Missing data repair method of water system based on graph signal sampling theory

(Excerpted Version)

Undergraduate Thesis, Tsinghua University

Supervisor: Shuming Liu, Division of Water Supply & Drainage

Xuanting Ji

1 abstract

Data quality is always a limiting factor for the development of water systems. Based on graph signal sampling, this paper carries out repair calculation for the missing data of water network.

Graph signal sampling theory is a feature frequency analysis method based on graph theory, which arranges the time series data according to the adjacent correlation and periodicity of hydraulic data, so as to integrate the characteristics of hydraulic data. The graph data obtained by time series sorting is transformed into time Laplace matrix, and the characteristic frequency analysis and missing data estimation calculation are carried out by using the low frequency property of hydraulic data, so as to obtain hydraulic data of the whole time series.

The implementation and test results

The method shows feasibility on a considerable proportion of missing data. For discrete missing, when the missing proportion is within 30%, the R^2 of the results is more than 0.9, and the MSE is less than 0.1. For 3 consecutive days of missing data in 30 days, the R^2 is about 0.9 and the mean absolute error is about 20%. The method is also compared with other calculating methods such as linear calculation, periodic method and time-change pattern method.

The graph signal sampling theory eliminates the need for intricate data pre-processing and prior data training. It simply requires the organization of the relevant time series data. The repaired data provides basis for a wide range of applications in urban water management, such as pipe burst detection and flow prediction.

2 Introduction

Monitoring data of water distribution network is the basis of intelligent application of pipe network. The water companies or other online monitoring management departments obtain the monitoring data of each node, evaluate the operation status of the water distribution network, and timely identify the abnormal situation of the water distribution network, such as leakage and pipe explosion, so as to carry out maintenance as soon as possible and ensure the normal life and production of urban residents. At the same time, there are numerous related studies based on water distribution network information data. Therefore, it is necessary to obtain accurate state information data in urban water distribution networks.

There are two main difficulties in Water distribution network data monitoring.

- **Poor quality:** Monitoring data often appear abnormal, missing and other problems. About 10% of the flow and pressure monitoring data of the pipeline network are missing, and 3% to 35% of the data are abnormal.
- **Transmission technology limitation :** The underground water distribution network are numerous nodes and complex topology, and the cost of maintenance and charging of the sensors of the detection equipment is also complex. Whether from the cost of water supply network management or monitoring equipment technology itself, do not support the water supply network for continuous monitoring of all nodes of data

The existing missing data estimation methods of water supply network mainly include two categories.

- **hydraulic model:** Pipeline network modeling softwares, such as EPANET (US) and InfoWorks (UK), simulate the flow state inside the pipeline through pipeline hydraulic calculations. However, the relevant parameters such as pipe diameter and roughness coefficient change dynamically, and this change is difficult to correct in time. What's more, the number of parameters such as node demand and node flow is large and difficult to determine accurately.
- **Prediction/Calculation:** Various Calculative algorithms are used to calculate the unknown hydraulic information of network in time dimension, including alternatives based on periodicity, regression analysis, Computational methods based on deep learning. Achieving a balance between enhancing precision and minimizing computational expenses poses a challenging endeavor.

The current techniques for approximating water network data primarily rely on parameter simulations or pre-existing data training corrections. However, these methods often result in error introduction or escalated expenses. This study proposes a data patching approach, leveraging the collective attributes of available monitoring data and temporal series correlations. By numerically estimating absent data, it aims to reconstruct the entire time series of water supply network monitoring data. The method could furnish a foundational dataset for flow prediction, burst detection, and intelligent water supply network management.

3 Methodology

The graph signal sampling theory is based on the graph signal processing technology. Through the graph characteristic frequency analysis and according to the characteristics of the system data, the relevant frequencies and corresponding characteristic frequency vectors that can represent the characteristics of the system data in the full frequency sequence are selected, and this process is called "graph signal sampling" to remove the interference information in the system data, which is similar to the "filtering" process in the signal processing process. Then the selected characteristic frequency information is harnessed for the comprehensive analysis and computation of the entire system data.

