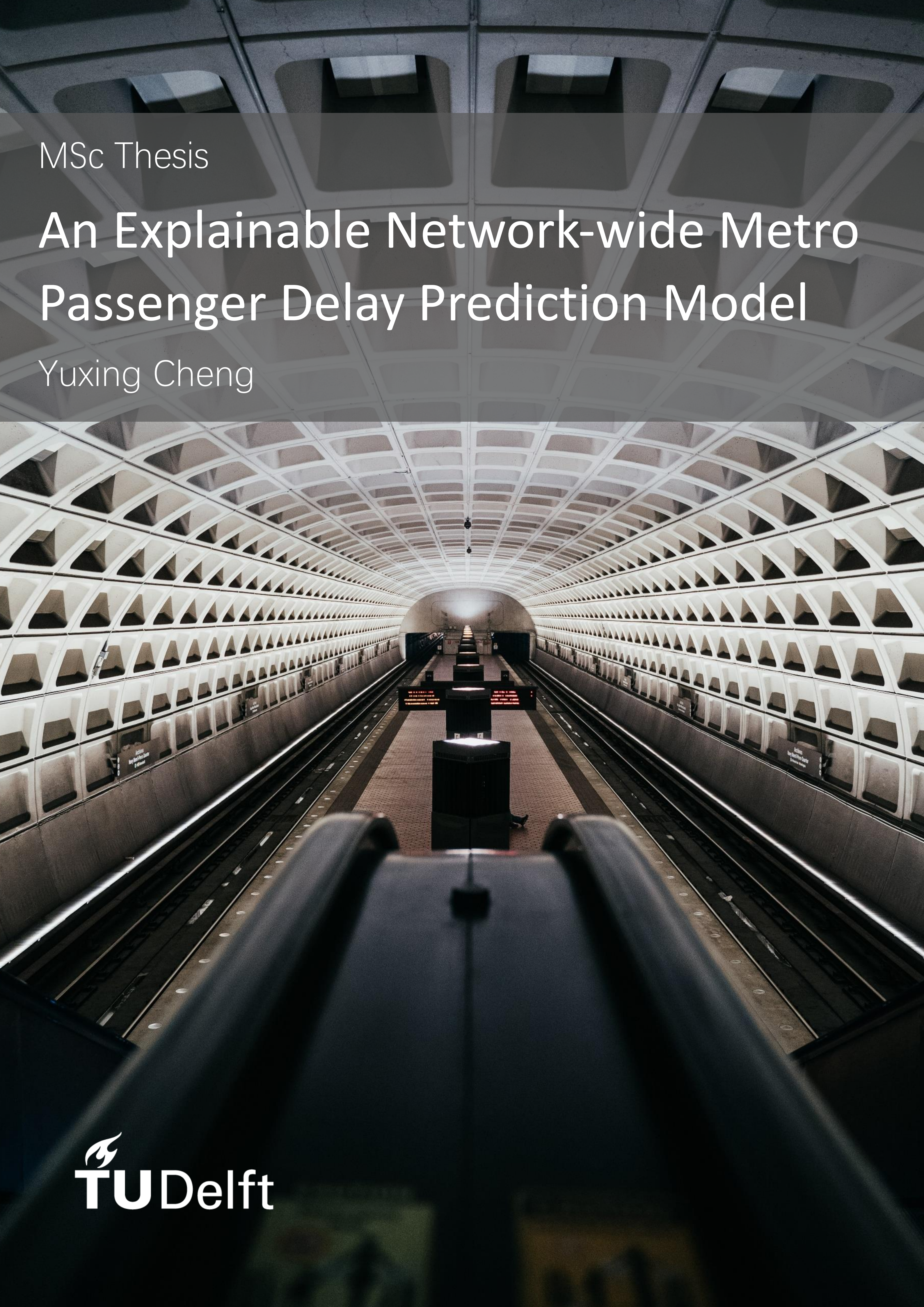
MSc Thesis

An Explainable Network-wide Metro Passenger Delay Prediction Model

Yuxing Cheng



TU Delft

An Explainable Network-wide Metro Passenger Delay Prediction Model

By

Yuxing Cheng

in partial fulfillment of the requirements for the degree of

Master of Science
in Civil Engineering

at the Delft University of Technology,
to be defended publicly on Monday, August 28, 2023.

Thesis committee:	Prof. Dr. ir. Oded Cats (Committee Chair)
	Dr. ir. Panchamy Krishnakumari (Daily Supervisor)
	Dr. Guopeng Li
Department:	Transport and planning
Student number:	5612233

Preface

This thesis stands as the culmination of my journey as a student in the Transport & Planning master's program at TU Delft. My passion for transportation studies has driven my continuous learning over the past six years, leading me to delve into my master's dissertation on this subject. Reflecting on my journey within the master's program, I'm grateful to realize that the last two years have marked the most significant progress in my academic career so far. Within these pages, I've compiled most of my recent research accomplishments, aimed at investigating the potential of explainable AI in enhancing insights for public transport performance analysis.

Please let me express my deep gratitude to my committee members. Firstly, I would like to appreciate my committee chair, Prof. Oded Cats, for offering me much guidance on this dissertation. Your dedication as a front-line researcher in public transportation has left a profound impression on me. Your expertise and research insights have inspired me to channel my interest into meaningful outcomes. Your assistance also greatly supported my PhD application, and our fruitful collaboration remains a cherished memory. Then, I express my sincere thanks to Dr. Guopeng Li for his constructive feedback and insightful perspectives on my work. Engagements with you have consistently aided me in pinpointing areas for improvement and in focusing my efforts on enhancing my present and future endeavors. You are always the model for my journey to follow as a PhD student and future researcher. Next, I would express my heartfelt gratitude to my daily supervisor Dr. Panchamy Krishnakumari. I greatly enjoy working together with you. Your mentorship always helps me overcome obstacles and challenges I met. As the mentor I've collaborated with most closely in the past year, your unwavering support during my first conference paper and dissertation has played a crucial role. These modest achievements have laid the foundation for both my forthcoming research pursuits and my PhD position application. Without your dedicated guidance, my progress over the past year, which is so significant to me in my just-started research journey, would not have been possible. In closing, I extend my appreciation to my friends and family for their continued encouragement and support, and for always believing in me.

The insights gleaned from the papers are always shallow only if exploring them by hand. Through this dissertation, I've come to realize that the path ahead might be tortuous, yet the prospect is bright. There will be a long way to go for an incoming PhD student to an independent researcher. I am fortunate to have invested two transformative years in the Department of Transport and Planning at TU Delft, conducting research on my dedicated topic under the guidance of my committee. With this culmination, I hold my belief that a promising tomorrow awaits.

Yuxing Cheng
Delft, August 2023

Executive Summary

Background and research objectives

In response to the need for reliable service planning, efficient schedule adjustments, and more effective operational strategies, recent years have witnessed a notable emphasis on offering operators and passengers more accurate delay prediction. To achieve improved accuracy in predicting transit delays and ensure consistent service reliability, various prediction methods have been developed, prominently among them being data-driven approaches such as Deep Neural Networks (DNNs). These data-driven models have showcased superior predictive precision compared to conventional methods. However, the lack of explainability in these models poses challenges for transit providers, hindering their ability to comprehend predictions and make informed decisions.

Understanding prediction mechanisms is vital to identify delay causes and take proactive measures. By analyzing delay-contributing factors, providers can optimize operations and enhance service reliability. The inscrutability of DNNs and similar models undermines trust, impacting decision-making support. While event-based models offer easier-to-interpret delay predictions, they fall short in capturing intricate spatial-temporal interactions.

Explainable AI (XAI) has arisen as a solution to this issue. XAI aims to clarify complex models' predictions by providing insights into their decision processes, potentially aided by domain expertise. Previous studies have sought to uncover reasons for metro delays, encompassing exogenous (external events) and endogenous (internal factors) causes. Yet, these studies rely on diverse input datasets, limiting universal insights into delay prediction.

This study focuses on achieving comprehensive short-term metro passenger delay prediction using a deep graph neural network, which only requires passenger origin-destination data, metro schedule, and network topologic information as input. It centers on the Washington DC metro network, aiming to explain prediction outcomes, understand spatial-temporal delay patterns, and assess the dynamic correlations of delay occurrence. The main research question delves into extracting spatial-temporal explanations from data-driven black-box models based on historical passenger travel data. To answer this question, this research subdivides into investigating spatial-temporal explanatory variables for network passenger delay, gauging the impact of historical delay data's spatial-temporal features on predictions, and scrutinizing spatial-temporal correlations across different metro network segments throughout the day.

Research approach

The research approach involves three steps: Data processing, Delay prediction, and Prediction explanation. Firstly, the Washington DC metro network's topological information and passenger data drive the estimation of delay occurrence and its mapping onto the network structure. This informs the construction of spatial-temporal graphs of delay occurrences. A Spatial-temporal Graph Neural Network (STGCN) is then utilized to provide delay prediction. After that, post-hoc analysis on the black box STGCN model provides explanatory variables of the model, which reveal the significance of spatial-temporal features in historical delay data. Finally, based on the explanatory variables, we applied pattern recognition techniques and purposed a series of data processing to explore the dynamic spatial-temporal correlation of delay across various locations.

The process of transferring the explanatory variables to the explanation of delay occurrence dependencies is the essence of the purposed approach, which is depicted in Figure 1. First, the black box STGCN model provides the prediction result of each node as time series data. Each time step has a corresponding predicted delay value of that node. Secondly, we select the prediction outcome of one node at a specific time step as the target. The post-hoc analysis is implemented to reflect how the input time-series delay data of each node in the graph contribute to the target prediction result we are interested in. Thus, each node has a time-series importance score that reflects each node's contribution dynamics on time scope. These contribution scores reflect the node-to-node dependency at a specific time step. Finally, we pile the importance score data of all the nodes to form the explanatory variable matrix of the target prediction result, which is the base for spatial-temporal dependencies analysis.

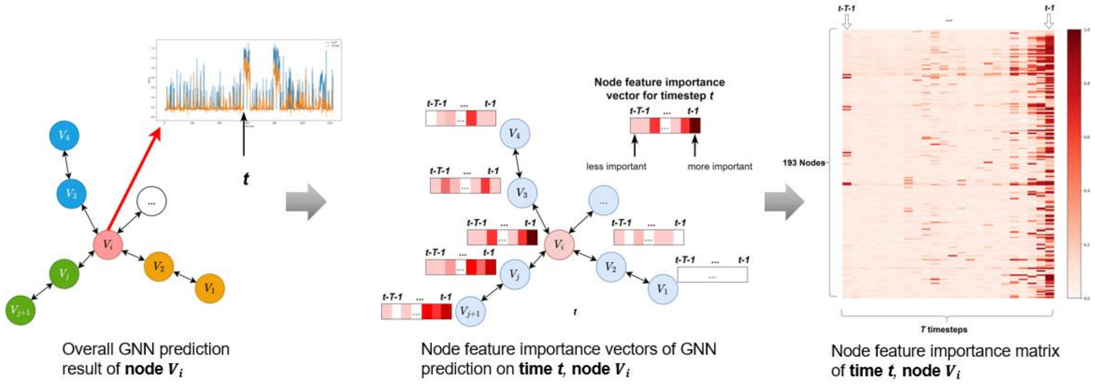

V_4
V_3
...
V_i
V_j
V_{j+1}
V_2
V_1
t
Overall GNN prediction result of **node V_i**
t-T-1 ... t-1
Node feature importance vector for timestep t
less important
more important
t
Node feature importance vectors of GNN prediction on **time t, node V_i**
t-T-1
...
t-1
193 Nodes
T timesteps
Node feature importance matrix of **time t, node V_i**

Figure 1 Mechanism of GNN Explainer explains predict the result of a specific node $\boldsymbol{V_i}$ at a particular timestep t

Case study

The dataset employed for the case study was sourced from WMATA and provides a comprehensive perspective of the metro network. It encompasses rail trip data, detailing journeys between origin and destination stations as well as intermediate movements. This dataset spans 277 days, commencing from August 19, 2017, to May 22, 2018. Additionally, passenger journey information from smart card records was integrated, revealing passenger origins and destinations. By combining these details with rail movements, individual passenger routes were inferred. To facilitate analysis, the dataset underwent pre-processing steps previously defined, transforming disaggregated raw data into a network-wide perspective.

Upon the completion of data preprocessing, a refined dataset spanning 260 days is derived. This dataset comprises two distinct facets: the topological details of 6 lines, 91 stations, 184 links, and 9 transfer stations, coupled with the estimated average delay experienced by passengers at various travel stages, encompassing waiting delay, track delay, transfer delay, and their linear sum. This data is meticulously charted onto a link graph that mirrors the directional pathways within the metro network.

The explanatory variables derived from the purposed post-hoc explanation approach reveal how the model makes specific predictions based on input time-series data. We further explore how these variables could uncover the spatial-temporal correlation of delay occurrence among different segments in the metro network.

Conclusion

By dissecting the explanatory variables extracted from the delay prediction of a specific day (May 18th, 2018), discernible patterns emerged in terms of the spatial-temporal significance of features and the interconnections between delay occurrences across different sections of the metro network. These revelations unveiled three notable temporal significance patterns: "Recent Dominance," "Partial Long Lasting," and "Temporal Global Relevant”. Similarly, the analysis identified three distinctive spatial significance patterns named "Key Nodes Dominate," "Spatial Imbalance," and "Spatial Global Relevant”. These insights showcase the varying contributions and impacts of different segments within the network. Notably, the network segments with significant contributions to the delay prediction are usually located near the transfer station and the terminal station. An exploration of the temporal distribution of these patterns highlights their close correlation with peak and non-peak hours, offering valuable insights into how these patterns are intertwined with the network's operational dynamics.

Furthermore, spatial-temporal correlations among delay occurrence are revealed by the following findings:

- Spatial-temporal correlation matrix analysis highlights imbalances in delay influence among nodes.
- Spatial-temporal correlations appear fluctuating strengths, with long-lasting patterns related to key nodes and peak periods.
- Different clusters exhibit significant interdependencies during peak hours, revealing tidal patterns.
- Nodes near terminal and transfer stations demonstrate stronger correlations and interdependencies.

Furthermore, our findings align with prior research and demonstrate the presence of spatial-temporal correlation patterns in delay occurrence, consistent with similar phenomena observed in traffic networks. The dynamic correlation underscores the significance of core node clusters, influencing neighboring nodes and contributing to delay propagation throughout the network.

