Monitoring data identification for a water distribution system based on data self-recognition approach

Han Che
Tsinghua University
cheh12@mails.tsinghua.edu.cn

Shuming Liu
Tsinghua University
shumingliu@tsinghua.edu.cn

ABSTRACT

Detecting the occurrence of hydraulic accidents or contamination events in the shortest time has always been a significant but difficult task. The simple and efficient way is to identify the sudden changes or outliers hidden in the vast amounts of monitoring data produced minute by minute, which is unpractical for human. A new method, which employs a data self-recognition approach to achieve that automatically, has been proposed in this paper. The autoregressive moving average (ARMA) model was employed in this research to construct the self-recognition model. 56 months monitoring data from Changping water distribution network in Beijing, which was firstly cut into different time-slice series, was used to establish the ARMA model. This provided a prediction confidence interval in order to identify the outliers in the test data series. The results showed a good performance in outlier identification and the accuracy ranges from 90% to 95%.Thus, the ARMA model showed great potential in dealing with monitoring data and achieving the expected performance of data self-recognition technology.

Key words:

data self-recognition; outlier identification; ARMA model

INTRODUCTION

Urban water distribution networks play a crucial role in the deployment of urban water resources. The easiest and most efficient way to monitor the status of a network is to check the monitoring data observed by independent online monitoring sensors, which exist in most water distribution networks. The most widely-used hydraulic monitoring indexes include pressure, flow rate, etc. Vast amounts of data are produced during consecutive 24-hour monitoring, which brings numerous data-identification tasks that are beyond the capability of dealing with it timely. When pipe explosion accidents occur, the pressure or flow rate data may fluctuate greatly or reveal weird value. However, detecting the sudden changes in vast data amounts accurately and timely is quite difficult by human. Therefore, a new method, which employs a data self-recognition approach to identify vast amounts of monitoring data, has been proposed. Data self-recognition approach can deal with all real-time data automatically and detect the outliers appear in the series in the shortest possible time when hydraulic accidents or contamination events happen.

The autoregressive moving average (ARMA) model, as the most widely used traditional statistical model and data-driven model, has been used for prediction in water distribution system fields in the past decades. Smith (1988) built time-series models to describe municipal water demand using similar methods. Jain et al. (2001) used autoregressive models to forecast weekly water demand. Zhou et al. (2000) proposed time series models to forecast short-term water demand variations based on four factors: trend, seasonality, climatic correlation and autocorrelation. The model efficiency of R^2 = 89.6% and a standard error of about 8% is considered acceptable. In 2002, another time series model was developed for water demand projection, in which long-term demand variations and short-term demand variations were expressed as a Fourier series and a climatic regression, respectively (Zhou et al. 2002). The performance R^2 of the model is only 75%, which is not adequate enough to be used in water distribution systems. Babel et al. (2007) developed a multiple coefficient water demand forecast and management model for the domestic sector considering various socio-economic, climatic and policy-related factors. Other data-driven models, including multi-linear regression model, artificial neural network, have also been used in water forecasting field. Chau (2006) reviewed the development and current progress of the integration of artificial intelligence into water quality modeling.

However, few attempts have been made for monitoring data identification using data-driven models. Data

self-recognition technology utilizes historical monitoring data from one independent monitoring site to assess and identify the validity of sensor readings. The objectives of the present study are to develop a data self-recognition model using ARMA modeling and evaluate the performance of the model on data identification.

MATERIALS AND METHODS

Study area and monitoring data

Changping district locates in the northeast of Beijing. The water distribution network in Changping district has over 50 monitoring stations and each dataset comprises several parameters monitored over 5 years. In the present study, Daoxiangcun monitoring station was selected for the analysis. The selected water parameters include pressure, instantaneous flow rate and integrating flow rate, 15-minute basis data series from 2006 to 2010. In this study, the first 56 months of data (90% of the whole data series) were utilized as training data for model construction. The remaining is used for model validation.

Data pre-processing

Data pre-processing involves three steps: replacing missing data, removing outliers and cutting consecutive time series into slices. In the initial data series, some data was missing due to the monitoring sensor's malfunction or the data transmission problem. Moving average method was used to replace the missing data to ensure the continuation of the series, which was the prerequisite of establishing the ARMA model. Outliers caused due to sensor malfunction were removed after justification with local engineers. Outliers were confirmed using a z-score criterion, which is:

$$z = \frac{x_i - \overline{x}}{d}$$

where x_i represents each historical value, $\overline{x}$ represents the average of the historical series, and d is the standard deviation of the historical value.