The time series data in water system is represented as a topological diagram composed of a series of points and edges, as shown in Fig.1. The horizontal represents the time data of the day, the vertical represents different dates, each node represents the corresponding day and time, and the connected edge represents the correlation between the hydraulic data corresponding to the two moments.

Different matrices are used to describe the correlation of data points.

With different connection between vertexes V_p and V_q, time topological matrix A_t:

$$A_t(p,q) = \begin{cases} 0, if\ V_p, V_q\ not\ connected \\ 1, if\ V_P, V_q\ connected \end{cases} \tag{1}$$

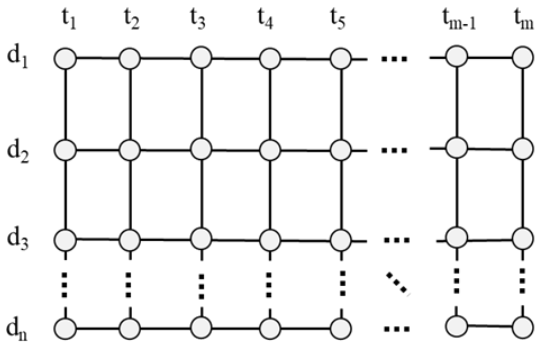


Figure 1: Topological diagram of time series data

Degree matrix D_t is a diagonal matrix, representing the sum of the weights of this vertex and other vertices.

$$D_t(p,q) = \begin{cases} 0, if\ p \neq q \\ \sum_{k=1}^{m \times n} A_t(p,k), if\ p = q \end{cases} \tag{2}$$

Time Laplacian matrix L_t describes the cumulative effects on other nodes when other nodes generate disturbances. Conduct eigenvalue decomposition of the time Laplacian matrix. The eigenvector corresponding to the smaller eigenvalue is the low-frequency term of the pipe network data change.

$$\begin{aligned} L_t &= D_t - A_t \\ &= U_t \Lambda U_t^T \end{aligned} \tag{3}$$

μ_t represents the low frequency feature vector in U_t. According to the low frequency of hydraulic data of pipe network, the low frequency term of Fourier change can be calculated from part of the monitoring data. y_{t_0} is the monitoring data after the removal of outliers and missing values; X_t is the characteristic spectral coefficient of pipe network data in time. $\widetilde{y_t}$ is the patched data.

$$\widetilde{y_t} = U_t \begin{bmatrix} X_t \\ 0 \end{bmatrix} \tag{4}$$

4 Applications and Comparisons

The method was tested on actual monitoring data of 30 days. There are no anomalous data after anomaly identification, which facilitates data preparation and post-calculation error analysis. The interval for uploading data at the monitoring point is 15 minutes, resulting in 96 data points per day. Different proportions of data were selected with *random* function in Python, and were set to blank. The remaining data were taken as inputs to the method. The results of the calculations are obtained, and the mean absolute error is obtained by calculating with the original real data, as shown in Fig.2 and Fig.3. The results show the trend that the higher the proportion of missing, the higher the error of calculation results. The average flow data of the monitoring point is 1.9L/s. When the proportion of missing was less than 70%, the error was less than 1.0L/s, as shown in Figure 3.5. When the missing proportion is less than 30%, the error is less than 0.3L/s.

The number of frequencies with the smallest average error is taken as the optimal number of frequencies. Repeat this calculation process 30 times to ensure the rationality of the calculation results. The higher

