbers from 1 to 77 denote different nodes and the values are represented by grayscale. Because of the fact that $\rho(X, Y) = \rho(Y, X)$, only the upper triangular part is presented. The diagonal elements are all 1. It should be noted that negative correlation coefficients are set as zero (not used during the network division), and hence displayed as black in the figure because the negative correlations are mostly weak (97% of the negative PCCs have absolute values less than 0.5). When the parameter *r_thres* is set as 0.8, the sensor network is divided into 47 sub-networks and a center node has, at most, eight correlated nodes within a sub-network. Three typical sub-networks are shown in Figure 3. A sensitivity analysis for the parameter *r_thres* is addressed in the 'Discussion' section.

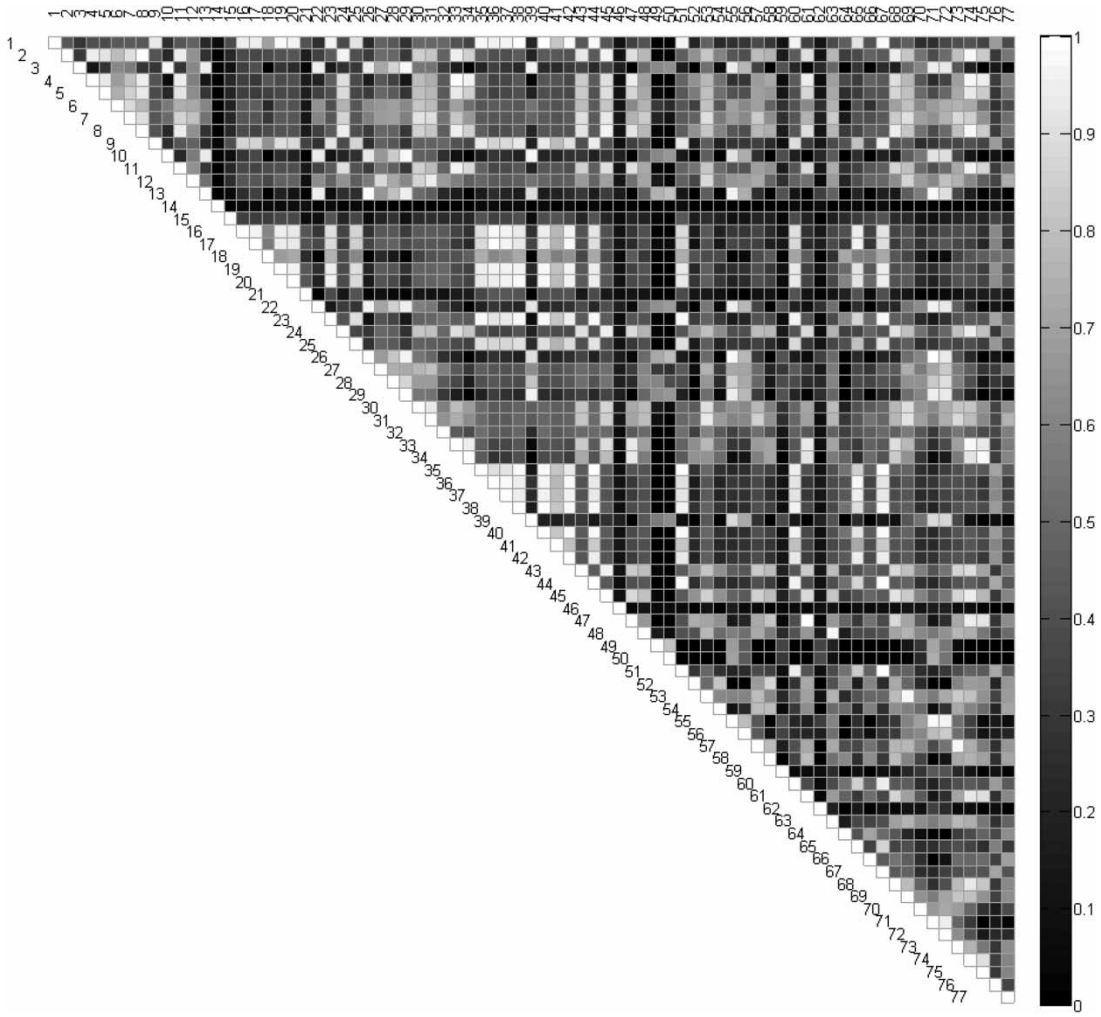

Figure 2 | The matrix of PCC values calculated for every pair of pressure measuring nodes.