The insights derived from this study hold considerable potential to assist metro operators to better understand the intricate dynamics of delay occurrences in the metro network. Operators can gain a comprehensive view of how delays propagate and impact various parts of the system. This understanding enables more informed decision-making, allowing operators to pinpoint critical nodes that exert significant influence on delay propagation. With this knowledge, operators can proactively implement measures to manage congestion, isolate bottleneck nodes, and optimize schedules during peak periods. Moreover, the identified patterns provide valuable guidance for developing targeted strategies to mitigate delays and enhance overall network reliability. Incorporating these insights into operational planning and decision-making processes can lead to more efficient and effective management of the metro system, ultimately improving the passenger experience and system performance.

This study's advantage lies in its ability to capture complex delay dynamics using basic passenger travel and schedule data through a data-driven model. However, the model's simplicity may limit its capacity to comprehensively capture delay occurrences, especially due to imbalanced data distributions. Future research could address this limitation by improving the model's performance with more balanced datasets.

In the future, comparing these phenomena across different metro networks and investigating the relationship between node cluster correlations and cluster size could provide valuable insights. Incorporating additional data, such as passenger counts, into the model might enhance its accuracy as well. These insights hold the potential for advancing delay recovery and scheduling management strategies.

Contents

List of Figures

1. Introduction

1.1 Background

In recent years, the demand for enhanced service reliability from transit users has grown substantially, significantly influencing transit performance, ridership numbers, and overall user satisfaction (Bešinović, 2020.; Krishnakumari et al., 2020). The comprehension of delay occurrences within the transport system provides valuable insights that can facilitate improved service planning, efficient schedule adjustments, and more effective operational strategies. When future delays are anticipated, the proactive management of delays and optimized resource allocation becomes achievable. To ensure operational accuracy and service reliability, numerous prediction approaches have emerged in the realm of public transport delay prediction. Data-driven approaches, such as Deep Neural Networks (DNNs), have been shown to achieve higher prediction accuracy than traditional methods (Spanninger et al., 2022; Tang et al., 2022; Tiong et al., 2023). However, the lack of explainability in these models is a significant challenge for transit providers. How the model arrived at its predictions and makes informed decisions based on input information is covered. This would limit their ability to understand the underlying dynamic of delay occurrence and propagation, and make well-informed decisions such as rescheduling and delay noticing based on this crucial information.

Understanding how the model makes its predictions is crucial for identifying the dynamics of delay occurrence and taking proactive measures to mitigate them. By analyzing the patterns of factors that contribute to delays, transit providers can make informed decisions about how to optimize their operations, allocate resources, and improve service reliability. Moreover, the lack of explainability in data-driven forecasting models, such as DNNs, is a significant concern for transit providers, as it hinders their ability to explain the model's predictions to stakeholders. This limitation can erode trust in the model and make it less reliable for supporting decision-making. Furthermore, although traditional approaches like event-based modeling that explicitly capture dependencies of train events (departure, arrival, and pass-through) can provide easier-to-interpret predictions for train delays, data-driven approaches can capture complex nonlinear spatial-temporal variable interactions that are difficult to achieve with event-based approaches (Spanninger et al., 2022; Heglund et al., 2020). Therefore, it is essential to understand how these black box models make their predictions and gain insights into the factors contributing to them.

In recent years, the concept of Explainable AI (XAI) has emerged as a potential solution to this problem. XAI aims to make complex machine learning models more interpretable and transparent by providing insights into how they arrive at their predictions (Adadi and Berrada, 2018; Arrieta et al., 2019). Thus, XAI companies with domain knowledge and expertise in the dynamics of the railway system have the potential to interpret the black-box model's prediction results with explanatory variables.

Previous studies have attempted to identify various exogenous and endogenous causes of metro delay occurrences based on black-box prediction results. The factors contributing to delays can vary depending on the data attributes analyzed. Exogenous factors refer to external events or circumstances that may impact the metro system's performance, such as weather conditions, accidents, or public events. Endogenous factors, on the other hand, are internal factors within the metro system, such as the frequency and reliability of trains, the capacity of stations, the cascading effect of delays, and the maintenance schedule (Cats and Hijner, 2021; Louie et al., 2017; Yap and Cats, 2021). However, existing research on influential factors of delay occurrence mostly relies on multiple data sources with various data attributes involved, resulting in limited and case-specific knowledge gained from delay prediction (Spanninger et al., 2022; Tiong et al., 2023; Zhang et al., 2023). This limitation highlights the need for research focused on extracting delay prediction explanations based on attributes existing in metro networks with only topological information, scheduling data and elementary passenger travel origin-destination data. By understanding how the model makes specific delay predictions, we can gain knowledge of delay occurrence characteristics, spatial-temporal interrelationships between delays occurring on different segments of the entire network, and the spatial-temporal importance of specific historical data that contributes to the delay prediction result. This knowledge can enhance the reliance on operators' decision-making to increase the overall service quality of the metro system in the long term.

The objective of this study is to provide a comprehensive understanding of the short-term metro passenger delay prediction of deep graph neural network, based on the topological information of metro network and passenger travel data. The study will focus on conducting a detailed case study of the Washington DC metro network, to explain the prediction results and identify the spatial-temporal characteristics and importance of delay occurrence prediction in this specific metro network.

1.2 Objective and research questions

The overall objective of this study is to explore the spatial-temporal correlation of short-term networkwide metro passenger delay prediction by an explainable AI approach. To achieve this objective, the main research question is needed to be solved:

- **What insights on spatial-temporal correlation could be acquired from the data-driven black-box prediction model for network passenger delay predictions based on historical passenger travel data?**

To answer the main research question, the follow sub-questions should be answered:

- **What are the spatial-temporal explanatory variables of network passenger delay occurrence prediction?**

- **How does the spatial-temporal feature of historical delay contribute to the model prediction?**

- **How does the spatial-temporal correlation of different metro network segments vary throughout the day?**

1.3 Approach

The overview of the research approach is as Figure 1 shown, which is divided into three steps: Data processing, Delay prediction, and Prediction explanation. In this research, the case study will be based on the metro network in Washington DC, USA, and its data will be described in section 4.1. In the first step, the input dataset includes the metro network's topological information and smart card data of metro passengers, which are used to estimate the delay occurrence and map it on the network structure. Based on the estimated delay value, spatial-temporal graphs of metro network delay occurrence are constructed. In the second step, a specific GNN model called STGCN (Spatial-temporal graph neural network) will be used to make delay explainable predictions of three categories (waiting time delay, transfer time delay, and track delay) based on a post-hoc explanation approach. This will extract the spatial-temporal feature importance as explanatory variables of historical delay data. Finally, in the third step, we explore the pattern of spatial-temporal feature importance of historical delay with pattern recognition technique, and derive the dynamic of spatial-temporal correlations of delay occurring at different locations. These results provide insights into the internal correlation characteristic of delay in the metro network.

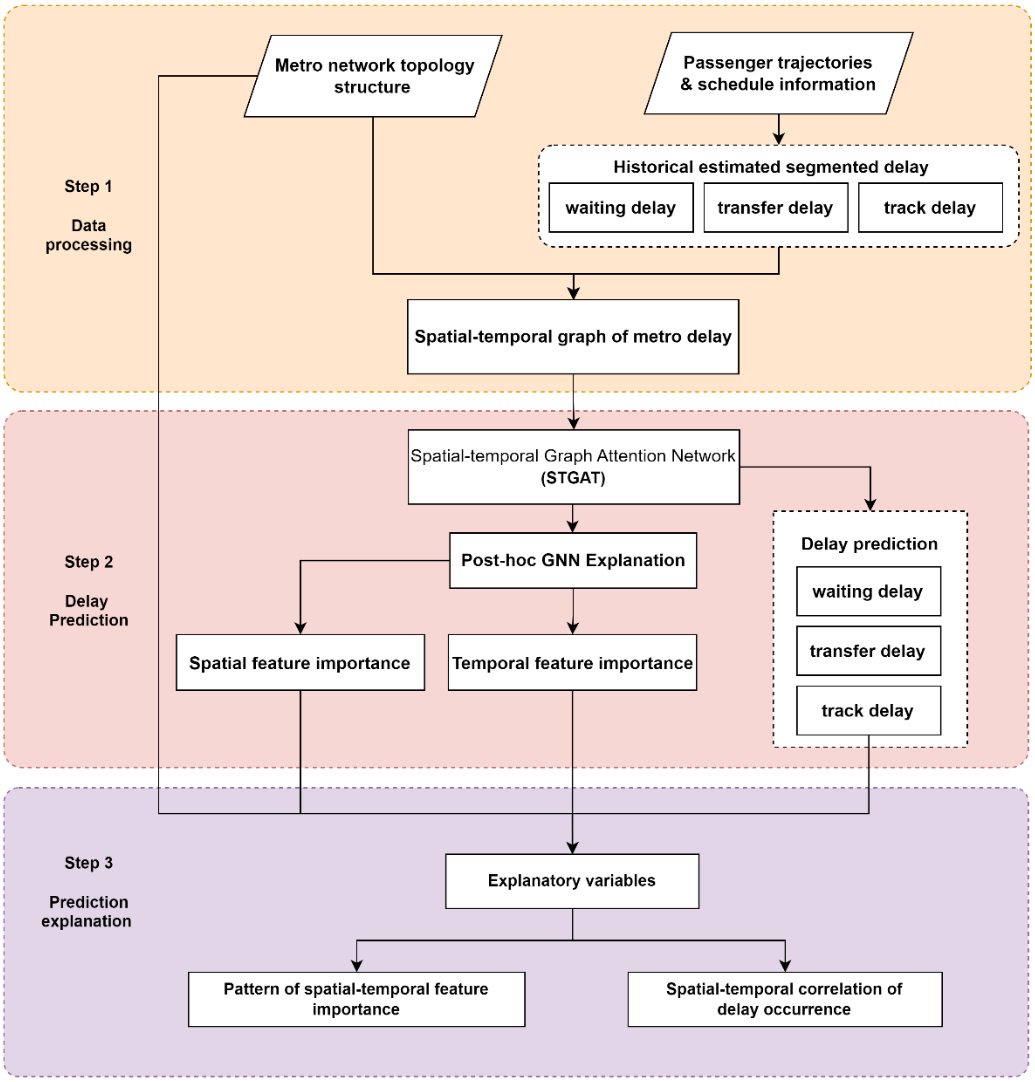

Step 1
Data processing
Metro network topology structure
Passenger trajectories & schedule information
Historical estimated segmented delay
waiting delay
transfer delay
track delay
Spatial-temporal graph of metro delay
Step 2
Delay Prediction
Spatial-temporal Graph Attention Network (STGAT)
Post-hoc GNN Explanation
Spatial feature importance
Temporal feature importance
Delay prediction
waiting delay
transfer delay
track delay
Step 3
Prediction explanation
Explanatory variables
Pattern of spatial-temporal feature importance
Spatial-temporal correlation of delay occurrence

Figure 1 Overview of research methodology

1.4 Thesis structure

The thesis is organized as follows: In Section 2, an extensive review of relevant literature is presented to establish the research context. Section 3 outlines the methodology employed, encompassing the mapping of passenger delay onto network links, passenger delay prediction using a black-box model, and post-hoc explanation extraction methods. Section 5 presents the results of the model prediction explanation analyzed from different angles, including the identification of spatial-temporal feature importance of historical delay data and the exploratory analysis of network topological structure. Finally, in Section 6, the study's findings and results are discussed, including their alignment with prior research, the study's contributions and limitations, and recommendations for future research directions.

2. Literature review

To explain the black box model's output of metro delay prediction, it is essential to review the existing research on how Deep Neural Networks (DNN) achieve accurate prediction, what approaches are available to explain their output, and what potential explanations for delay correlations can be expected.

The scope of the literature review covers the research that proposes approaches that can be utilized for transport network-wide forecasting, including but not limited to the metro network. Despite the nuanced differences between metro and railway service in geographical scale of the system and service frequency, there is a great deal of similarity between these transportation systems. Both metro and railway systems rely on similar data sources, such as passenger counts, train schedules, and real-time tracking data, to make predictions about potential delays. Besides, the techniques used for delay prediction in both metro and railway systems are based on machine learning and statistical modeling approaches, such as artificial neural networks and decision trees (Spanninger et al., 2022; Tiong et al., 2023). These techniques can be applied to both transportation systems with little modification, as the underlying principles for prediction are the same. Therefore, the insights and methodologies from the literature on railway or similar transport network delay prediction (road traffic network) can be applied to metro delay prediction.

2.1 Data-driven approach for spatiotemporal traffic delay forecasting.

Tiong et al. summarize previous studies on train delay modeling into two major categories: event-driven approach and data-driven approach. It is argued that the data-driven method is anticipated to deliver state-of-the-art results in terms of prediction accuracy, as compared to the event-driven approach. The two most widely used data-driven methods for train delay prediction are decision trees and deep neural networks (DNN) (Tiong et al., 2023). The main advantages of DNNs are their ability to automatically learn complex, non-linear relationships between input features and the output variable, without relying on hand-engineered features. This makes DNNs more flexible to different datasets with varying attributes, as they can adapt to the data and capture intricate patterns that may not be easily identifiable using traditional machine learning models such as decision trees or supervised regression. In addition, DNNs have shown superior performance in the task of railway delay prediction (Spanninger et al., 2022). On the other hand, decision trees and other machine learning models are limited in their ability to handle complex relationships and may require careful selection and engineering of input features.

Recent studies have shown that graph neural networks (GNN), specifically spatial-temporal graph convolutional networks (STGCN) first purposed by (Yu et al., 2018), are particularly effective for predicting delays in metro network systems. This is because metro networks can be modeled as spatial-temporal graphs, where the nodes represent stations, and the edges represent connections between stations. The time series data of metro network delays can be represented as signals on the graph nodes. STGCN achieves the prediction by using graph convolutions, which can learn the spatial relationships between the nodes in a graph, and temporal convolutions, which can capture the time-dependent patterns in the data. This approach enables STGCN to model the complex interactions between different nodes in the graph over time, leading to more accurate and reliable node predictions(Dai et al., 2020; Yu et al., 2018). Such a model has been proved well-suit for traffic forecasting. For instance, (Heglund et al., 2020) applied STCGN on the British railway network and outperforms other statistical models which do not explicitly account for interactions on the rail network. Besides, a series of research purposed various novel network architectures based on STGCN, and update the performance of the traffic forecasting on baseline road traffic condition datasets including PEMS, METR-LA, and UVDS (Bui et al., 2021; Li and Zhu, 2021; Shao et al., 2022). Among these STGCN-based model improvements, some of the approaches may require multi-source data with more advanced attributes beyond elementary attributes in the mentioned baseline datasets, including GPS trajectory data, weather data, and multimedia data (Li and Zhu, 2021).