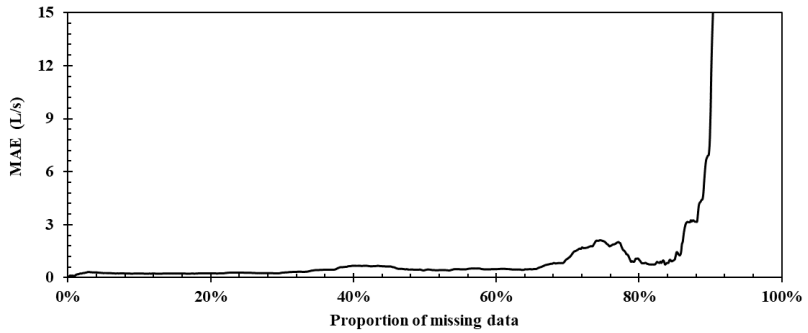


Figure 2: Absolute error of different proportion (within 100%) of missing data

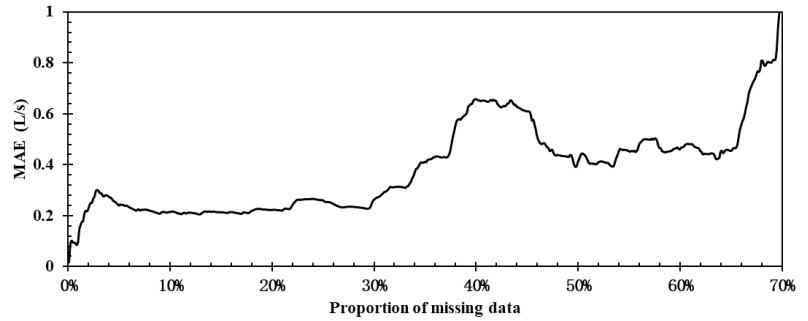


Figure 3: Absolute error of different proportion (within 70%) of missing data

the proportion of missing data, the lower the optimal low-frequency amount. This is because too many frequency components bring some of the non-low-frequency interfering information into the calculation process, leading to the deviation of the calculation results from the accurate values.

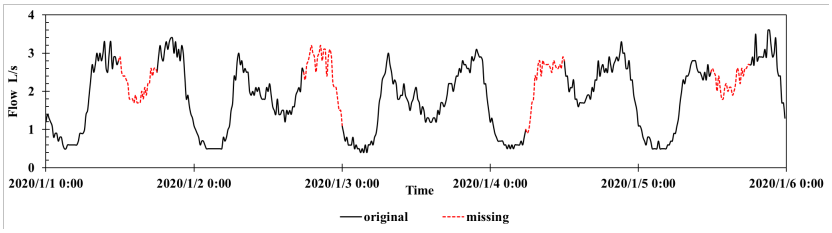
In the actual data monitoring process, there are two common categories of missing data: discrete missing and continuous missing, as shown in 4. For discrete missing, segments with different length are common. Therefore, different groups of data were randomly selected and set blank in the overall data, with 24 data points in each group. For continuous missing, data of several days are selected and set blank.

For discrete missing data, the calculation is compared under different methods with missing proportions: linear estimation method, periodicity based replacing method, time variation pattern based calculation method, and graph signal sampling (GSS) based calculation method. The mean squared value are shown as Fig.5.

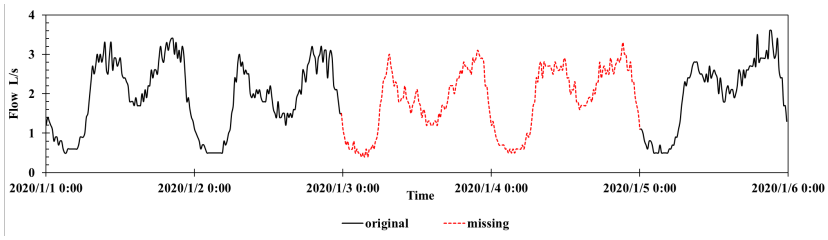
The determination coefficient R^2 between the calculated values of the four methods and the actual monitored values was calculated respectively. The repairing effect of linear method and alternative method is relatively poor, and the R^2 is about $0.3 \sim 0.5$. The repairing results of the graph signal sampling method and the time-change pattern method are relatively better, and the R^2 is about 0.8, which decreases slightly with the increase of the missing proportion.

For continuous missing data, the hydraulic data will show periodic fluctuation in a long time series, and the linear estimation method is no longer applicable. The errors of three methods are compared under different missing proportions, as shown in Fig.6.