MODEL CONSTRUCTION AND VALIDATION

For each sub-network, a linear model which takes the form of Equation (3) is constructed to quantify the pressure transfer from correlated nodes to the center node

$$y = \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon = \mathbf{x}^{\mathbf{T}}\boldsymbol{\beta} + \varepsilon \tag{3}$$

where model input $\boldsymbol{x}$ is the pressure measurements of the correlated nodes at time t, the output y the pressure estimate of the center node at the same time, $\boldsymbol{\beta}$ the vector of regression coefficients, and ε the error term. Different constructions of the model input are tested in the Discussion.

The parameter vector $\boldsymbol{\beta}$ is estimated using least squares with the data of the first two days, that is, $2 \times 6 \times 24 = 288$ input–output pairs. The rest of the data are left for testing. The accuracy of the model is measured with mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{1}{M}\sum_{i=1}^{M}\left|\frac{p_{\text{mdl},i} - p_{\text{mea},i}}{p_{\text{mea},i}}\right| \tag{4}$$

where p_{mdl} and p_{mea} are the model output and the measured pressure at the center node, respectively.

Among the 47 sub-networks, the largest value of MAPE is 2.38%. In addition, the model parameters are not

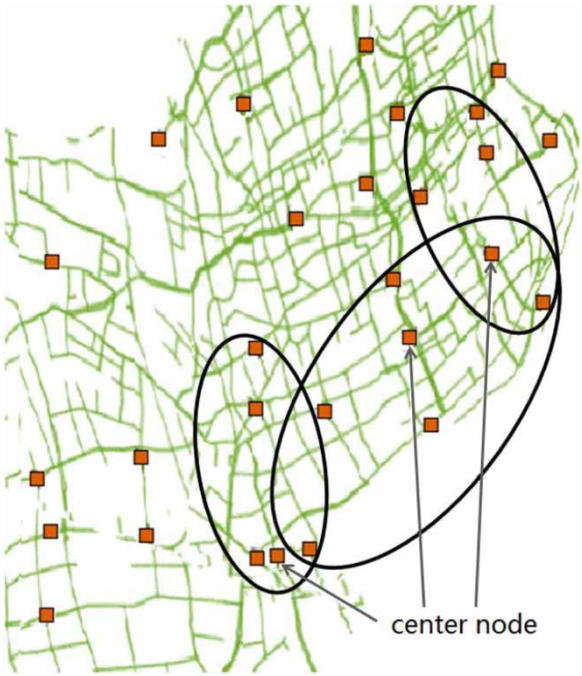

Figure 3 | Three sub-networks obtained through the division.

updated after the calibration. In other words, the pressure transfer relationship described by the model seems to be time independent. Figure 4 shows the error assessment result and Figure 5 shows the model output compared to the pressure measurements on a chosen day. The worst case is shown in Figure 5(c), where the error still seems acceptable.

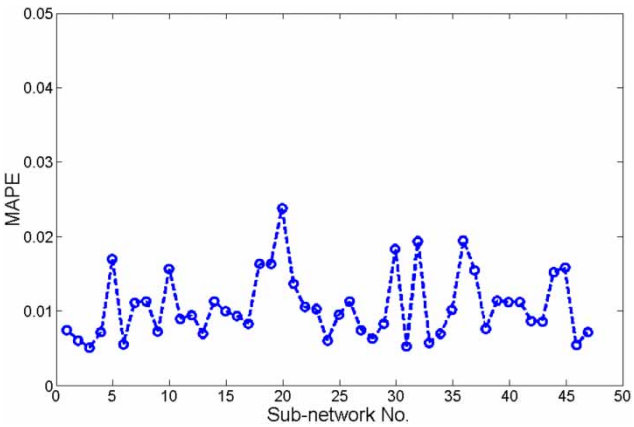

Figure 4 | Error of the linear models constructed for the 47 sub-networks.