However, some studies have pointed out that a major drawback of data-driven models is their black-box nature. These GNN models with stronger predictive capabilities usually provide unexplainable outcomes, leaving researchers and practitioners unsure of the actual reasoning behind a prediction(Arrieta et al., 2019; Tiong et al., 2023). Although some research may offer partial explainability for GNN-based prediction results, such explanations are often specific to a particular dataset and may not apply to other networks with different input data attributes (Zhang et al., 2023). To gain insights and explanations from these models, extra efforts are required to interpret and analyze their results, and additional domain knowledge is required to interpret and analyze the output generated by the model(Menno Yap and Oded Cats, 2021).

2.2 Explainable Artificial intelligence (XAI) in public transport delay prediction

The lack of explainability in deep learning models is a growing concern for developers and engineers across various fields. This is particularly true for black-box DL models that are utilized to make critical predictions. As a result, stakeholders in the field of artificial intelligence (AI) are calling for greater explainability in these models (Adadi and Berrada, 2018; Arrieta et al., 2019). This has led to the development of Explainable AI (XAI) to address this challenge. To tackle this challenge, Explainable AI (XAI) has been developed. The primary objective of XAI, though varying based on the intended audience, is to create data-driven techniques that yield more explainable models while maintaining high learning performance. This approach enables human users to understand, appropriately trust, and effectively manage the emerging generation of AI partners (Arrieta et al., 2019).

In the context of predicting public transport delays, such as those in metro or railway systems, the explainability of the delay prediction result is vital to the metro operator and passengers(Dalmau et al., 2021; Rößler et al., 2021). By answering how specific features of input data contribute to the prediction result, XAI can guide the operator in designing and operating the public transport system more efficiently. Moreover, an explainable model can help verify whether the model behavior is as expected and establish a trusting relationship with the model user (metro operator). Therefore, developing explainable AI models for predicting public transport delays can be highly beneficial for both the operator and the passengers.

In the field of XAI, two types of models are widely recognized in terms of explainability: self-explainable models, which are inherently explainable by design (e.g., decision trees or linear models), and post-hoc explainability models, which require external XAI techniques for explanation (Adadi and Berrada, 2018; Arrieta et al., 2019). Self-explainable models have transparent structures that are easily understood and interpreted by humans. Conversely, post-hoc explanation methods do not modify the underlying model but instead scrutinize the model's behavior to identify the most crucial features that influence the predictions. Complex models like neural networks or deep learning models are frequently regarded as black boxes because their internal operations are not transparent to human understanding. As a result, post-hoc explanation methods are employed to decipher these models' reasoning behind their predictions.

Among the post-hoc explanation methods, LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) are the two classic methods for feature relevance explanation that are widely implemented in the previous research (Fontoura et al., 2020; Lundberg et al., 2019). Both LIME and SHAP generate feature importance weights that help understand the impact of different input features on a model's predictions. LIME provides local, interpretable explanations for individual predictions, while SHAP computes the contribution of each feature to the prediction across the entire dataset. Both methods are model agnostic, meaning they can be used with any type of machine learning model. However, they do have certain limitations. Firstly, LIME and SHAP can only explain predictions in terms of the input dataset features that the model was trained on, which may not capture all the relevant factors in the real world that affect the outcome. In the context of explainable train/metro delay prediction, the explanations provided by these methods are restricted to the attributes included in the input dataset. Consequently, the key factors that lead to delay occurrences may differ depending on the particular input dataset being used for analysis (Rößler et al., 2021; Taleongpong et al., 2022, 2022; Tiong et al., 2023; Wales and Marinov, 2015). Additionally, it is challenging to include all the possible external or internal factors in the input dataset, which may have a potential impact on the outcome distribution characteristics. Those unexpected factors may not be captured by the input features and therefore may not be explainable by LIME and SHAP (Kumar et al., 2020).

Previous studies on delay prediction have primarily focused on identifying the causes of delays from various aspects, such as weather or unexpected disruption events (Lee et al., 2016; Yap and Cats, 2021; Wei et al., 2015). However, such explanations for forecasting may not necessarily imply the causation of delay occurrence, which may not be sufficient to provide clear and direct insights for operators to prevent delay occurrence. To address this limitation, the focus should shift to the internal correlation and dependencies of delay distribution itself, rather than explore delay causation by analyzing exogenous variables. Some previous research has touched upon this idea. For instance, Zhang et al. proposed an end-to-end spatial-temporal deep learning model for multistep prediction without relying on exogenous variables like weather or feature engineering, which could capture the temporal heterogeneity of traffic patterns. However, previous similar studies sparsely provide a further post-hoc explanation of the model that could reveal the relative spatial-temporal feature importance of different segments of the network or different traffic variables.

2.3 Spatial-temporal correlation of transport networks delay occurrence

In recent years, the spatial-temporal correlation of metro delay occurrence has emerged as a subject of significant research interest, complementing the multitude of latent reasons for delay occurrence revealed by previous studies. While this type of explanation does not necessarily imply causation, it can still provide valuable insights for improving the reliability of the metro system (Cats and Hijner, 2021). By integrating information about the spatial-temporal inter-correlation of delays into delay prediction models, metro operators can gain a deeper understanding of the relationships between different system components' reliability. This understanding enables them to proactively take measures to prevent delays from happening. Studies exploring the mechanism of spatial-temporal correlation within rail/metro networks have investigated various angles, such as delay cascading effects, network resilience, and delay mitigation (Krishnakumari et al., 2020; Menno Yap and Oded Cats, 2021; Wales and Marinov, 2015).

Moreover, similar correlation patterns likely exist in passenger transport networks across multiple modes, including road traffic networks and metro networks, as partially revealed by existing research (Faroqi et al., 2017; Su et al., 2017; Sun et al., 2011; Yang et al., 2017). For example, studies in road networks have identified positive or negative spatial-temporal correlations of traffic links, depending on traffic regimes (Ermagun et al., 2017). Other research has highlighted the impact of jammed cores in traffic networks, which can trigger cascading effects and congest neighboring nodes (Petri et al., 2009). The result of this study was derived via a minimal network flow model. And it suggested from the theoretical point of view that, quarantining the congested node in the network will help diminish the congestion propagation. An urban congestion pattern recognition approach has also been proposed, synthesizing the road network of a city into a few consensual 3D speed maps, revealing intercorrelation patterns across different network segments on different days (Lopez et al., 2017).

While extensive efforts, particularly those focused on road traffic data, have been devoted to exploring spatial-temporal correlation patterns to improve traffic status forecasting, there remains a gap in investigating the specific contribution of spatial-temporal correlation in individual network segments to overall traffic status, such as delays, congestion, or passenger flow (Lu and Lin, 2019). Additionally, little attention has been given to understanding how delays at different stages of an individual's metro travel collectively contribute to the overall delay experienced by passengers.

This gap may be attributed to previous research primarily focusing on enhancing prediction accuracy rather than gaining insights into network vulnerabilities. Furthermore, the data available for these studies often consist of limited attributes, resembling initial data obtained from sensor methods, such as flow counts or smart card data. Incorporating data inference approaches like those proposed by Krishnakumari et al. (2020), which map estimated passenger delays onto network edges, could greatly enrich our understanding of passenger travel delay experiences and network delay correlations. Addressing these research gaps will be instrumental in developing more robust and efficient strategies for delay prevention and management in metro systems.

3. Methodology

In this section, the applied method is described. Section 3.1 introduces how the passenger trajectory delay is estimated. Section 3.2 conceptually stated the mechanism of a black-box graph neural network for graph time-series data prediction. These two sections cover the method for producing the input of this study. After that, the post-hoc explanation approach that led to model explanatory variables is described in Section 3.3, and the pattern recognition and classification approach for exploring spatial-temporal patterns of delay occurrence is described in Section 3.4.

3.1 Metro passenger delay estimation

The metro passenger delay estimation method proposed by Krishnakumari et al. is capable of mapping passenger delay into network elements. This approach decomposes the delay along a passenger trajectory into its corresponding track segment delay, initial waiting time, and transfer delay, as Figure 2 indicated. The data required for the delay estimation is the Passenger-to-Train-Assignment (Zhu et al., 2017) result derived from the ODX method described in (Sánchez-Martínez, 2017). Therefore, the initial data required for metro passenger delay estimation is the smart card data, which has been reviewed by Gordon et al.（2013). The work introduced in section 3.1 is processed by Krishnakumari et al. (2020).

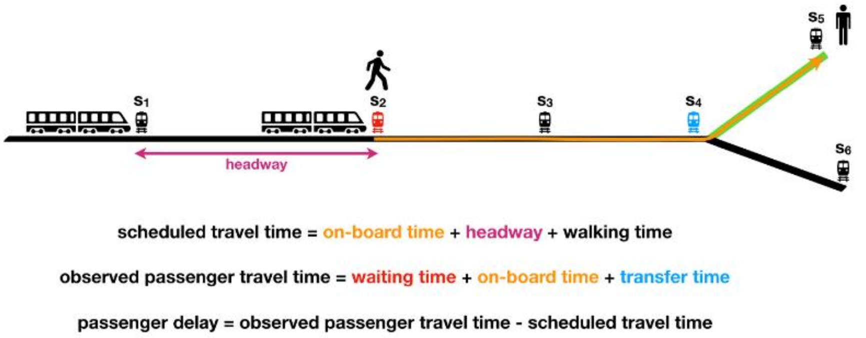

S5
S1
S2
S3
S4
S6
headway
scheduled travel time = on-board time + headway + walking time
observed passenger travel time = waiting time + on-board time + transfer time
passenger delay = observed passenger travel time - scheduled travel time

Figure 2 Schematic picture of scheduled and observed travel time between station s_2 and s_5 (Krishnakumari et al., 2020).

Without losing the generality, the metro network can be modeled as a directed graph $G = (S, E, X(t))$, where the node set S contains interconnected stations $s_1 \dots s_n$, the edge set E represents the track segments between stations, and $X = x_1 \dots x_T$ is the set of attributes of the whole network in time scale. Any train that runs on the network has a specific route consisting of multiple stations. The link set for the same journey is defined as $l_1 \dots l_n$. Besides, the smart card dataset includes the trip origin and destination information of passengers $n \in 1 \dots N$, where N is the total number of passengers. The passenger $n \in 1 \dots N$ experienced travel trajectory from origin station S_o to destination station S_d is denoted by $r_{S_o,S_d,n}$.

The initial steps in passenger delay estimation involved determining the passengers who experienced delays during their trips and the extent of the delay incurred. This was achieved by comparing the scheduled travel times to the observed travel times as Equation 1 descript:

$$\boldsymbol{d^{n}_{S_o,S_d,k}} = \begin{cases} \boldsymbol{t^{n}_{S_o,S_d,k} - \tilde{t}^{n}_{S_o,S_d,k}} & \boldsymbol{if\ t^{n}_{S_o,S_d,k} > \tilde{t}^{n}_{S_o,S_d,k}} \\ \boldsymbol{0} & \boldsymbol{otherwise} \end{cases} \tag{1}$$

Where $d^{n}_{S_o,S_d,k}$ refers to the delay of passenger n experienced during the trip from origin station S_o to destination station S_d at the time period k, in which the passenger checked in at S_o; $t^{n}_{S_o,S_d,k}$ refers to the observed travel time for passenger n departing at time period k and can be obtained by finding the difference between the tap-in and tap-out time of that passenger; $\tilde{t}^{n}_{S_o,S_d,k}$ is the summation of in-vehicle time of the estimated route, the transfer time if a transfer is included, and the headway of each line traversed. A positive value of $d^{n}_{S_o,S_d,k}$ indicates that the passenger was delayed, while a negative value would be set to 0, as it indicates the passenger arrived earlier than scheduled. Incorporating the headway was deemed necessary to account for the randomness in the arrival of passengers concerning the schedule, which is a plausible assumption for metro networks with short headways.

The aim is to estimate the delay that occurred at each link and station in a directed metro network. Specifically, the delay incurred at the origin station and transfer station of a passenger was distinguished. To accomplish this, Equation 2 is formulated that incorporated the delay experienced at the initial station, transfer stations, and links. The result should be equal to the overall delay at the passenger's destination for a single journey.

$$\boldsymbol{d^{n}_{S_o,S_d,k} = d^{wait}_{S_o} + \sum_{S_x,S_{x+1}}^{E} b_{S_x,S_{x+1}} \cdot d^{on-board}_{(S_x,S_{x+1}),k} + \sum_{S_i}^{I} b_{S_i} \cdot d^{trans}_{S_i,k}} \tag{2}$$

$d^{n}_{S_o,S_d,k}$ is the total estimated delay of passenger n traveling between origin station S_o and destination station S_d, departed at period k; $d^{wait}_{S_o}$ is the passenger initial waiting time at origin station S_o; $b_{S_x,S_{x+1}}$ is a binary value indicating whether the track between stations S_x and S_{x+1}; $d^{on-board}_{(S_x,S_{x+1}),k}$ is the on-board delay on the track between stations S_x and S_{x+1}; b_{S_i} is a binary value indicating whether transfer station S_i is path of the path; and $d^{trans}_{S_i,k}$ is the delay at transfer station S_i. The definition of passenger delay may vary depending on the fare validation scheme employed. For instance, in a surface bus system, the initial waiting time delay cannot be discerned from the AFC data. Therefore, in such cases, the waiting delay component will be excluded from equation (1). Besides, the passenger delays estimated for each network element are assumed to be common or average for all individuals.

The process of delay estimation involved creating a set of equations to account for each passenger. However, solving this set of equations continuously proved impossible. To overcome this challenge, the data was discretized over time, with 30-minute time slices being used. This decision was based on the fact that all headways (gaps between consecutive vehicles) on all lines are shorter than 30 minutes, ensuring that a transportation service is always available within each time slice. Further details on solving the equations for delay estimation were done by Krishnakumari et al. (2019). Passenger data was allocated to the corresponding time slice based on their check-in time. It is important to note that passengers may cross into the next time slice during their journey, introducing dependencies between information in adjacent time slices. These dependencies should be taken into consideration when analyzing or utilizing this data (Krishnakumari et al., 2019).