When the absolute value of z is above 2, the data is called common outlier. While the absolute value of z is above 3, the data is called highly outlier (Abraham et al. 1989). The range of monitoring data fluctuates greatly due to the change of seasons and time of the day. To minimize the impact, the original data series were re-grouped according to season and time. The data of a certain time and season from the consecutive time series was extracted to form new series. In this paper, the 24 hours continuous dataset from one sensor were cut into 96 pieces, each of which represents the data monitored at every 15 minutes, e.g. from 0:00 to 0:15. These data were then used to constitute 96 new data series, which represents the historical observed data for specific time slice. To reflect the variation of different seasons, these 96 data series were then treated to generate 384 new data series on the basis of the division of season. Each final data series includes 350 to 400 values. If the series was continued cutting into more series considering the effect of weekdays and holidays, the amount of the data could not guarantee the reliability of the model. Therefore, 384 data series considering the effect of certain time and season were produced.

Data self-recognition

In this study, an auto regressive and moving average modeling (ARMA) approach (Box et al. 1991) was employed for data self-recognition.

A general ARMA(p,q) model, is shown below:

$$\begin{cases} x_t = \varphi_0 + \varphi_1 x_{t-1} + \cdots + \varphi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \\ \varphi_p \neq 0, \quad \theta_q \neq 0 \\ E(\varepsilon_t) = 0, \quad Var(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E x_s \varepsilon_t = 0, \forall s < t \end{cases}$$

where p,q is the order of the model, φ is the AR operator of order p, and θ is the MA operator of order q.

The ARMA modeling approach involves the following steps: determining the model structure using the minimum Akaike Information Criterion (AIC)(Bozdogan,1987), estimating the parameters of model and

examining the residuals of the model using Durbin-Watson statistics (Montgomery et al. 2001)in order to verify if the model is an adequate one for the series.

Data identification

After establishing the ARMA models, model predictions are used to identify whether the reading is a an outlier or not. The interval with a 95% confidence level of one-step prediction is adopted to assess the abnormality and shown below:

$$\left(\hat{x}_t(1) \mp 1.96 \cdot \sqrt{Var[e_t(1)]}\right)$$

where $\hat{x}_t(1)$ is the one-step prediction value; $e_t(1)$ is the series of the residual error between the fitting data and observed data in training dataset; Var means variance;

The value that falls outside the interval will be recognized as an outlier, while that falls inside the interval will be recognized as a normal value, which indicates that the network around the monitoring station works with no problem. It is worth noticing that all the "normal" words that mention below represent the situation that value falls inside the interval, but not refers to the Gaussian distribution in statistical field.

Data feedback correction

Outliers or errors would influence the accuracy of subsequent identification. For data self-recognition, outliers impact the trend of prediction interval, resulting in possible consecutive error judgment on data identification, not only for outliers but for normal data as well. Therefore, it is quite essential to have data feedback correction. Similar researches (Tian et al. 2004) have been conducted the same analysis on short-term load data forecast in electric power system field. Commonly used methods for data feedback correction include moving average method, data horizontal and longitudinal comparison, probability method, etc. In this study, moving average method is used as data feedback correction function to replace the outliers or errors to ensure the precision of models.

Model performance evaluation

The performance of data identification in different ARMA models are evaluated using the mean absolute percentage error (*MAPE*) and the outlier recognition rate. The MAPE criterion, which is defined as below, is used to evaluate the fitting performance of the model:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f_i - o_i}{o_i}\right| \times 100\%$$

where f_i and o_i are the fitting and observed values, respectively, and n is the number of total training data. The value of *MAPE* can reflect the effectiveness of the models.

The outlier recognition rate (*ORR*) is defined as:

$$ORR = \frac{m_f}{m_t} \times 100\%$$

where m_f and m_t are the number of recognized outliers and the number of total outliers, respectively.

In this study, artificial simulated outlier sequences were added to the original series to test the identification performance of the models. Each original series contain three types of artificial simulated outlier sequences, each of which consist of 50 random data. 5,10,20 outliers were added to different types of sequences respectively. The identification performance was demonstrated by the outlier recognition rate of each sequence.