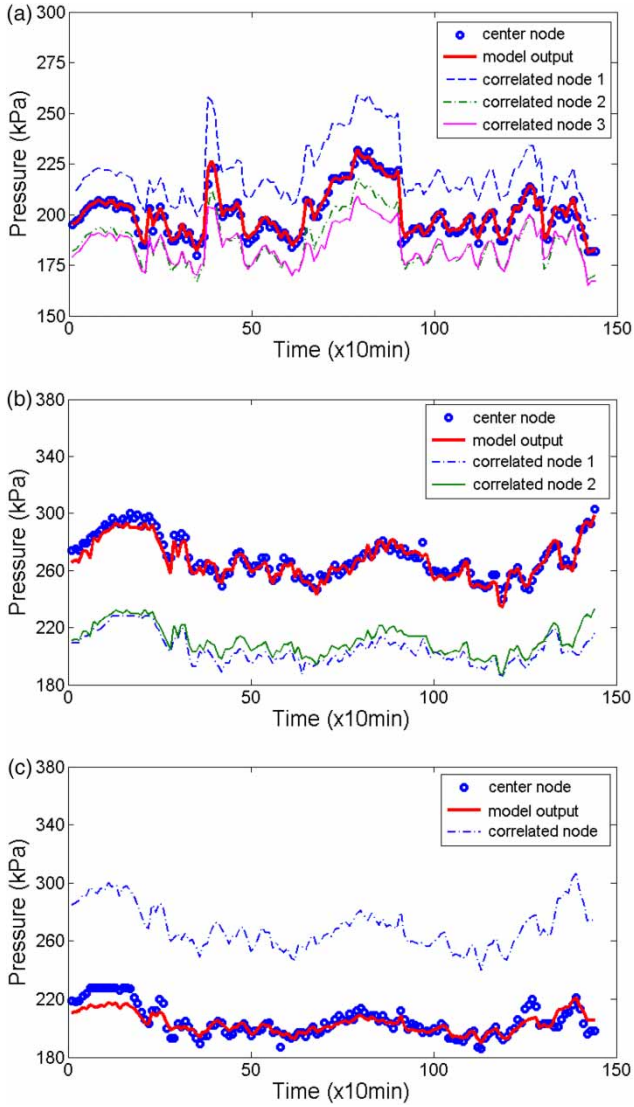

Figure 5 | Model output and pressure measurements for three center nodes on a certain day: (a) center node No. 3 with MAPE 0.50%; (b) center node No. 10 with MAPE 1.56%; and (c) center node No. 20 with MAPE 2.38%.

MODEL APPLICATION

Among the total 47 identified center nodes, 12 are at pump stations measuring the pump outlet pressure. Application of these pump-centered sub-networks is proposed as follows. Suppose that the pressure received from a node is too low and needs to be increased by Δp (kPa), the pressure regulating procedure will be initiated as follows. First, the pump station that is the most correlated to the node of present

interest is located and the corresponding linear model for the pump-centered sub-network is available. Then, the input vector to the model is constructed as $\boldsymbol{p}+\Delta\boldsymbol{p}$, where the vector $\boldsymbol{p}$ contains the current pressure measurements at the correlated nodes of the pump station and $\Delta\boldsymbol{p}=\Delta p\times[1,\ 1,\ldots,\ 1]^{\mathrm{T}}$ with a proper dimension. The output of the model is the suggested pump outlet pressure. The reason for constructing the model input in such a way is that the correlated nodes of a pump station should follow the same trend of pressure change. A numerical example is given below to test the proposition.

