https://doi.org/10.71573/v8x9h144

Graph-Based Model for Efficient Data Retrieval in Incomplete Stormwater Networks

Mohsen Hajibabaei [1] https://orcid.org/0000-0002-0047-9715, *Sina Hesarkazzazi* [1] https://orcid.org/0000-0002-0828-9182 *& Robert Sitzenfrei* [1,*] https://orcid.org/0000-0003-1093-6040

[1]Unit of Environmental Engineering, Department of Infrastructure Engineering, University of Innsbruck, 6020, Innsbruck, Austria
**Corresponding author email: Robert.Sitzenfrei@uibk.ac.at*

Abstract
Modelling urban stormwater networks (USNs) provides valuable insights into their performance and assists in improving management strategies. However, a common and significant challenge arises from incomplete information about USNs, where essential network data (e.g., sewer diameters) are unavailable, hindering reliable hydrodynamic analysis. To address this issue, we propose an efficient and fully automated graph-based data retrieval model for USNs with incomplete information. The model automatically infers missing physical attributes, including sewer diameter and slope, by considering the topological features (e.g., connectivity) and hierarchical patterns observed in sewer diameter variations. The framework was tested on a real-world USN with a complete dataset, where data gaps were artificially introduced by randomly removing sewer diameter and slope information ranging from 10% to 90%. Each data gap scenario was repeated 100 times, resulting in 900 incomplete USN configurations. The results demonstrated that the model efficiently retrieved missing data with high accuracy, achieving up to 90% data recovery while accurately reproducing hydrodynamic attributes, such as flow rates. This model provides an efficient tool for water utilities managing incomplete USNs, enabling them to conduct various hydrodynamic analyses.

Highlights
- A graph-based model to retrieve unavailable diameter and slope (invert elevation) data.
- Real-world USN investigations revealed robust performance despite significant data gaps.
- Achieves up to 90% data recovery within seconds, with reliable flow rate predictions.

Introduction
Urban stormwater networks (USNs) are critical infrastructure that protect communities and urban environments against pluvial flooding. Effective management of USNs requires hydrodynamic analyses, which depend on both high-quality and sufficient data. However, due to inconsistent documentation, many aging networks lack essential data (e.g., sewer diameter), and when data is available, it frequently lacks sufficient quality (Hajibabaei et al., 2024). Therefore, reliable methods for recovering missing data or even verifying existing data in USNs are crucial for accurate hydrodynamic analysis. To address this challenge, some studies have implemented approaches such as machine learning to infer missing sewer diameters and material data in incomplete networks (Belghaddar et al., 2021). However, these machine learning-based methods require extensive data preparation and training, limiting their applicability, particularly for large-scale USNs. Other studies have employed virtual sewer generators to recover network information (Montalvo et al. 2024). Such network generators rely on iterative design procedures based on local regulations, which demand high

computational effort, especially for large networks. To overcome these limitations, this paper proposes a computationally efficient and automated model that does not require extensive surrogate data for retrieval. Built on graph theory, the model enables reliable data recovery for incomplete USN information, providing a practical tool for water utilities to enhance stormwater management.

Methodology

To illustrate the proposed model, an example is presented in Figure 1. In Figure 1a, a USN with complete data is shown, while in Figure 1b, it is assumed that 70% of the edges, represented in grey, have unknown diameter information. Red nodes in Figure 1b indicate manholes with unknown invert elevations. In this study, a manhole's invert elevation is considered unknown when all sewers connected to it lack diameter data. The data required for the retrieval model include the layout of the USN, sewer lengths, and runoff areas for the manholes. The model consists of five modules: Uniformity, Hierarchy, Completion, Elevation, and Hydrodynamic. In each module, a value called the Level of Confidence (LC) is assigned to each sewer to indicate the model's confidence in the retrieved data. For sewers with available diameter data, LC is set to 1 (100% confidence). For sewers with missing diameter data, LC is initially set to 0 and then updated during the retrieval process, as described in Equation (1). Consider the shortest path (SP) between two graph nodes where the edge weight (sewer length) is minimal. If multiple edges (denoted as e_k with k = 1, 2, ..., n) along the SP have unavailable diameter information, then the missing values can be estimated using two reference edges, one at the upstream end and one at the downstream end. In this case, the LC of e_k is calculated as follows:

$$LC(e_k) = \begin{cases} \alpha_{up} \cdot (1 - k \cdot \delta) & for\ \ k \leq \lfloor n/2 \rfloor \\ \dfrac{\alpha_{up} \cdot (1 - k \cdot \delta) + \alpha_{down} \cdot (1 - (n - k + 1) \cdot \delta)}{2} & with\ k = \dfrac{n+1}{2}, if\ n\ is\ odd \\ \alpha_{down} \cdot (1 - (n - k + 1) \cdot \delta) & for\ \ k > \lceil n/2 \rceil \end{cases} \quad (1)$$

Where, α_{down} and α_{up} are the availability factors associated with the reference edges, and δ is a confidence decay factor that quantifies the loss of confidence as data is propagated from known values to unknown ones. For reference edges with the LC=1, α is set to $\#\,available\ diameter/\#\ total\ edges$. For reference edges with 0 < LC <1, α is set to their LC values. Additionally, δ represents the ratio of the missing gap in the SP to the overall gap in the USN, as well as uncertainties in each module, calculated as $\delta = (n/\#unavailable\ diameter) \cdot (1 + \beta)$, where β is the module-specific decay factor, ranging from 0 to 1. The summarized explanation of each module is as follows. (1) Uniformity Module: As shown in Figure 1c, this module fills the unknown diameters of edges e1-e3 between two sewers (A and B with identical known diameters) that are connected via a directed path. The same diameter value is inferred for these edges (e1-e3) using the SP approach. For this module, β is set to 0 due to low uncertainty. (2) Hierarchy Module: This module employs a customized centrality metric, runoff edge betweenness centrality (EBC^R) (Hesarkazzazi et al., 2022), that spatially distributes the runoff area of each node through sewers and provides insight into the aggregate runoff areas of each sewer (see Figure 1d), which is proportional to pipe capacity. The process starts with the second largest diameter (0.6 m in Figure 1e), and the SP is conducted between nodes N5 and the outlet. Following the hieratical patterns of diameters and using EBC^R values obtained in Figure 1e, the model cross-references the EBC^R values of p1 to p3 with those of reference edges H and I to estimate their diameters. For the module, β is set to 0.5. (3) Completion module: when no reference edge exists upstream of the edge, as seen in edges like t1 and t2 in Fig. 2f, the Completion module is applied to complete the diameter retrieval. Here, edges with unknown diameter information are typically assigned the same diameter as their downstream reference edge when EBC^R values are similar, as with t1 and t2. For edges like p1 and p2, where EBC^R values differ significantly, diameters are inferred by comparing their EBC^R values with the average values of edges with known diameters. Due to high uncertainty, β in the module is set to 1. (4) Elevation Module: The invert elevations, such as J1 in Figure 1h, are estimated by considering the available slopes of neighbouring sewers (edges e1 to e3 in Figure 1h) and using minimum slope requirements obtained from the retrieved diameter (edges e4 to e7). (5)

Hydrodynamic Module: Once the data is retrieved, a hydrodynamic model of the completed USN is constructed by converting the graph into a Stormwater Management Model (SWMM). This model ensures that the reconstructed USN can handle real operational conditions by checking for discrepancies in the flow depth-to-diameter ratio. The retrieval process is repeated for specific sewers if significant discrepancies are found. Finally, the model concludes with the completion of diameter and slope retrieval.