RESULTS AND DISCUSSIONS

384 ARMA models were developed using instantaneous flow rate data by MATLAB software. Due to the limitation of length, 16 typical models, which were built by the data series at 0:00, 06:00, 12:00, 18:00 in four seasons respectively, were mainly discussed in this paper. ARMA model at 06:00 in autumn is shown in Figure.1 as an example. The MAPE value of the model in Figure.1 is 8.73%, which showed a good fitting

performance. The average MAPE value of most established models were approximately 10%, except for some rare series whose MAPE value was nearly 20%.

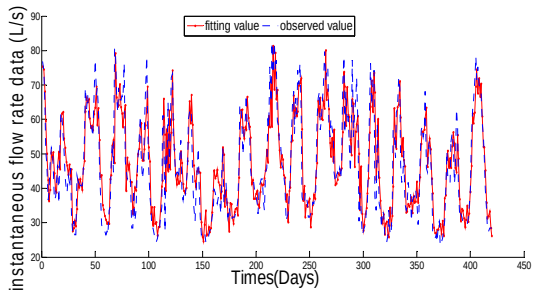

Figure.1 ARMA model at 06:00 in autumn

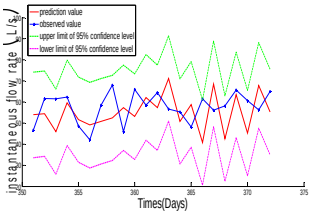

Figure.2(a)

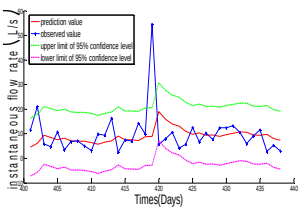

Figure.2(b)

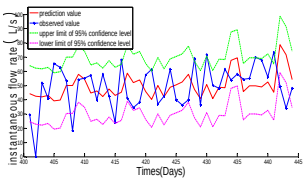

Figure.2(c)

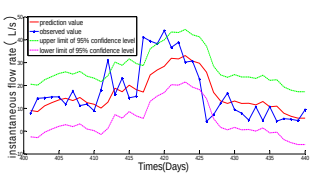

Figure.2(d)

Figure.2 data identification performances in four seasons

The data identification graphs without data feedback correction for models at 12:00 in four seasons are shown in Figure.2. Comparing the graphs in Figure.2, it is observed that four models in different seasons correspond to four different types of identification performance. In the spring models (Figure. 2(a)), the data series is stable and all data are within the prediction confidence interval, which indicates that this part of the network is in normal status. In the summer models (Figure. 2(b)), the data series is also stable. Only few data increase suddenly outside the interval, which indicates that these data are recognized as outliers. In the autumn models (Figure. 2(c)), the data series fluctuates frequently resulting in an opposite trend for prediction interval due to the lag of the model. Thus, it is hard to tell whether the data recognized by the model is really an outlier or not. In the winter models (Figure. 2(d)), the series is not stationary and suddenly-increased data appear consecutively. It can also be seen from Figure.2 that series during different periods of time and seasons have different trends, which proved the necessity of time series slice cutting.

To test the performance of data identification in quantitative perspective, 5,10,20 outliers were added to the original series, respectively. For each series, prediction and identification were made in two ways: with data feedback correction and without it. The results for series at 06:00 in summer with 10 added outliers were shown in Figure.3.

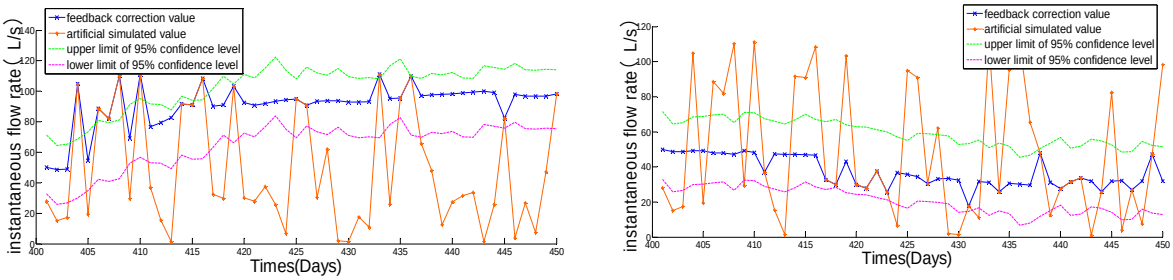

Figure.3 data identification performance at 06:00 in summer with/without data feedback correction

As shown in Figure.3, all outlier recognition rates for models with data feedback correction are 100%. For models without data feedback correction, the outlier recognition rates are 80%, 50% and 25% respectively. It is obvious that the models with correction showed greater performance in outlier identification. A summary for the ORR performances of other models is given in Table 1.