Under the empirical scheduling strategy performed at pump stations, the water pressure measured at each sensor node varies cyclically corresponding to the water demand at different times of day. Typically, the pressure is raised up high in the daytime for intensive industrial and domestic water consumption while late at night the pressure is reduced to its lowest level. Thus, for each correlated node in the 12 pump-centered sub-networks, those times when pressure changes significantly are found to simulate the pressure regulating. As shown in Figure 6, the nodal pressure after change is the desired pressure and the pump outlet pressure at the same time is the desired output of the linear model. Model output is compared to the measurement (desired output) using MAPE. For one correlated node of a pump station, as indicated in Figure 7, 67 pressure regulating incidents are found during 29 days (time horizon for model validation) and the model output can fairly

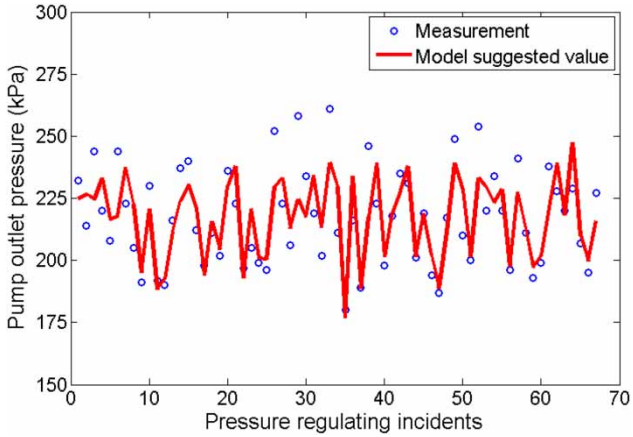

Figure 7 | Pressure regulating at a correlated node, model suggested pump outlet pressure versus measurements (MAPE = 4.36%).

determine the pump discharge. The procedure is repeated for each correlated node in the pump-centered sub-networks and the error assessment result is shown in Figure 8.

It can be seen that the MAPE values are all less than 5%, which means the proposed pressure regulating method has the potential to support the pump operation. The limitation of the method is that it can only be applied to those nodes which are already included in the model. The proportion of (non-pump) nodes covered by the pump-centered sub-networks is nearly half (22/45).

Another possible application of the model is to monitor the operating state of the water supply network. When the pressure data from the correlated nodes in a sub-network are received, they are forwarded to the corresponding linear model to get an output, which is then compared to the

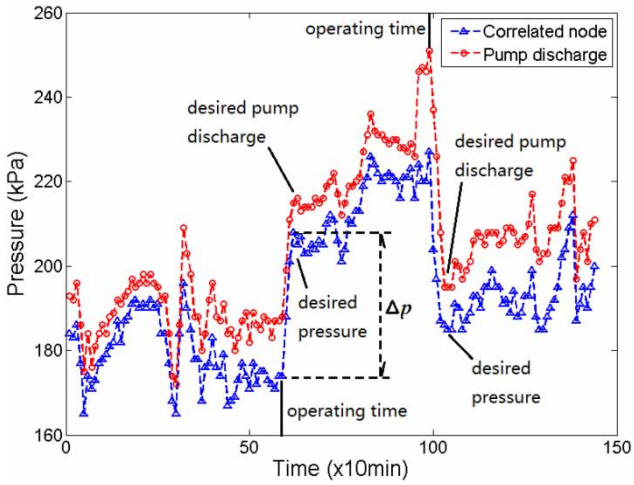

Figure 6 | Pressure at the correlated node affected by the pump discharge.