3.2 Delay prediction model

3.2.1 GNN prediction model

The directed graph $G = \big(S, E, X(t)\big)$ representation mentioned in section 3.1 allows for input the metro network information into the graph neural networks (GNNs). The delay prediction on the metro network can be formulated as a time-series regression task, in which observed the delays on E, during the previous T_{past} time step, are used to predict the most likely delay at $t + T_{past}$ time step. Based on the definition, the regression task can be represented as Equation 3.

$$v_t = \, argmax_{v_t} \, log \, P\left(v_t \mid v_{t-T_{past}}, \dots, v_l \right) \tag{3}$$

Where $v_t \in R^{N \times F}$ is the tensor of F delay features on N links of the metro network at time t. In this study, v_t is equal to $d^{n}_{S_o, S_d, t}$, which represents the sum of three component delay: the waiting delay d_t^{wait}, the on-board delay $d_t^{on-board}$ and the transfer delay d_t^{trans}. Thus, each delay can be regarded as a data attribute, leading to $F = 3$.

The aforementioned graph formulation considers the delays that have been mapped onto links in the metro network as edgewise features of G, via the delay estimated approach introduced in section 3.1. At the same time, no node feature of G is involved. Thus, the nodes and edges of G can be inverted to the link graph $L_G = (V, E, X)$ enable the use of architectures with only nodewise features. Where V is the node in L_G which represents track links in the metro network, E is the edge in L_G which represent station in the metro network and without attributes, X is the node attributes set of L_G, which storage estimated historical delay data. This link graph then has an adjacency matrix A_L, defined as a binary variable. A_L equals t if track i and j are connected by a station, and equal to 0 and vice versa. Based on the link graph formulation, the historical delay data could be put in the ST-GCN model to predict node attributes. The L2 loss function is used to train the model, which is defined as Equation 4.

$$L(\hat{y}; \theta) = \sum_{t} \left\| f\left(v_{t-N_{\text{past}}}, \dots, v_t; \theta \right) - v_{t+N_{\text{future}}} \right\|^2 \tag{4}$$

Where θ denotes the trainable parameters, $v_{t+N_{\text{future}}}$ denotes the ground truth delay value, and $f(\cdot)$ denotes the model prediction based on the historical data in past $periods$ period.

Specifically, the spatial-temporal graph convolutional network (STGCN) realized the spatial and temporal convolution via the structure indicated in Figure 3. The overall model includes 2 spatial-temporal blocks and an output layer. STGCN is first purposed by Yu et al., (2018).

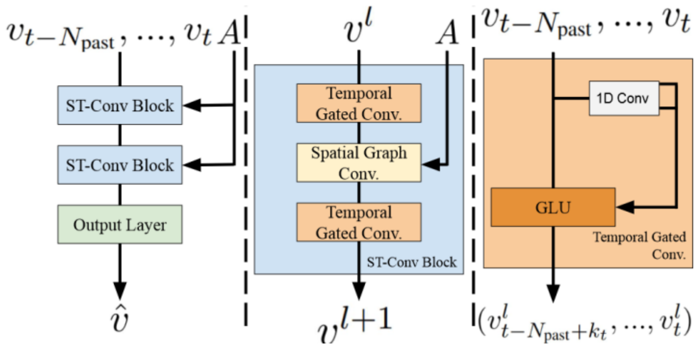


Figure 3 The STGCN model architecture developed by Yu et al., (2018). From left to right are the overall model structure, ST-Conv block, and Temporal Gated Conv block, respectively.

For spatial convolution blocks, it captures the spatial relationships in the data by utilizing graph convolutions. Spectral graph theory offers a method, known as the graph Fourier transform, to generalize the convolution operation for graph-structured data. The key aspect of this approach is the analysis of the eigenvalues of the normalized graph Laplacian matrix, which is defined as follows:

$$L = I_N - D^{-1/2} A_L D^{-1/2} \tag{5}$$

Where I_N denotes the identity matrix; D is the diagonal degree matrix of the adjacency matrix A_L such that $D_{ii} = \sum_j A_{ij}$. The standard convolution for regular grids is not applicable to general graphs. In GNN based model, the essence is the graph convolution operator "$*_G$", which is based on the conception of spectral graph convolution, to generalize CNNs to structured data forms. The graph convolution" $*_G$" is defined as the multiplication of the graph signal x with kernel Θ, as Equation 6 denotes.

$$\Theta *_G x = \Theta(L)x = \Theta\left(U \Lambda U^T\right)x = U\Theta\left(\Lambda\right)U^T x \tag{6}$$

Where the graph Fourier basis $U \in R^{(N \times N)}$ is the matrix of eigenvectors of the normalized graph Laplacian, $\Lambda \in R^{(N \times N)}$ is the diagonal matrix of eigenvalues of L, and kernel $\Theta\left(\Lambda\right)$ is a diagonal matrix. All the convolution operations are on the link graph L_G.

The computation of Θ involves $O(n^2)$ operations, which limits its practical use for large-scale graphs. To address this issue, an approximation method was introduced by Hammond et al. (2011). that restricts the graph kernel Θ to the set of Chebyshev Polynomials. Additionally, (Kipf and Welling, 2017) introduced a first-order approximation for the graph kernel. Both of these approximations have been incorporated into the STGCN architecture, after being generalized for use with multidimensional tensors. For brevity, the details of these approximations and the generalization of graph convolution have not been included in this paper, but can be found in Defferrard et al., (2017).

The temporal convolution block, it aims to capture temporal relationships in the data. Recurrent Neural Networks (RNNs) are often used for this purpose; however, these networks can be difficult to train due to the" vanishing gradient" problem. Additionally, Oord et al., (2016) have shown that a 1D convolution along the temporal dimension of data can be more effective than an RNN on shorter sequences, while at the same time being quicker to train. As shown in Figure 2 (right part), the temporal convolutional layer of each ST-Conv block contains a 1D causal convolution with a kernel of size kt and a gated-linear unit (GLU) nonlinear activation. Like the gating present in RNN models, namely LSTM and GRU, the nonlinear activation provides a gating that determines the importance of past inputs on future predictions. The resulting temporal convolution is defined in Equation 8.

$$\Gamma *_T Y = P \odot \sigma(Q) \tag{8}$$

Where the input is split along the "channels" dimension to obtain P and Q. More information about temporal convolution and its generalization to 3D tensors can be found in (Yu et al., 2018).

Accordingly, the output of block l is as Equation 9 defined.

$$v^{l+1} = \Gamma_1^l *_T ReLU\left(\Theta^l *_G \left(\Gamma_0^l *_T v^l\right)\right) \tag{9}$$

Where Θ^l is the spectral kernel of the graph convolution, and Γ_0^l , Γ_1^l are the temporal kernels within block l. $ReLU$ denotes a rectified linear unit activation. After stacking two ST-Conv blocks, we attach an extra temporal convolution layer with a fully connected layer as the output layer in the end, as the left of Figure 3 indicates. The temporal convolution layer maps outputs of the last ST-Conv block to a single-step prediction.

In summary, the STGCN model is a universal framework that can be used for structured time series analysis, including transport network modeling and spatial-temporal sequence learning(Yu et al., 2018). Its spatial-temporal block combines graph convolutions and gated temporal convolutions to extract useful spatial and essential temporal features. The model is entirely composed of convolutional structures, making it faster and more efficient in handling large-scale networks with fewer parameters.

3.2.2 Benchmark prediction model

To ensure that the STGCN model gives a relatively more accurate forecasting explanation than other simpler or self-explainable models, the performance of STGCN should be compared with the benchmark models. Different from the GNN model that considers network topologic information, the benchmark model should make predictions only based on historical delay data as an input feature. Two benchmark models particularly well-suited for tabular data prediction are selected for assessing the prediction performance of the STGCN model, which are XGBoost and LSTM regression.

XGBoost is a popular machine learning algorithm that has gained significant attention in recent years due to its high accuracy and efficiency in handling large-scale datasets. XGBoost stands for "Extreme Gradient Boosting," and it is an ensemble learning method that combines multiple decision trees to improve prediction performance. XGBoost is a boosting algorithm, which means that it trains a sequence of models that build upon the weaknesses of the previous models. In XGBoost, each model is a decision tree, and each tree is built using a gradient boosting technique that minimizes the loss function of the previous tree. This process continues until a certain stopping criterion is met, such as reaching a specified number of trees or achieving a certain level of accuracy.

One of the key advantages of XGBoost is its scalability, which allows it to handle high-dimensional data with sparse features. This is accomplished through a regularization term in the objective function, which penalizes large model coefficients and helps prevent overfitting. Additionally, XGBoost supports multiple loss functions, including regression, classification, and ranking, making it versatile for a wide range of applications. Another advantage of XGBoost is its explainability. The algorithm provides feature important scores, which indicate the contribution of each input feature to the model's overall prediction. This can be useful for understanding the underlying factors that influence the outcome and can help with feature selection and interpretation.

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) that has been widely used in sequence modeling tasks due to its ability to effectively capture long-term dependencies in sequential data. The key difference between LSTMs and traditional RNNs is the addition of memory cells, which allow LSTMs to selectively forget or retain information over time. This makes LSTMs particularly effective in modeling sequences with long-term dependencies or sequences with multiple layers of temporal dependencies.

One of the key advantages of LSTMs is their interpretability, which allows for a better understanding and analysis of the model's predictions. This is because LSTMs can visualize the activations and memory states of each cell, making it easier to identify the specific features or patterns that the model is using to make predictions. LSTMs are often chosen as benchmark models in sequence modeling tasks due to their strong performance and interpretability. They have been shown to outperform traditional RNNs and other sequence modeling approaches in many applications, and their interpretability makes them a valuable tool for understanding the underlying patterns and features in sequential data.

3.3 GNN post-hoc explanation approach

To extract information on how the model makes predictions based on the input time-series data, it is essential to acquire the explanatory variables that unveil the significance of each input data point’s contribution to the prediction outcome. This process enables us to identify the most influential nodes in the input graph and explore explanations for the GNN model's specific predictions. By doing so, we gain valuable insights into the model's decision-making process and understand the dependencies among time-series data in the graph.

Post-hoc explanations are techniques used after a model has made predictions, enabling us to gain insights into its internal workings without modifying the model itself. The primary benefit of using a post-hoc approach is that it provides model explanations that are model-agnostic. In other words, these explanations are not tied to a specific model architecture and can be applied to different types of models. This flexibility is especially valuable when working with complex models like GNNs, as it allows us to interpret their predictions without the need for intricate model-specific interpretation techniques.

The Graph Neural Network (GNN) explainer(Ying et al., 2019) is a post-hoc explanation approach for GNNs, which aims to provide insights into how GNNs make predictions on graph-structured data. The approach is based on a novel graph perturbation technique that systematically removes nodes from the graph and measures the resulting change in the GNN's output. This process allows the GNN explainer to identify the most influential nodes in the input graph and provide explanations for the GNN's predictions.

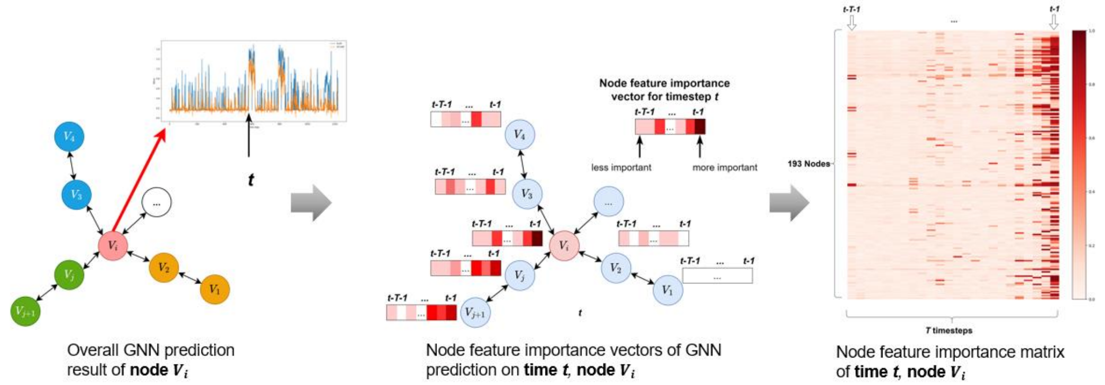
Same as the previous section, the link graph $L_G = (V, E, X)$ is the spatial-temporal graph which represents the metro network with delay data as node attributes, and act as input of STGCN model. For the sake of simplicity and aligning with the mainstream of research, in the following statements, G is used to represent the link graph L_G, and the input of STGCN can be represented as $G = (V, E, X)$.

The primary objective for GNN Explainer is to generate a minimal graph that explains the decision for a node or a graph. To achieve this goal, the problem can be defined as finding a subgraph in the computation graph, that minimizes the difference in the prediction scores using the whole computation graph and the minimal graph. Specifically, for each node v, v's computation graph $G_c(v)$ is defined as the neighborhood-based aggregation that fully determines all the information the GNN used to make the prediction $\hat{y}$ at node v. For $G_c(v)$, there is associated binary adjacency matrix $A_c(v) \in 0,1^{n\times n}$, and associated feature set $X_c(v) = x_j | v_j \in G_c(v)$. Thus, prediction result of GNN is $\hat{y} = f\big(G_c(v), X_c(v)\big)$. Formally, GNN Explainer explain the model prediction $\hat{y}$ as $(G_S,\ X_S^F)$, where G_S is the minimal graph of the computational graph $G_c(v)$, and X_S is the feature of G_S, and X_S^F a small subset of node features, masked out by the mask F that indicate the importance of node features in X_S.

The objective of explaining a graph is to identify the critical features or structures that influence the decisions of a neural network. In detail, the notion of feature importance can be denoted by mutual information MI, and the explanation process can be formulated as following optimization framework:

$$\max_{G_S} M\,I\big(Y, (G_S, X_S)\big) = H(Y) - H(Y \mid G = G_S, X = X_S)$$

A feature is considered important if its removal or replacement significantly changes the prediction outcome. Conversely, if removing or altering a feature does not affect the prediction, the feature is deemed non-essential and should not be included in the explanation of the graph. For node v, MI quantifies the change in the probability of prediction $\hat{y} = f\big(G_c(v), X_c(v)\big)$, given the computation graph is limited to explanation subgraph G_S and corresponding node features X_S. For brevity, the details of the optimization framework have not been included in this paper, but can be found in (Ying et al., 2019).



Overall GNN prediction result of **node V_i**

Node feature importance vectors of GNN prediction on **time t**, **node V_i**

Node feature importance matrix of **time t**, **node V_i**

Figure 4 Mechanism of GNN Explainer explains predict result of a specific node V_i at a particular timestep t. The node feature importance matrix reflect how the historical estimated delay of all the node contributes to the delay prediction.