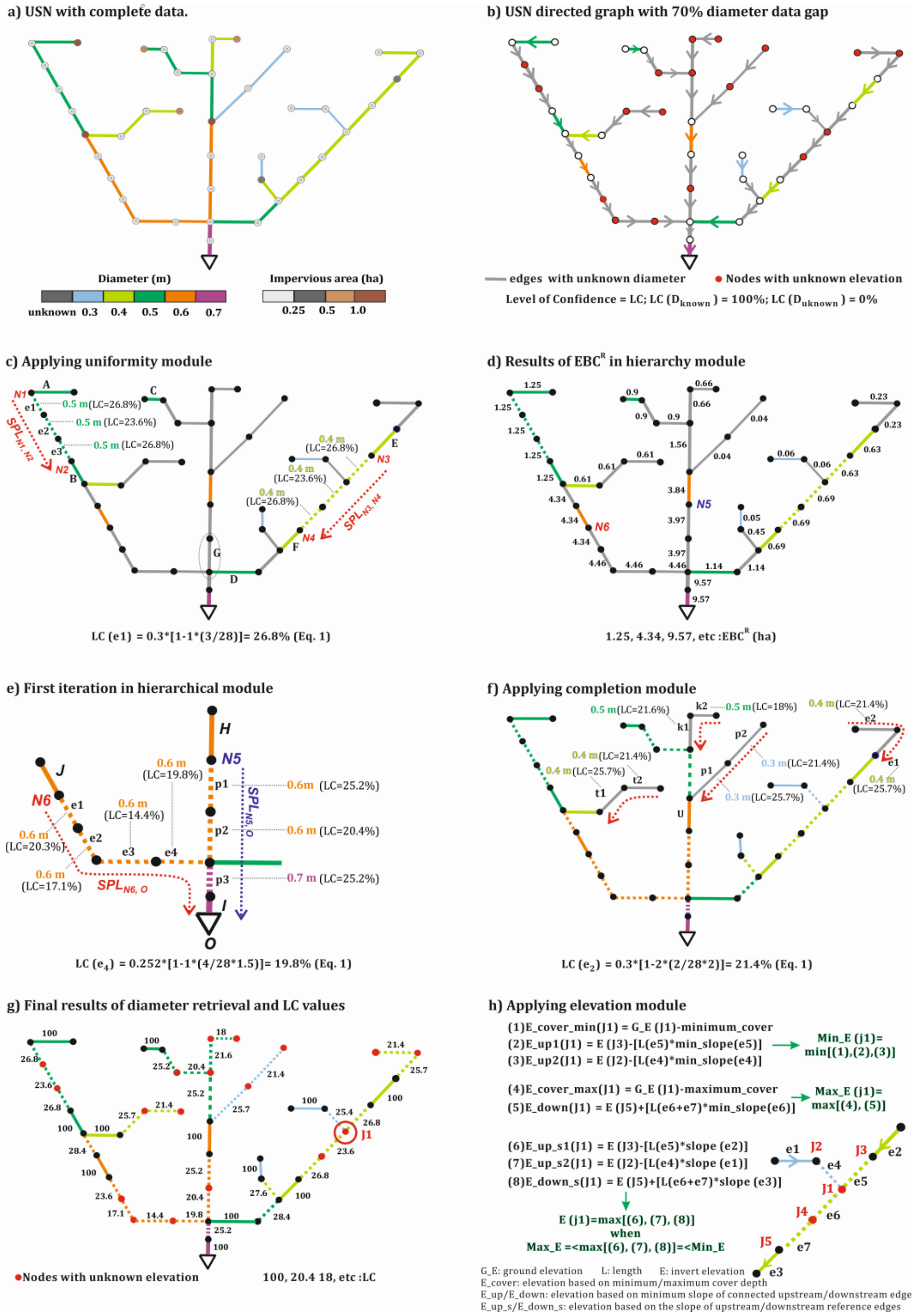

Figure 1. An illustrative example of the data retrieval model (Hajibabaei et al., 2025) [CC-BY license].

Case study

A real-world USN in Austria was selected as the case study. This USN is a branched network consisting of 237 junctions, with an average burial depth of 1.8 meters, and 236 sewers categorized into nine distinct diameter classes ranging from 0.25 m to 0.8 m. Detailed information about this USN can be found in Hajibabaei et al. (2025). This USN was investigated by artificially introducing data gaps generated by randomly removing sewer diameter and slope information ranging from 10% to 90%. Each gap scenario was repeated 100 times, resulting in 900 incomplete USN configurations.