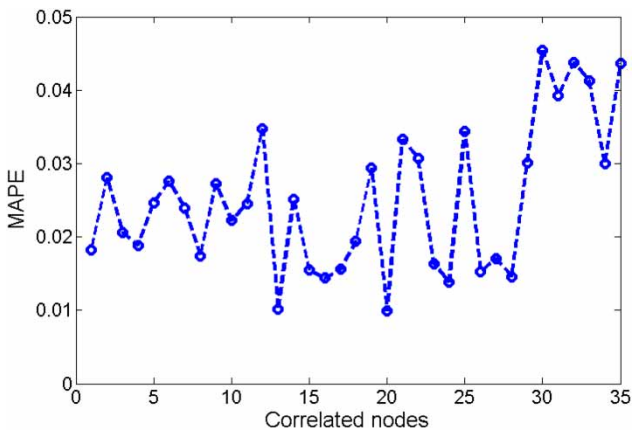

Figure 8 | Error assessment for the simulated pressure regulating.

center node pressure measurement. If the difference is greater than a pre-specified threshold, an unexpected operating condition is detected, which may be a leak within the current sub-network. Due to the lack of data of such accidents, quantified analysis remains to be done. In addition, if high-rate pressure data are available, leak detection and localization can be achieved through analysis of the pressure transients (Whittle *et al.* 2010). For the same purpose, acoustic techniques are well developed and practiced (Stoianov *et al.* 2006, 2007). It must be addressed that since a leak will probably impact more than one sub-network because of the overlap, applying our model directly for leak detection seems challenging.

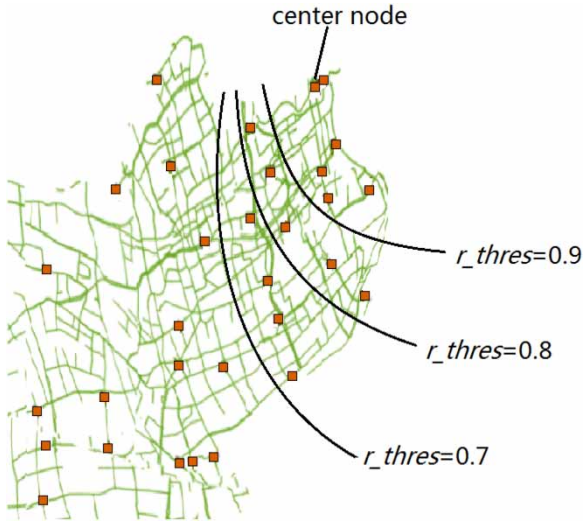

Figure 10 | Nodes included in a sub-network under different *r_thres*.

DISCUSSION

About *r_thres*

Previously, the parameter *r_thres* is set as 0.8 for the network division and a good result is obtained. To make this assignment seem less random, tests for different values of *r_thres* are carried out. Three center nodes are selected and correlated nodes are assigned to them under different *r_thres*. Then the model is recalibrated and the resulting model accuracy is shown in Figure 9. The lines display a common trend that the error goes down as *r_thres* increases.

However, as *r_thres* increases, a sub-network is likely to contain fewer correlated nodes, as shown schematically in Figure 10. Also, the number of nodes which are not contained in any sub-network is 12 under $r\_thres = 0.8$, and even 16 under $r\_thres = 0.9$. Therefore, a trade-off between the model accuracy and the node inclusion in the model must be made to determine *r_thres*. On the other hand, some nodes have low correlations (PCC < 0.5) with all the other nodes, and consequently, data at these nodes are discarded if *r_thres* is chosen to be greater than 0.5. The isolation of these nodes revealed by the data is probably due to the structure of the pipe network and the sensor deployment which is relatively sparse compared to the scale of the network. Although not considered in our paper, design of the sensor network can greatly affect the performance of the sensors, as presented in the work of Ediriweera & Marshall (2010).

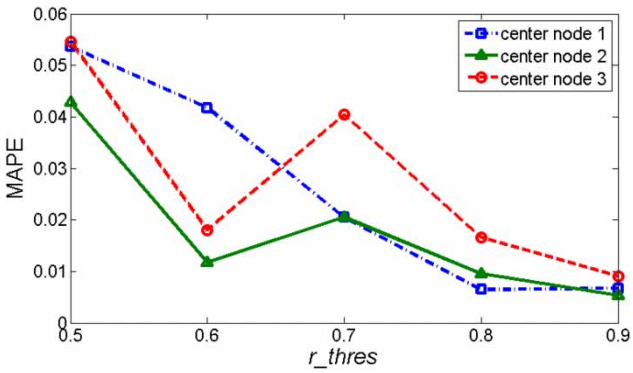

Figure 9 | Error of the model under different values of *r_thres*. Tests are performed for three center nodes.