The MAE of the graph signal sampling method is about 20% under different miss lengths, while the mean square error MSE is lower than 0.2, which is lower than the other two methods, and the error gap is larger with the increase of miss length.

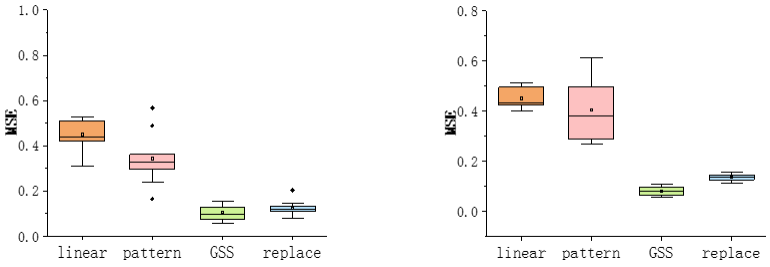


(a) Diagram of discrete loss



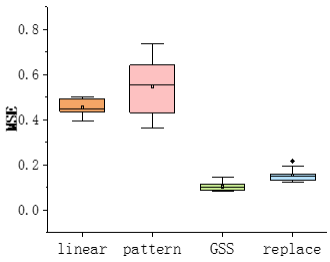
(b) Diagram of continuous loss

Figure 4: different categories of loss

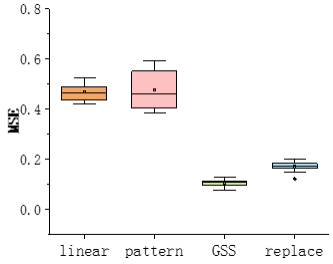


(a) 10% missing

(b) 20% missing

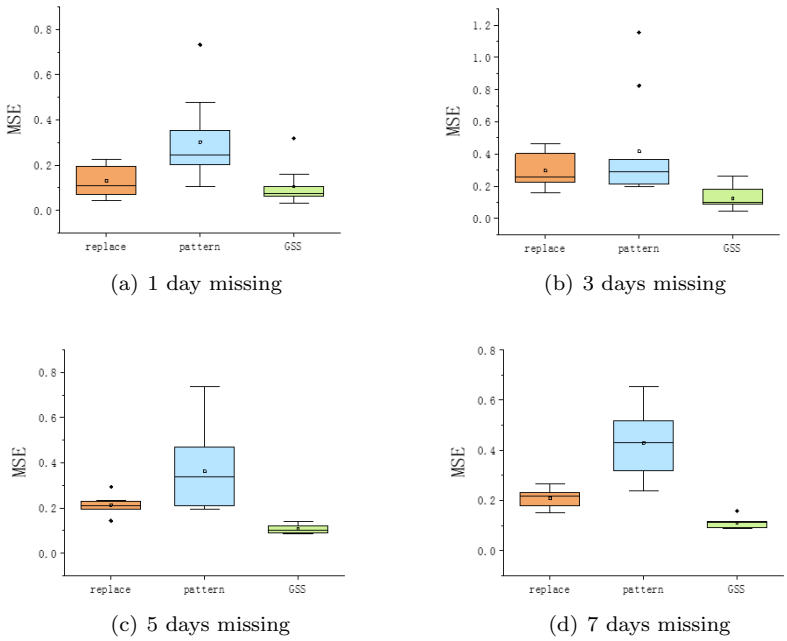


(c) 30% missing



(d) 40% missing

Figure 5: MSE of different missing proportions



(a) 1 day missing

(b) 3 days missing

(c) 5 days missing

(d) 7 days missing

Figure 6: MSE of different missing days

5 Results

In this paper, the graph signal sampling method is applied to data repair in water networks and the following conclusions are obtained.

- The proportion of missing data was negatively correlated with the repair effect.
- By comparing with other methods, the data repair error of the graph signal sampling method is lower and better fitted to the real monitoring data.
- The number of feature frequencies used has essential impact on the results. The optimal feature frequency should be used in the process of using the method to obtain the optimal effect.