Results and discussion

The results of applying the proposed model to the case study across nine data gap scenarios, each including 100 incomplete USNs, are presented in Figure 2. As shown in Figure 2a, the median value of the Mean Absolute Percentage Error (MAPE) between the existing and retrieved diameters remains below 10% for data gaps up to 80%. Additionally, the median value of the Mean Absolute Error (MAE) between the existing and retrieved invert elevations, shown in Figure 2b, stays below 0.27 m, representing 0.15% of the average burial depth in the actual network. These results indicate that the proposed model can accurately retrieve data for incomplete USNs, even with substantial data gaps. To further validate the model, Figure 2c presents a statistical analysis of the hydrographs at three selected nodes for the 80% data gap scenarios. As shown, the inflow hydrographs of the retrieved USNs closely match those of the existing network. In addition, the Nash–Sutcliffe model efficiency coefficient values in Figure 2c are between 0.99 and 1, demonstrating the model's ability to accurately estimate hydrodynamic characteristics. Regarding computational efficiency, for an 80% data gap, the complete retrieval process takes around 17 seconds, with limited considerations for code efficiency.

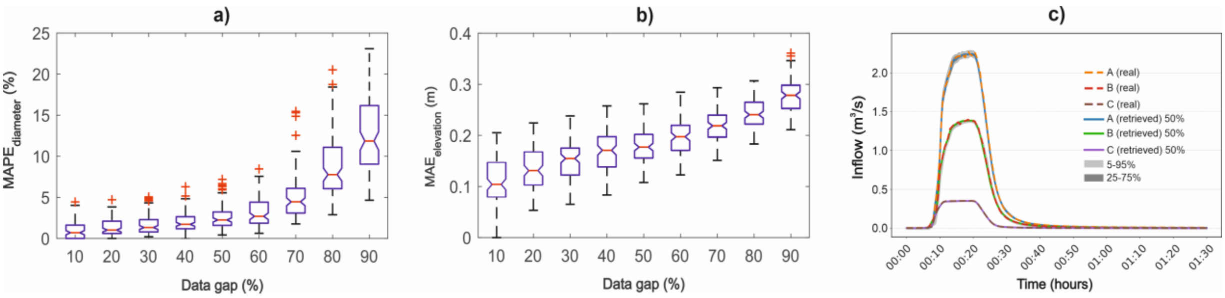

Figure 2. a) MAPE between the existing and retrieved sewer diameter; b) MAE between the existing and retrieved invert elevations; c) Validation of inflow hydrographs in 80% data gap for selected nodes (A, B, C). (Hajibabaei et al., 2025)

Conclusions and future work

This paper proposed a new graph-based retrieval model that automatically fills data gaps in incomplete USNs. By utilizing network topology and customized centrality metrics, the model efficiently recovers diameter and slope information up to the 90% data gap. Further results on the applicability of the proposed model to larger networks and its computational efficiency can be found in Hajibabaei et al. (2025). Future work will focus on extending the framework to include looped USNs.

Acknowledgement

This research was funded by the Austrian Science Fund (FWF) [10.55776/P36737].

References

Belghaddar, Y., Chahinian, N., Seriai, A., Begdouri, A., Abdou, R., Delenne, C., 2021. Graph convolutional networks: Application to database completion of wastewater networks. Water 13, 1681.

Hajibabaei, M., Hesarkazzazi, S., Minaei, A., Dastgir, A., Sitzenfrei, R., 2024. Using complex network theory for missing data reconstruction in water distribution networks. Sustain. Cities Soc. 101, 105114.

Hajibabaei, M., Hesarkazzazi, S., Sitzenfrei, R., 2025. Filling data gaps in urban drainage networks: An automated graph theory

framework for data collection and reconstruction. Water Res. 124272.
Hesarkazzazi, S., Hajibabaei, M., Bakhshipour, A.E., Dittmer, U., Haghighi, A., Sitzenfrei, R., 2022. Generation of optimal (de) centralized layouts for urban drainage systems: A graph-theory-based combinatorial multi-objective optimization framework. Sustain. Cities Soc. 81, 103827.
Montalvo, C., Reyes-Silva, J.D., Sañudo, E., Cea, L., Puertas, J., 2024. Urban pluvial flood modelling in the absence of sewer drainage network data: A physics-based approach. J. Hydrol. 634, 131043.