About the construction of the model input

We have also considered introducing more input variables for the linear model: the historical pressure values of the correlated nodes and the center node. Specifically, in order to estimate the center node pressure at time *t*, the input variables to the model consist of pressure values of the center node and the correlated nodes at time *t*-1, *t*-2, …, *t*-*k*, i.e., the input dimension is $k \times$ (number of correlated nodes + 1). Three sub-networks are chosen for the test and the resulting model accuracy is shown in Figure 11 as *k* varies from 0 to 5, where 0 denotes the input–output mapping adopted in the previous section.

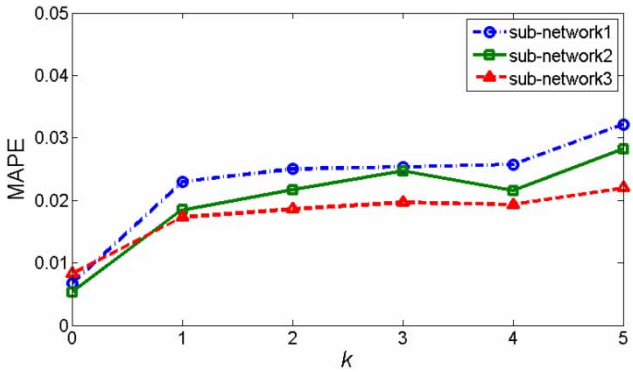

Figure 11 | Error of the model under different constructions of the input. Tests are performed for three sub-networks.

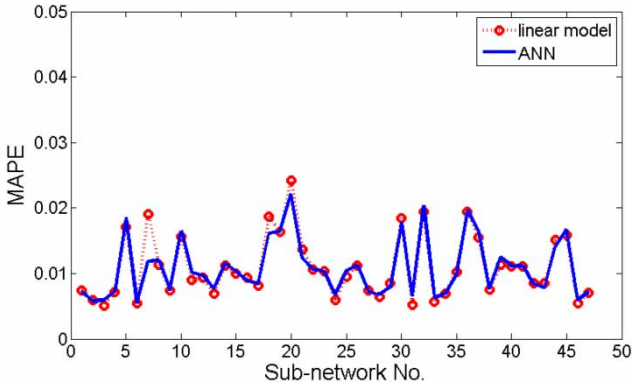

Figure 12 | Prediction error of the ANNs constructed for the sub-networks in comparison with the linear model.

It can be seen that the error becomes larger as more historical pressure values are introduced into the model input. However, different choices of k do not change the decision time, that is, we can decide whether the operating state is normal or an accident may happen only when we have the pressure measurement at time t. Therefore, the most accurate one, namely, $k = 0$ should be preferable.

ANN approximation for the pressure transfer

For each sub-network, an ANN is constructed instead of a linear model to describe the pressure transfer. The ANN adopts a typical architecture of a multi-layer perceptron with a single hidden layer and the input and output are the same as described previously. Again, data of the first 2 days are used for ANN training and the rest for testing. The backpropagation learning algorithm with momentum updating (Ham & Kostanic 2003) is applied to train the ANN and the number of training epochs is no more than 5,000. The prediction error of the ANNs is displayed in Figure 12 together with the error of the linear models. It can be found that the ANN does not outperform the linear model basically. Therefore, it seems to be unnecessary to employ a relatively complex structure for pressure transfer modeling.

CONCLUSION

A data-driven approach for estimating nodal pressures within a water supply network is presented without using a hydraulic model. Sensor nodes are grouped together into a series of sub-networks according to the correlation analysis of the data, and then a linear model is constructed to estimate the pressure of a center node using the observed measurements of the correlated nodes. The average relative error of the estimation is found to be less than 3%. Based on the model, a pressure regulating method is proposed and tested with the real data of the nodes in pump-centered sub-networks. The result shows the potential of the model for the pump scheduling.