Table 1 the ORR performances of models

ORR / Model	5-outlier		10-outlier		20-outlier	
	With DFC	Without DFC	With DFC	Without DFC	With DFC	Without DFC
00:00 of spring	100%	100%	100%	70%	100%	35%
06:00 of summer	100%	80%	100%	50%	100%	25%
12:00 of autumn	100%	80%	100%	60%	90%	40%
18:00 of winter	100%	60%	60%	20%	95%	15%
Average	100%	80%	90%	50%	96.3%	28.8%

DFC= Data Feedback Correction

For the 5-outliers artificially simulated sequences, the average ORR is 100% and 80% corresponding to the models with data feedback correction and without it, respectively. For the 10-outliers artificially simulated sequences, the average ORR is 90% and 50% respectively. For the 20-outliers artificially simulated sequences, the average ORR is 96.3% and 28.8% respectively. It can obviously be noticed that models with data feedback correction step have great potential in outlier identification with an average performance over 90%. Meanwhile performances of models without data feedback correction are poor in outlier identification with an average performance only about 53%.

CONCLUSIONS

A data self-recognition method based on ARMA modeling was proposed in this research for online monitoring data quality control. The results showed that series in different periods of time and seasons have different trends, which proved the necessity of time series slice cutting.

Results show that models with data feedback correction step had great potential in outlier identification with an average performance over 90%, while models without that step were poor in outlier identification with an average performance of only about 53%, which can't satisfy the expected effect.

Based on the findings of this study, it is concluded that data self-recognition modeling can be successfully applied to establish reliable monitoring data identification models for detecting outliers timely when hydraulic accidents or contamination events happen. It is also concluded that data feedback correction is essential for data self-recognition and can reduce workload significantly.

ACKNOWLEDGMENTS

The authors would like to thank Tsinghua Independent Research Project (2011080993) for its financial support to this research.

REFERENCES

1. Abraham, B., Chuang, A. (1989) Outlier detection and time series modeling, *Technometrics*, 31, 2, 241-248.
2. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., (1991) Time Series Analysis, Forecasting and Control, Prentice Hall, Englewood Cliffs, NJ, USA.
3. Bozdogan, H., (1987) Model Selection and Akaike's Information Criterion(AIC): The General Theory and Its Analytical Extensions, *Psychometrika*, 52, 345-370.
4. Babel, M. S., Gupta, A. D., Pradhan, P., (2007) A multivariate econometric approach for domestic water demand modeling: an application to Kathmandu, Nepal, *Water Resources Management*, 21, 573-589.
5. Chau, K. W., (2006) A review on integration of artificial intelligence into water quality modeling, *Marine Pollution Bulletin*, 52, 726-733.
6. Jain, A., Vershney, A. K., Joshi, U. C., (2001) Short-term water demand forecast modeling at IIT Kanpur using artificial neural network s, *Water Resources Management*, 15, 299-321.
7. Matrix Laboratory, Mathworks Inc.
8. Montgomery, D. C., Peck, E. A., Vining, G. G., (2001) Introduction to Linear Regression Analysis, 3rd Edition, John Wiley & Sons, New York, New York.
9. Palani, S., Liong, S. Y., Tkalich, P., (2008) An ANN application for water quality forecasting, *Marine Pollution Bulletin*, 56, 1586-1597.
10. Smith, J. A., (1988) A model of daily municipal water use for short-term forecasting, *Water Resources Research*, 24, 2, 201-206.
11. Tian, Z. Y., Zhang, M. L., Zhao, R., (2004) Identification and manipulation of anomalous data of load sequence in short-term load forecasting, *Jilin Electric Power*, 175, 21-23.
12. Zhou, S. L., McMahon, T. A., Walton, A., Lewis, J., (2000) Forecasting daily urban water demand: a case study of Melbourne, *Journal of Hydrology*, 236, 153-164.
13. Zhou, S. L., McMahon, T. A., Walton, A., Lewis, J., (2002) Forecasting operational demand for an urban water supply zone, *Journal of Hydrology*, 259, 189-202.