The proposed process is depicted in Figure 4 with an example of network prediction. Assuming that we are interested in how the model makes predictions of node V_i at timestep t, we implement the above-mentioned approach and acquire the node feature importance vector of each node, which reflect how the time-series historical data of different node in the past time window T contribute to the target node V_i at timestep t. The color blocks in the feature importance vectors shown in Figure 4 denote MI, which can reflect the importance of the corresponding data point of the historical data step (from *t-T-1* to *t-1*) to the prediction outcome. At the same time, data of some nodes may be over trivial (for example, node V_1 in white) and can be excluded from the minimal subgraph G_S. Finally, all the vectors of all the n nodes are piled along y-axis of the matrix mask F as the colored matrix in Figure 4, which reflect the node features importance of G_S. The x-axis of matrix mask F is the time scale of input data which covers T timesteps, and the y-axis indicates the number of nodes.

3.4 Pattern recognition and classification

For the prediction result of a specific node at a specific timestep, we calculated the importance values of input node feature X_S as explanatory variables using the post-hoc explanation method purposed in section 3.3. The distribution pattern of the input feature importance values may vary across different node predictions, resulting in a feature importance matrix image with diverse patterns.

To address this variability and classify the correlation patterns into distinct types with discernible characteristics, we introduce an image pattern recognition technique to classify images with similar patterns. By leveraging this method, we aim to classify the feature importance distribution patterns into elementary representative types that possess unique attributes. The obtained clustering results offer a comprehensive overview of the diverse correlation patterns present in different delay predict outcomes.

By gaining insights into these feature importance distribution patterns, researchers can develop a deeper understanding of the relationships between input features and prediction outcomes. The clustering results serve as valuable prior knowledge for further exploration and analysis of these correlations. This knowledge can inform and guide subsequent investigations, allowing for a more informed exploration of the correlation patterns within the context of delay prediction for nodes.

For many image clustering or classification problems, replacing raw image data with features extracted by a pre-trained convolutional neural network (CNN) leads to better clustering performance(Guérin and Boots, 2018). Previous research compared multiple neural network architectures and proved that the ResNet50 could perform relatively better than other prevailing architectures. Residual Network is a classic neural network used as a backbone for many computer vision tasks, which was first proposed by (He et al., 2016). ResNet-50 is a convolutional neural network with 50 layers. The pre-trained Resnet-50 deep neural network architecture could effectively recognize the features of the images and has been widely used in computer vision, including image classification and detection applications. The process of implementing the ResNet50 can be done by the PyTorch deep learning framework.

4. Case study

In this section we present a case study involving the Washington DC metro network to demonstrate the application of the proposed methodology outlined in the preceding sections. Section 4.1 describes the input data utilized for the case study. Section 4.2 describe how the data are preprocessed before being used in the model. Section 4.3 introduces the parameters setting of the prediction model and GNN Explainer in detail.

4.1 Data description

This study utilized a comprehensive dataset obtained from WMATA, encompassing various aspects of the metro network. The dataset included information on rail trips, providing details of journeys from the origin station to the destination station, as well as intermediate movements between adjacent stations. The data covered a period starting from August 19, 2017, to May 22, 2018, spanning a total of 277 days.

In addition to rail trip data, passenger journey information was also incorporated into the analysis. This data, derived from smart card records, offered insights into the origins and destinations of passengers. By combining the known origin and destination with the rail movements, it was possible to infer the specific route traveled by each passenger.

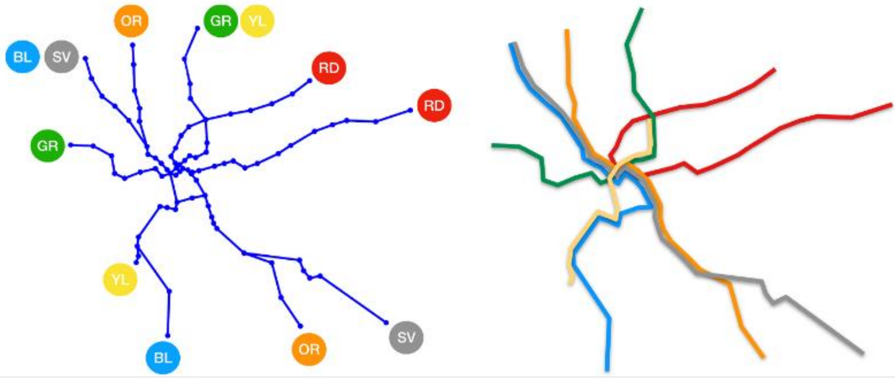


Figure 5 Washington metro network(Krishnakumari et al., 2020)

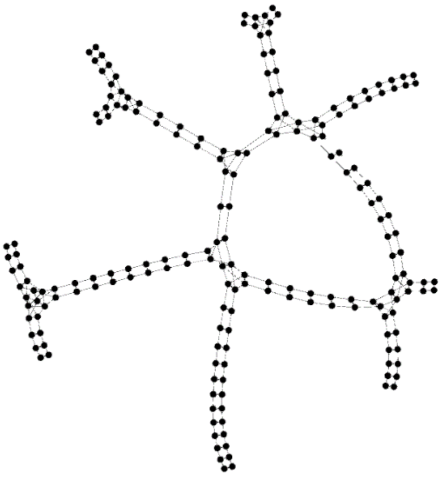


Figure 6 Link graph of Washington metro network.

Krishnakumari et al.(2020) pre-processed the dataset before its utilization in the current study. Specifically, they implemented the steps outlined in section 3.1, which involved transforming the disaggregated raw data of passenger and rail trajectories into a network-wide aggregation perspective. The estimated delay outcomes are mapped in each network link as mentioned in section 3.1. Accordingly, the dataset available for this research includes two parts: the topology information of 6 lines, 91 stations, 184 links and 9 transfer stations as Figure 5 indicates, and the estimated average delay experienced by per passenger experienced (estimated delay dataset in the below. The estimated delay data cover every stage of travel, including estimated waiting delay at the initial station ("waiting delay"), delay during the vehicle running on the track ("track delay"), delay at the transfer station ("transfer delay"), and their linear sum ("total delay"). These data are mapped in 193 nodes in the link graph, as Figure 6 denotes. The nodes in the Link graph Figure 6 represent the directional track in the metro network depicted in Figure 5

4.2 Data preprocessing

The occurrence of irregular delay occurrence patterns in transport systems can be attributed to various reasons, including national holidays, maintenance work, or check-in system failures. For instance, days with passenger counts that are ten times lower than average and without a discernible morning and evening peak pattern can be attributed to irregular events like holidays. Additionally, days with over half of the data showing zero passengers in the system can be discounted as special holidays, maintenance work, check-in system, or data processing failures. It is difficult to determine when the earlier data on such days is still relevant or may have been influenced, making it necessary to discount these days as well. Since a complete dataset spanning 277 days is available, the removal of data under irregular conditions still leaves a substantial amount of data for robust analysis. Consequently, the study disregards data from dates featuring any of the aforementioned irregular conditions, as illustrated in Table 1. This refinement results in a dataset of 260 days, which serves as the foundation for the study's analysis.

Table 1 The dates that were excluded and the corresponding reason.

Date(dd-mm-yyyy)	Reason
04-09-2017	Labor Day
23-10-2017	No afternoon data
23-11-2017	Thanksgiving
24-11-2017	Thanksgiving
08-12-2017	No afternoon data
13-12-2017	No morning data
25-12-2017	Christmas
01-01-2018	New Year's
15-01-2018	Martin Luther King Jr. day
19-02-2018	President's day
02-03-2018	Very little data
09-03-2018	No data
21-03-2018	Very little data
26-04-2018	No data
19-06-2018	No data
04-07-2018	Independence day
24-08-2018	No morning data

To train the delay prediction model, we separate the whole dataset of 260 days into a train set, validation set, and test set, with the ratio of 60%, 20%, and 20%. Correspondingly, the train set covers 156 days, the date from August 19th, 2017, to January 22, 2018; the validation set covers 52 days, beginning from January 23, 2018, to March 16, 2018; the test set covers 52 days, beginning from March 17, 2018, to May 22, 2018.

Furthermore, we find that the outliers in the dataset are extremely large which is unreasonable, these 0.1% extreme data are deleted from the dataset. The finally used datasets distributions are depicted in Figure 7. The x-axis refers to the delay value in seconds. The y-axis refers to the cumulative distribution function (CDF) of each dataset. The figure illustrates that more than nearly 80% of delays occur on individual network segments that last no more than 50 seconds.

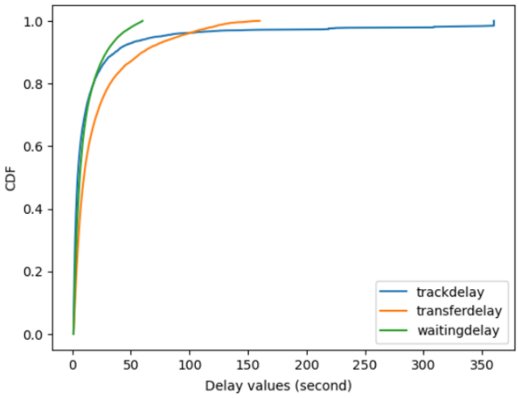


Figure 7 Data distribution of estimated delay dataset. The label denotes the distribution of track, transfer, and waiting delay distribution respectively.

4.3 Model parameters setting

4.3.1 Prediction model parameter

In the ST-GCN model, we use the historical estimated delay data of the preceding 24 time steps, which is the delay in the past 12 hours, to predict the delay of the next time step, which is the delay in the next half an hour. The number of features of input spatial-temporal time series delay data is 3, which includes the data of estimated waiting time delay, transfer delay and track delay. The output feature is one of the estimated waiting time delays, transfer delays, track delays, or overall delays on each link, which has 1 channel every run. For the STGCN model, the size of the spatial kernel $k_s = 5$ and the size of the temporal kernel$k_t = 5$. In this study, the graph Laplacian is approximated using Chebyshev polynomials. The ST-Conv block contains bottleneck-shaped channels, with Block 1 consisting of (1, 32, 64) channels, Block 2 consisting of (64, 32, 128) channels, and the Output Block consisting of (128, 1) channels. The model is trained for 100 epochs with a batch size of 32 using ADAMW optimizer, and L2 loss with the initial learning rate r=0.001 while implementing a learning rate decay r=0.1r every 10 epochs. As the baseline models for comparison, the XGBoost and LSTM regression are also implemented based on the same time-series dataset of estimated delay, without any additional feature engineering or topologic information. These two models use the same input and output time window as the ST-GCN model. In this way, the comparison may highlight the advantage of the STGCN model in extracting spatial-temporal features based on graph convolution.

4.3.2 GNN Explainer parameter

The explanatory variables of prediction results are obtained using the Python library package PyTorch Geometric (PyG) library, which is a powerful tool built upon PyTorch for developing and training Graph Neural Networks (GNNs) to handle a wide range of applications dealing with structured data. PyG consists of various methods for deep learning on graphs and other irregular structures, also known as geometric deep learning. Specifically, in our approach, we leverage the **torch_geometric.explain** module, which provides a set of essential tools to interpret the predictions of a PyG model and comprehend the underlying phenomenon of a dataset (Fey and Lenssen, 2019).

To achieve this, we employ an explainer with the following parameters:

- **model**: The PyG model we want to explain the predictions for.
- **algorithm**: The explainer algorithm we use to interpret the model's behavior. In this case, we use **GNNExplainer**, which undergoes 200 epochs during the explanation process. The principle of GNNExplainer is introduced in section 3.3
- **explanation_type**: We set this to 'phenomenon' to focus on explaining the underlying phenomenon of the dataset, although the alternative option is 'model'.
- **node_mask_type**: The type of mask used for nodes during explanation. We set it to 'attributes', indicating that we use attribute-based masks. Alternatively, we could use

'object'.

- **edge_mask_type**: The type of mask used for edges during explanation. In this case, we set it to 'object', which highlights the importance of edges in the explanation. Alternatively, we can set it to 'None' to exclude edges from the explanation.
- **model_config**: A dictionary containing various configuration settings related to our PyG model. We specify that our model is used for 'regression' tasks, and we focus our explanations at the 'node' level. The 'return_type' is set to 'raw', indicating that our model returns raw values rather than log probabilities.

By configuring the explainer with these parameters, we gain valuable insights into the inner workings of our PyG model and understand how it arrives at its predictions. This level of interpretability is crucial for building trust and understanding in complex machine learning models, especially in applications involving structured data.

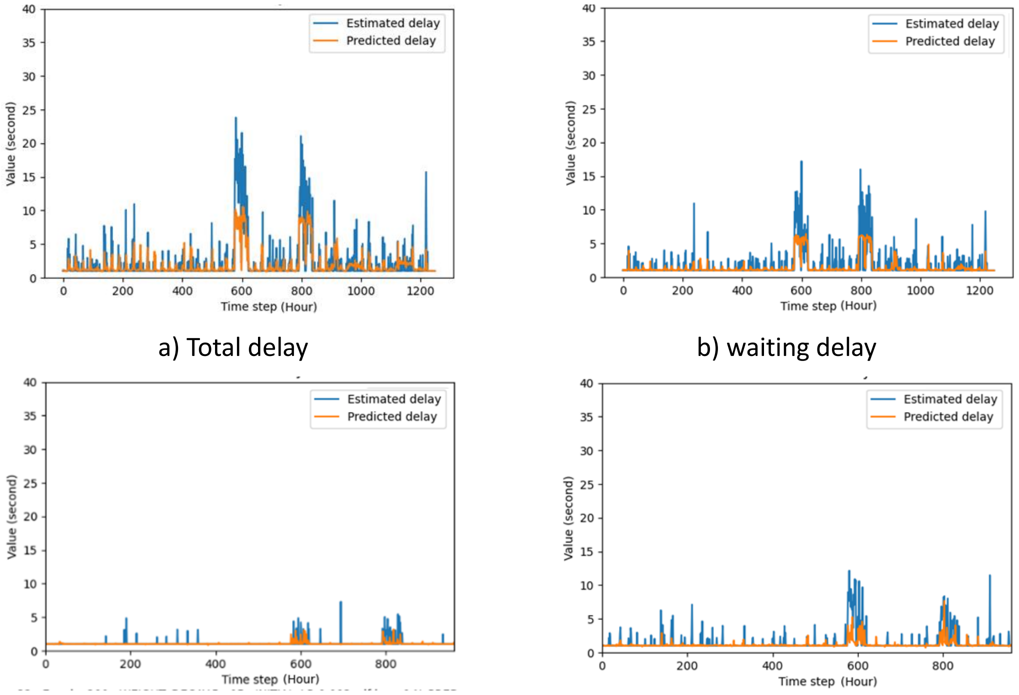
.