ACKNOWLEDGEMENTS

This study is supported by the National Science Foundation of China (grant no. 61233004) and the Doctoral Program of Higher Education of China (grant no. 20110073110018).

REFERENCES

Babovic, V., Drecourt, J., Keijzer, M. & Hansen, P. F. 2002 A data mining approach to modelling of water supply assets. *Urban Water* **4**, 401–414.

Ediriweera, D. D. & Marshall, I. W. 2010 Monitoring water distribution systems: understanding and managing sensor networks. *Drinking Water Eng. Sci.* **3**, 107–113.

Germanopoulos, G. & Jowitt, P. W. 1989 Leakage reduction by excess pressure minimization in a water supply network. *Proc. Instn. Civ. Engrs.* **2**, 195–214.

Ham, F. M. & Kostanic, I. 2003 *Principles of Neurocomputing for Science & Engineering*. China Machine Press, Beijing.

Jowitt, P. W. & Germanopoulos, G. 1992 Optimal pump scheduling in water-supply networks. *J. Water Resour. Plann. Manage.* **118**, 406–422.

Kritpiphat, W., Tontiwachwuthikul, P. & Chan, C. W. 1998 Pipeline network modeling and simulation for intelligent monitoring and control: A case study of a municipal water supply system. *Ind. Eng. Chem. Res.* **37**, 1033–1044.

Lee, S. J., Choi, G. B., Seo, J. C., Lee, J. M. & Lee, G. 2012 Fault detection of pipeline in water distribution network systems. *WASET* **64**, 1089–1094.

Mounce, S. R., Khan, A., Wood, A. S., Day, A. J., Widdop, P. D. & Machell, J. 2003 Sensor-fusion of hydraulic data for burst detection and location in a treated water distribution system. *Inform. Fusion* **4**, 217–229.

Ormsbee, L. E. & Lansey, K. E. 1994 Optimal control of water supply pumping systems. *Water Resour. Plann. Manage.* **120**, 237–252.

Rossman, L. A. 2000 *EPANET 2 USERS MANUAL*. Environmental Protection Agency, Washington, DC, USA.

Simpson, A. R. & Elhay, S. 2009 A framework for alternative formulations of the pipe network equations. World Environmental and Water Resources Congress, ASCE Conference Proceedings, Kansas City, MI, 17–21 May, 342, pp. 283–294.

Srirangarajan, S., Allen, M., Preis, A., Iqbal, M., Lim, H. B. & Whittle, A. J. 2013 Wavelet-based burst event detection and localization in water distribution systems. *J. Signal Process. Syst.* **72**, 1–16.

Stoianov, I., Nachman, L., Whittle, A., Madden, S. & Kling, R. 2006 Sensor networks for monitoring water supply and sewer systems: lessons from Boston. *Proceedings of the 8th Annual Water Distribution Systems Analysis Symposium*, Cincinnati, OH, 27–30 August, pp. 1–17.

Stoianov, I., Nachman, L., Madden, S. & Tokmouline, T. 2007 PipeNet: a wireless sensor network for pipeline monitoring. *Proceedings of the 6th International Conference on Information Processing in Sensor Networks (IPSN), ACM/IEEE, Cambridge, MA*, 25–27 April, pp. 264–273.

Whittle, A. J., Girod, L., Preis, A., Allen, M., Lim, H. B., Iqbal, M., Srirangarajan, S., Fu, C., Wong, K. J. & Goldsmith, D. 2010 Waterwise@SG: a testbed for continuous monitoring of the water distribution system in Singapore. *Proceedings of the 12th Annual Conference on Water Distribution System Analysis*, Tucson, AZ, 12–15 September, pp. 1362–1378.

Yu, G., Powell, R. S. & Sterling, M. J. H. 1994 Optimized pump scheduling in water distribution systems. *J. Optimiz. Theory Appl.* **83** (3), 463–488.

First received 13 March 2014; accepted in revised form 18 June 2014. Available online 24 July 2014