5. Results

This chapter presents the findings of the case study, along with the corresponding analysis. Section 5.1 includes an overview of the delay prediction results for the Washington DC metro network. Section 5.2 focuses on the explanatory variables derived from the model prediction using the post-hoc approach. This section includes a detailed analysis of the spatial and temporal feature importance distribution patterns of input historical data, as well as the spatial-temporal correlations among nodes in the metro network's link graph. The combination of these analyses allows us to gain a deeper understanding of the dynamics and dependencies within the metro network, shedding light on the factors influencing delay occurrences and providing valuable insights for network management and optimization.

5.1 Delay prediction

As mentioned in the previous chapter, the output of the ST-GCN model consists of the delay prediction for 193 links within the metro network at intervals of 30 minutes. The dataset encompasses data for three distinct attributes: waiting delay, transfer delay, and track delay. Consequently, from the ST-GCN model, we obtain the respective delay predictions for these three attributes, as well as the total delay prediction. The total delay prediction is derived by summing the individual predictions for waiting for delays, transfer delays, and track delays for all 193 links within the metro network. Figure 8 illustrates the result of the delay prediction of a sample node in the graph, which corresponds to the network segment starting from station *Metro Center* to *Gallery Pl-Chinatown*.



a) Total delay

b) waiting delay

c) track delay

d) transfer delay

Figure 8 Delay prediction result of a sample node (network segment from *Metro Center* to *Gallery Pl-Chinatown*) in the link graph. The label indicates the estimated and predicted delay.

Based on the prediction results, we can observe that the model effectively captures the major trends in delay occurrences, especially for longer-lasting delays. However, as the estimated delay increases, there is a tendency for larger discrepancies between the estimated delays and the model's predictions. This limitation becomes evident during highly severe or prolonged delay scenarios, where the model struggles to precisely estimate the extent of delays.

For instance, on the 12th day of the test set (timestep range from 576 to 624), there were significant delays in the metro network. The model correctly predicts the occurrence of serious delays during this period, but the exact extent of the delays is not accurately forecasted. This highlights the need for further enhancements to improve the model's performance in such extreme delay situations.

Besides, for most cases, both the estimated delay and predicted delay during waiting, on the track, and transferring delays fall below 10 seconds. Analyzing the values depicted in the four subfigures of Figure 8, we can ascertain that the total delay is predominantly influenced by the waiting time delay in a passenger's travel trajectory, followed by the transfer delay. On the other hand, the track delay appears to occur less frequently and has a lesser impact compared to other types of delays. Thus, for simplicity, we will analyze the prediction result of total delay data with more significant and stable delay phenomenon prediction.

To evaluate the model's performance, we compared it with baseline models XGBoost and LSTM using MAE and RMSE metrics. The results in Table 2 demonstrate that the ST-GCN model outperforms the baseline models in terms of prediction accuracy. This indicates that the ST-GCN model provides more accurate delay predictions compared to XGBoost and LSTM. A relatively accurate prediction model implies that the prior knowledge model learned from the real-world data is more reliable and accurate, which enhances our ability to provide compelling explanations for delay forecasts.

Table 2 Performance of different models on different attribute data

Model	Total delay		Waiting delay		Track delay		Transfer delay	
Metrics	MAE (s)	RMSE(s)	MAE(s)	RMSE(s)	MAE(s)	RMSE(s)	MAE(s)	RMSE(s)
ST-GCN	**0.388**	**0.651**	**0.281**	**0.504**	**0.118**	**0.265**	**0.206**	**0.458**
XGBoost	1.560	4.547	0.890	2.919	0.334	2.087	0.595	2.629
LSTM	1.301	4.496	0.781	2.879	0.394	1.225	0.540	2.619

Consequently, we implemented post-hoc explanation techniques based on the prediction results of the ST-GCN network, enabling us to gain further insights and interpretability into the key data points influencing the delay predictions.

5.2 Explanatory variable of the prediction result

In this section, we present the explanatory variables of the black-box ST-GCN model, which were obtained through a post-hoc explainability approach. These explanatory variables represent the importance of historical data, specifically the spatial-temporal features of the input historical delay data. These variables interiorly contain the information on delay severity, so extra weighting on these variables by the value of delay occurrence is not required. By examining these explanatory variables, we can explore how the spatial-temporal feature of historical delay data contributes to the model prediction and uncover the spatial-temporal correlations among the historical delay values of the links.

As described in Section 3.3, we can determine the importance of each input data to the prediction result for every node and time step. In our case study, the spatial-temporal graph consists of 193 nodes, and the historical data is collected over a time window of 24 timesteps. Consequently, the size of each feature importance matrix is 193 * 24, reflecting the relevance of each feature to the prediction at each time step. For each time step, we can generate 193 feature importance matrices that belong to corresponding 193 nodes. Furthermore, for a single day, we can obtain a total of 193 * 48 feature importance matrices.

Figure 9 is the image of a sample feature importance matrix, corresponding to the prediction result of node 0 at timestep *t*. The colors of each cell in the matrix from deep to light depict the relative importance score of historical data to the prediction result. The importance scores are linearly normalized to span 0 to 1 across the matrix, ensuring a consistent scale for comparison. The x-axis is the time scale of input data which range covers preceding *T* timesteps. The historical data of 193 nodes are piled up along the y-axis. The arrangement of these colored pixels reflects the importance distribution of input historical data to the prediction result. The pattern of the image could provide us the potential insight into the spatial-temporal feature importance of estimated historical delay data.

To facilitate interpretation, auxiliary lines are incorporated into the figure. The blue and green auxiliary lines indicate their significant spatial and temporal features of input historical data respectively. The blue dash line highlights when certain historical data points play a pivotal role in shaping the present prediction. Similarly, the green dash lines illuminate nodes that exhibit a lasting impact on the prediction outcome over a more extended duration. These insights on the spatial-temporal feature of input data indicate that the present delay prediction of node 0 exhibits relatively higher dependency on the key segments' delay status, and the particular historical period delay still has an influence on current delay prediction. Thus, in the following subsection, the result spatial-temporal feature is recognized and evaluated in depth based on the image shown in Figure 9.

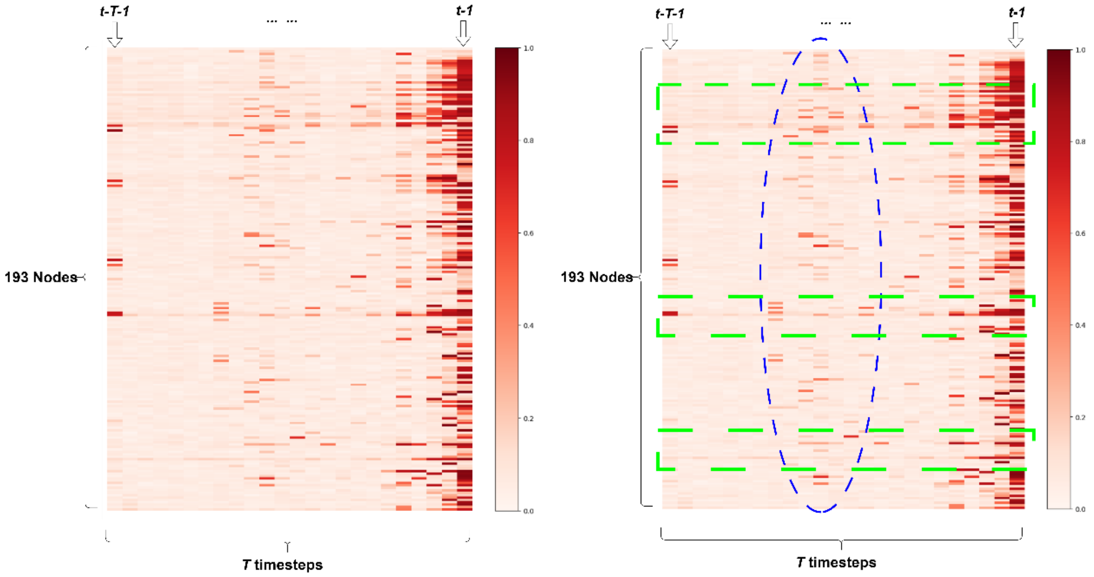


Figure 9 A sample of feature importance matrix image. The green and blue dash lines show the pixels that represent significant contributive data points.

However, it's important to note that the 2-dimensional form of the feature importance matrix solely reflects the temporal importance distribution of input data, disregarding the spatial importance distribution. Unlike temporal features, which can be easily organized sequentially along the horizontal axis, the current form of Figure 9 does not adequately convey the spatial distribution of the nodes. Thus, we will analyze the temporal and spatial feature importance separately in sections 5.2.1 and 5.2.2, with a similar methodology for pattern recognition but different forms of processing data. In section 5.2.3, based on the explanatory variables, we will explore the spatial-temporal delay correlation of the nodes in the link graph.

5.2.1 Pattern of temporal feature importance

The temporal feature importance is reflected by the distribution of importance values along the horizontal axis in the importance matrix. As there are prediction results of 193 nodes on 48 timesteps, and for each prediction result there is a corresponding feature importance matrix image reflecting different temporal feature patterns.

Thus, we implemented image pattern recognition as mentioned in section 3.4 to cluster the images with similar patterns to explore the representative patterns of temporal feature importance among prediction results. By implementing PCA (principal components analysis), the high-dimension feature of images is reduced to a limited number of principal dimensions, which aims to reduce the dimension of the image feature while mostly remain the image feature for clustering. Figure 10 shows the cumulative explained variance ratio by principal components. To mostly retain the feature of images, we choose 20 components of an image feature, which leads to more than 90% variance among the data covered.

The crucial step of clustering is to decide the number of clusters, or the value of K. Different K values for clustering are tested to calculate corresponding silhouette score and SSE (sum of squared errors). The result is shown in Figure 11. A clearer elbow point of the SSE curve in Figure 11a and a larger silhouette score in Figure 11b refer to a K value with better clustering performance. Thus, the best K value is 3. In Figure 12, the image clustering result is mapped into 3D space. The images with similar pattern are clustered and in the same color. Three axes reflect the value of the three most significant principal components (PC), which could mostly reflect the unitless relative distance among data points. The top 80% of data points in terms of their distance from the centroid are visualize.

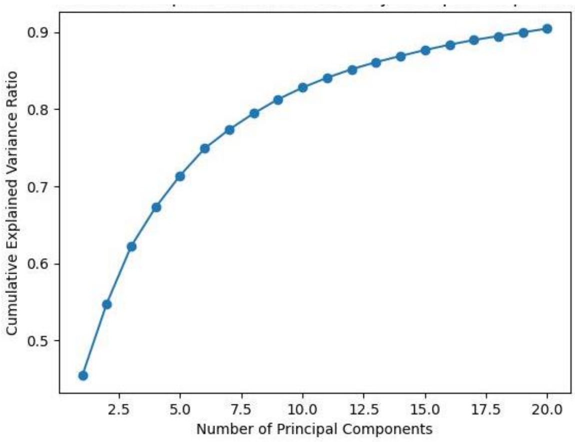


Figure 10 Cumulative explained variance ratio by principal components. After dimensional reduced to 20 principal components, more than 90% of variances among data are retained.

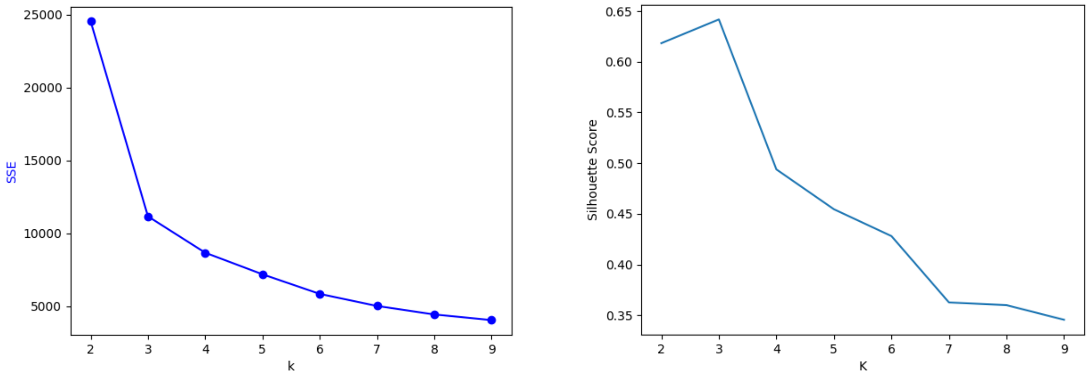


Figure 11 SSE and silhouette score of different K values. These figures show the optimum cluster number is 3.

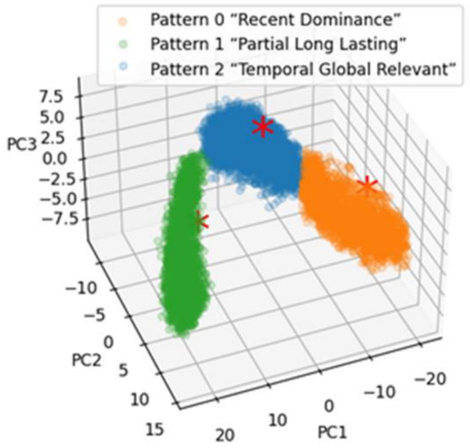
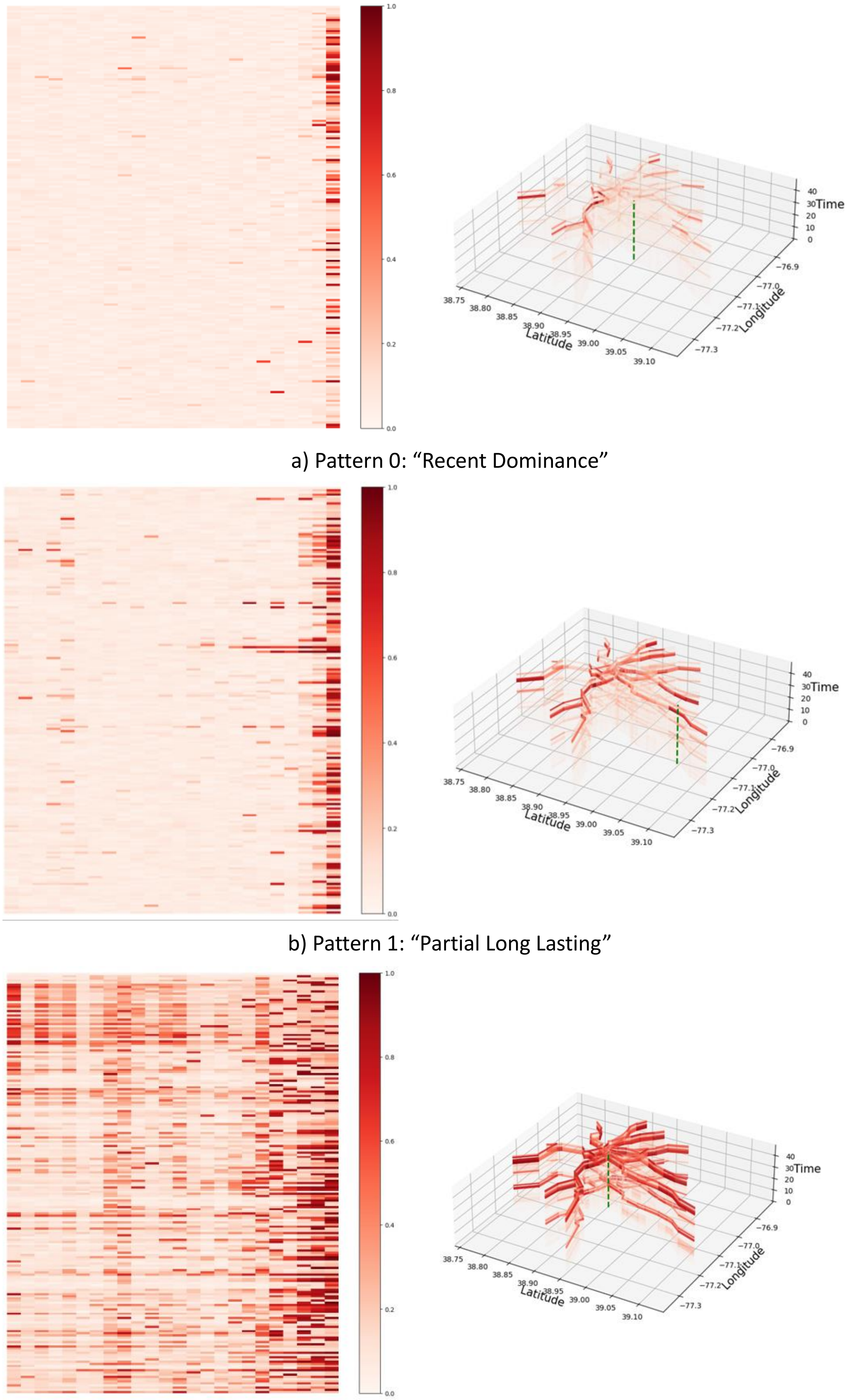


Figure 12 Clustering result of feature importance matrix. The red stars represent the centroid of each cluster.

The images within each cluster exhibit similarities in terms of their features, aligning with the corresponding temporal distribution patterns observed in the feature importance matrix. The proximity of a data point to the cluster centroid indicates the degree to which its image pattern represents the cluster as a whole. Consequently, for analysis and representation of the temporal feature importance pattern, we select the image that is closest to the centroid. In Figure 13, we showcase the most representative image for each cluster, accompanied by a 3D metro network mapping of the feature importance matrix values.

In Figure 13, each feature importance matrix corresponds to the prediction result of a specific node at a particular timestep. The color gradient, ranging from light to deep red, represents the linearly normalized importance value assigned to the input historical data. The latitude and longitude axes visualize the topological distribution of the network, while the time axis reflects the sequential arrangement of the input data's time steps. Vertically higher positions within the network image indicate data points that are closer in time to the predicted timestep. The green dash lines in each image denote the node that the prediction belongs to. To describe these patterns, we have assigned names to each of them based on their major features.

Notably, a remarkable observation emerges from Figure 13: each cluster exhibits a distinctive and significant pattern in terms of the distribution of temporal feature importance. The variations in color and positioning within the network image highlight the unique characteristics and specific contributions of different nodes and timesteps within each cluster. This analysis provides valuable insights into the temporal dynamics of input data contribution. To analyze the temporal feature importance, we only focus on the importance distribution along the time axis in Figure 13.



a) Pattern 0: “Recent Dominance”

b) Pattern 1: “Partial Long Lasting”

c) Pattern 2: “Temporal Global Relevant”

Figure 13 Identified three patterns according to temporal feature importance clustering outcomes.

In Figure 13 a) the pattern “Recent Dominance” is depicted. The midpoint image illustrates a unique trend where the most recent historical data to the prediction time holds a considerable diminish over the predicted outcome. Impactful influence is observed in the historical data of only a few nodes, with this influence dissipating within a span of two timesteps and rapidly diminishing thereafter. Besides, a large proportion of nodes show no contribution to the prediction. This pattern signifies that, within this specific context, primarily the latest recent delays of specific nodes, usually in the past 30 minutes, substantially affect the present delay prediction of the select node.

In Figure 13 b), the pattern “Recent Dominance” is depicted. In the context of this pattern, the historical delay data associated with certain nodes have a longer-lasting impact on the prediction outcomes. To these certain nodes, delays in the past 1 or more hours still appear correlation to the current delay. At the same time, other nodes show the impact that extends no more than a single timestep, which is less than 30 minutes. This implies that the delay on specific key nodes in the metro network graph shows a longer-lasting contribution to the predicted result, than other nodes. As time moves backward, the effect diminishes eventually, and the significant continuous effect usually not surpass nearly 4 hours.

For Figure 13 c), a notable shift is observed where a larger number of inputs datapoints make contributions to the prediction result with similar significance. While the diminishing temporal impact trend seen in pattern 1 and 2 persists but is less obvious. To some of the nodes, their historical delay shows equivalent impacts spanning many hours. This phenomenon reveals that the present delay demonstrates balanced dependence on historical delay within the past 24 hours across key nodes. As a result, cluster 2 reflects the pattern which can be named “Temporal Global Relevant”.

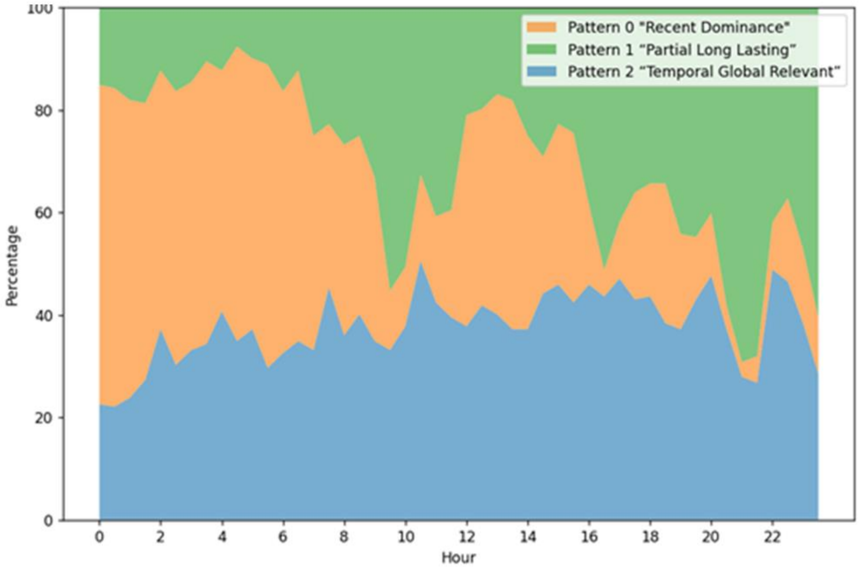


Figure 14 Temporal distribution of three feature patterns

Furthermore, we delve into the temporal distribution of these three patterns over 24 hours, which suggest a potential interplay between the different patterns over time. Figure 14 depicts the proportion of feature importance matrices corresponding to each pattern. The horizontal axis corresponds to the hours of the day, while the vertical axis reflects the percentage of each pattern at various times. This figure effectively highlights the temporal distribution disparities among the distinct patterns within a single day.

Figure 14 shows that the “Recent Dominance” pattern tends to occur more frequently before the morning peak. The count of this pattern decreases significantly during the morning and evening peaks. In contrast, pattern 1, referred to as "Partial Long Lasting," demonstrates an opposite and complementary trend. The count of pattern 1 reaches its peak during the morning and evening peaks, while the lowest point coincides with the peak of pattern 0. This observation indicates a complementary relationship between the two patterns, where one pattern's occurrence tends to be at its highest when the other pattern is at its lowest. Meanwhile, the "Temporal Global Impact" pattern (pattern 2) appears a more evenly distributed occurrence, with greater frequency during daylight hours. This pattern emerges when patterns 0 and 1 are spontaneously less prominent.

Thus, through an investigation into the unique temporal dependency patterns found in the historical data of delay predictions, we discern a common trend in how temporal features influence delay predictions throughout the day. We uncover three distinct patterns of temporal feature significance, primarily differing in the duration and spread of delay impact over time. And considering the distribution of these three patterns across the day, we can infer that delays during morning and evening peaks are often linked to historical delays at specific critical nodes over an extended period.

5.2.2 Pattern of spatial feature importance

To uncover the distribution of spatial feature importance, we condense the feature importance matrix along the temporal axis and project the aggregated importance values onto the metro network, as illustrated in Figure 15. This network significance graph pertains to the specific target node *Vi*. The color gradient assigned to each network segment corresponds to the cumulative of its importance score over the past *T* time, which indicates the time aggregate contribution importance of the segment's delay on the present prediction outcome of the target node *Vi.* The sum of each segment's importance score is linearly normalized across the entire graph, ranging from 0 to 1.

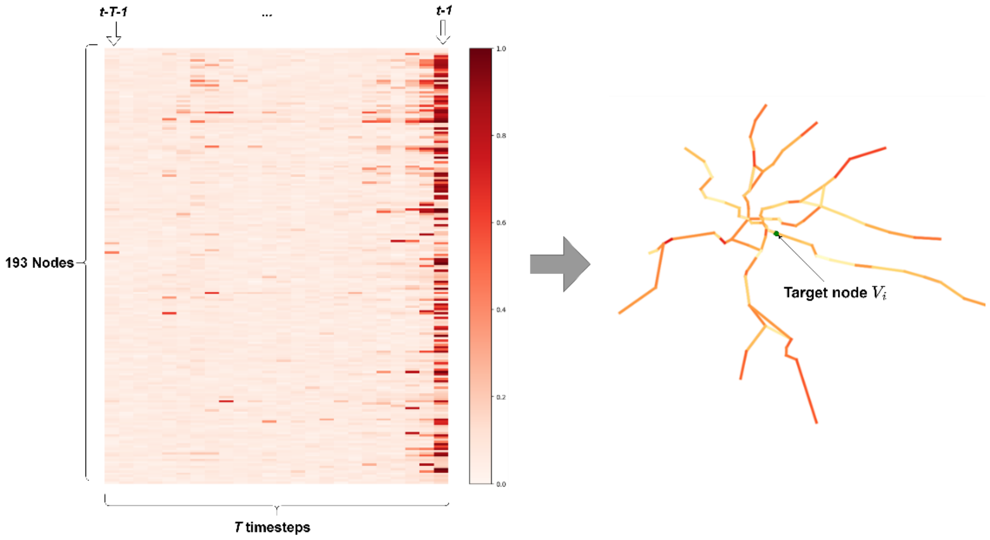
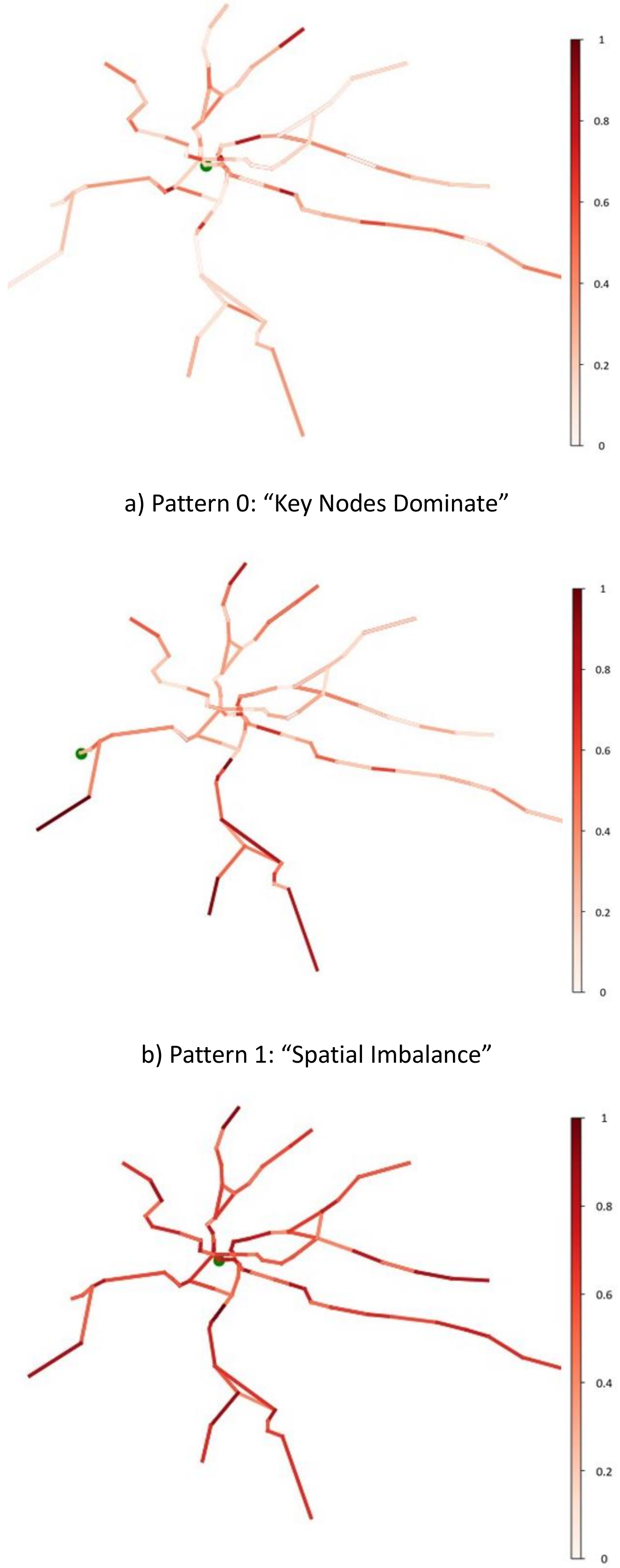


Figure 15 Compress the feature importance matrix for analyzing spatial feature importance of input data.

Thus, all the feature importance matrices are transferred to the form of 2D mapping. A similar approach as used in 5.2.1 for clustering is implemented based on the dataset of spatial importance graph depicted in Figure 15. Although each feature image indicates the feature importance of different prediction results, the pattern of the images reflects the relationship on how the historical data make a significant contribution to the result.

The most representative network images of each cluster are shown in Figure 16, showcasing the distinct identified patterns in the spatial feature importance distribution. The selected image is the one closest to the centroid within the corresponding cluster. The color gradient in the figures represents the contribution value of each link in the metro network, with contribution values being globally linearly normalized. The green dot on each image indicates the station for which the spatial feature distribution is associated to delay prediction. Each pattern reflects a unique distribution of input data contributions across different locations within the network. To aid in describing these patterns, we have assigned names to each of them based on their major features. By analyzing and comparing these different patterns observed in clusters 0, 1, and 2, we gain a deeper understanding of the spatial characteristics and the significance of the distribution among nodes within the metro network.



a) Pattern 0: “Key Nodes Dominate”

b) Pattern 1: “Spatial Imbalance”

c) Pattern 2: “Spatial global relevant”

Figure 16 Identified three patterns according to spatial feature importance clustering outcomes.

In Figure 16 a), the pattern of "Key Nodes Dominate" is depicted. In the context of this pattern, only a few key nodes make a significant contribution to the prediction results. Interestingly, these key nodes with significant contributions are evenly distributed across different lines and branches of the metro network. It is important to note that these key nodes tend to appear in proximity to the terminal stations, and near the transfer stations.

Figure 16 b) illustrates the pattern of "Spatial Imbalance." In this pattern, the segments with significant contributions are unevenly distributed among the lines and branches of the network. They tend to be more concentrated on specific lines or branches, leading to an imbalance in their spatial distribution. However, similar to the pattern "Key Nodes Dominate" depicted in Figure 16 a), key nodes near the terminal and transfer stations still appear relatively significant impact on delay prediction as appeared in pattern "Spatial Imbalance", although their contributions may exhibit an imbalance in extension within the cluster.

Figure 16 c) illustrates the pattern of "Spatial Global Relevant". In this pattern, most of the nodes show a higher contribution to the prediction result than other patterns appear. At the same time, certain key nodes at similar locations continue to play a crucial role in contributing to the predicted outcomes. However, the distinction between these significant nodes and other nodes is less pronounced compared to patterns "Key Nodes Dominate" and "Spatial Imbalance". This implies that the influence of individual nodes within the pattern "Spatial Global Relevant" is more evenly distributed, leading to a reduced gap in their impact on the prediction results.

Accordingly, by analyzing the spatial feature importance pattern, we identified three patterns with distinctive features, including "Key Nodes Dominate", "Spatial Imbalance" and "Spatial Global Relevant". However, a common regulation is revealed from these patterns: the key nodes with significant contributions to the delay prediction are usually located near the transfer station and the terminal station.

5.2.3 Spatial-temporal correlation of delay occurrence

In previous sections, we focused on feature importance matrices that depicted the contribution of historical data to the prediction results of individual edges within the metro network graph. This represents a many-to-one mapping, as it assesses the impact of historical data on a single node. However, to comprehensively explore the spatial-temporal correlation of delay occurrences, we extend our analysis to a many-to-many (or one-to-one) mapping. By considering how each node contributes to the prediction outcomes of every other node, we can effectively depict the spatial-temporal correlation of delay occurrences within the network.

To accomplish this, we aggregate all the feature importance matrices over a single day, resulting in a spatial-temporal correlation matrix. This matrix is shaped as 193 nodes by 193 nodes by 48 timesteps, encapsulating the spatial relationships and temporal dynamics of delay occurrences. Figure 17 presents the time-aggregated correlation matrix, where each value represents the sum of the correlation values at the corresponding position in the matrix. The color gradient ranging from light to deep represents the dependent extension from low to high. The red lines serve as auxiliary lines along the diagonal. The vertical and horizontal axes represent the source and target of dependencies, and the label on the axes indicate the index of nodes.

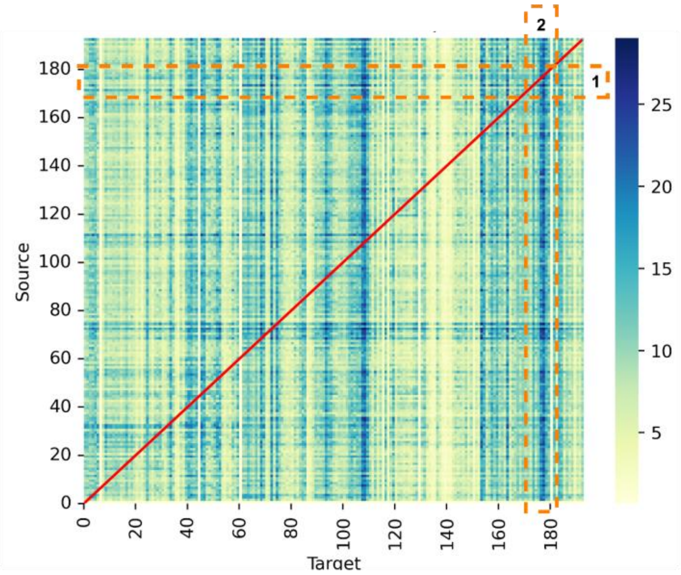


Figure 17 Time aggregated correlation matrix among nodes

It’s worth clarification that the distinction between the terms "feature importance value" and "contribution value." The former is utilized to describe the many-to-one mapping as previously mentioned, specifically referring to the significance of historical data in relation to a specific prediction outcome. The historical data distributed in time and space represent the feature of prediction input. To differentiate from this concept, we introduce the term "contribution value" in the following context to characterize how a particular node influences the delay prediction of other nodes, thereby reflecting the one-to-one relationship among all nodes in the graph.

Figure 17 provides a visual representation of the notable imbalance observed in the daily aggregated correlation values among nodes. For instance, the orange dashed frames labeled as 1 and 2 demonstrate that the node with an index close to 178 is more likely to be influenced by a majority of the nodes in the network (frame 2) while having less impact on other nodes (frame 1). This discrepancy highlights the varying degrees of influence and interconnections among nodes within the network, shedding light on the asymmetric relationships and potential dependencies that exist in the spatial-temporal correlation of delay occurrences.

To investigate the dynamics of correlation within a single day, we analyze the correlation matrix across different time steps. Figure 18 illustrates the distribution of linearly normalized contribution value in a whole day, presented through a historical plot and cumulative distribution function curve. To mitigate the influence of varying data quantities, the vertical axis of the histogram represents relative frequency or relative probability density, rather than absolute sample counts.

We visualize the correlation dynamics by mapping all the normalized contribution values of different times to the network graph. Figure 19 depicts the major significant correlation among nodes during the morning and evening peak hours on a sample day March 17th. Additionally, it is worth noting that more than 80 percent of the contribution values are lower than 0.6, as Figure 18 indicates. Given the complexity that would arise if all correlation values were directly depicted in the graph, we simplify the visualization and highlight stronger correlations by establishing a threshold of 0.6 for visualization. Only correlation values surpassing this threshold are depicted as directional edges in the graph.

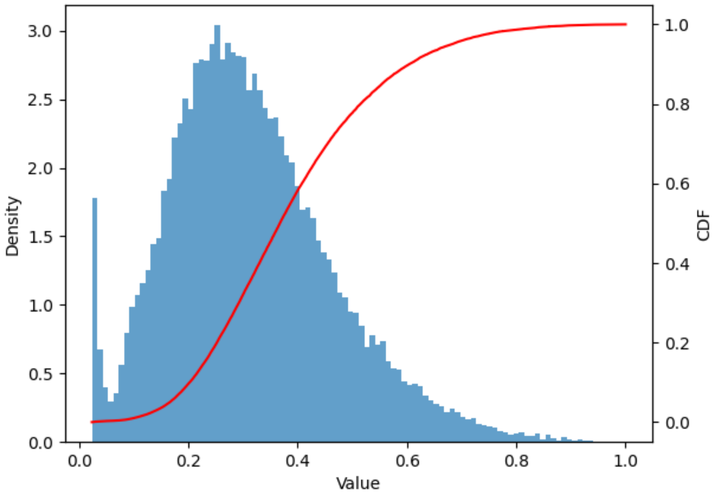
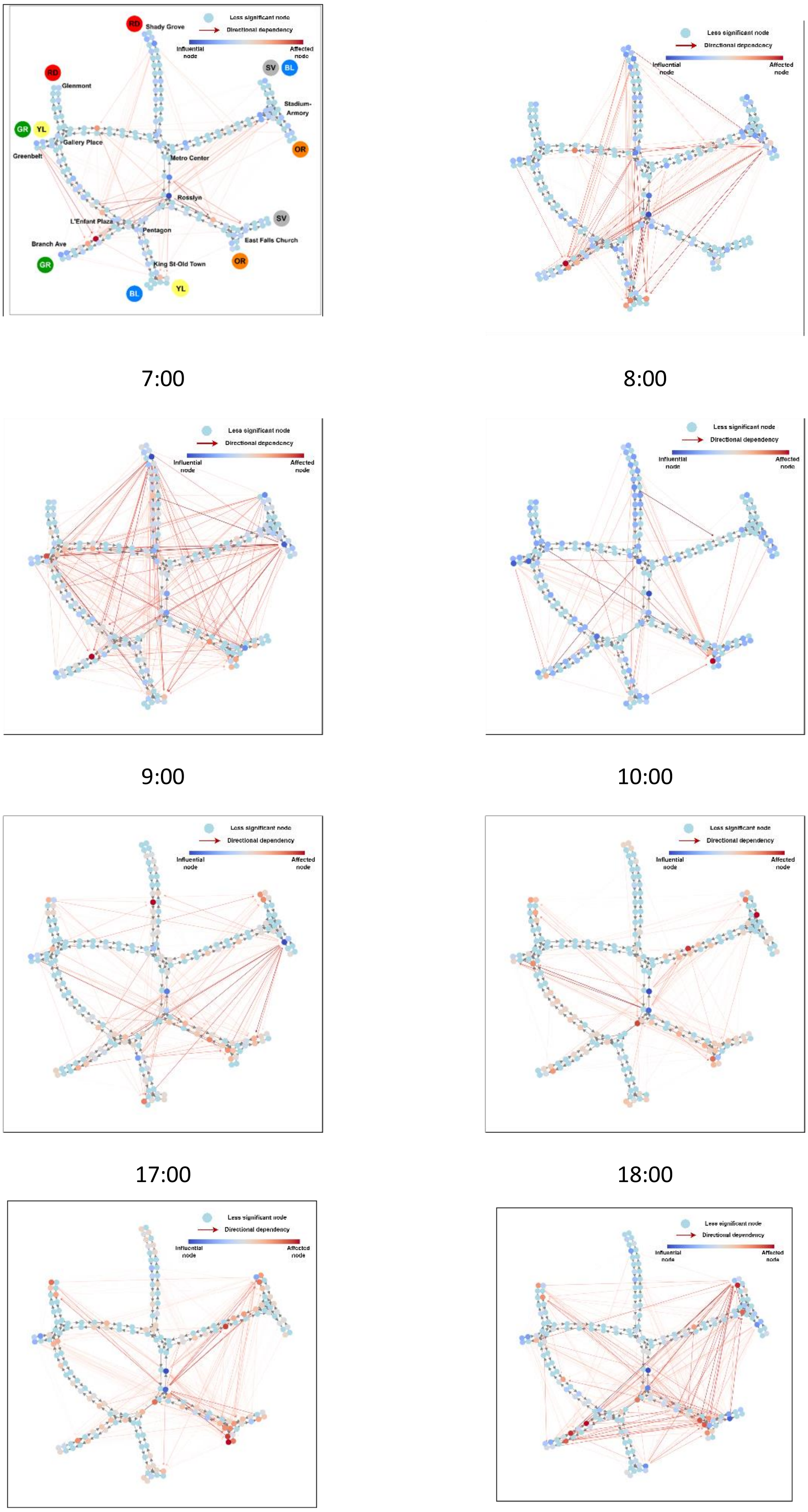


Figure 18 Distribution of normalized contribution value

In Figure 19, the label of lines, terminal stations and transfer stations are indicated based on the link graph previously mentioned in 4.1. Each node in the link graph represents the track segment of the metro network. The black directional lines represent the connection between nodes.

The directional edges represented in red symbolize the dependencies from the source node to the target node, surpassing the threshold value previously mentioned. These red directional edges indicate directional dependence between nodes, signifying the influence from source nodes to target nodes. In cases where a node is impacted by another, the dependency is designated as an inflow with a positive weight on the impacted node, and vice versa. The color gradient assigned to each node is determined by its node weight, reflecting its corresponding net flow. Nodes depicted in a red color gradient signify a positive net flow, and their node weight is positive. Conversely, nodes displayed in a blue color gradient indicate a negative net flow, with a negative node weight. Nodes without any red directional edges connected imply that they have no significant contribution to or are not impacted by other nodes, and they are depicted in light blue color.

Figure 19 provides insights into the flow dynamics and interdependencies among different nodes within the metro network. By analyzing the contribution values and net flows, we can identify clusters of nodes with significant impacts and observe the dynamics of nodes' delay interdependency.



7:00

8:00

9:00

10:00

17:00

18:00

19:00

20:00

Figure 19 The correlations among nodes during the morning peak hours

Figure 19 includes the subfigures of node correlations at different time periods (7:00-10:00, 17:00-20:00) during the morning peak hours. First, a notable feature is the varying amount of selected significant correlations. During the morning hours, the number of red directional edges first increases to its peak at 9:00 am and then gradually decreases. During the evening hours, the number of red directional edges first increases to its peak at 9:00 am and then gradually decreases. This phenomenon indicates that the strength of correlation among nodes may fluctuate significantly with time passing. Remarkably, this corresponds with the peak occurrence of the temporal feature pattern labeled as "Partial Long Lasting," as depicted in Figure 14 b). This suggests a plausible inference that the significant interdependencies often stem from the historical data of key nodes with long-lasting impacts on future delays.

Moreover, by examining the color distribution of nodes with red directional edges connected, we observe that most of these nodes have a negative weight, while fewer nodes have a positive weight. The peak amount of node weight distribution is below zero, revealing that most nodes are in blue and have a slight net outflow or impact on other nodes. Simultaneously, some nodes (such as nodes near *East Falls Church* during the evening peak hours) are significantly impacted by others, leading to high positive net inflow nodes in deep red color.

Furthermore, nodes with the same colors (in red or blue) often form clusters located at specific locations within the metro network, notably near transfer stations (e.g., nodes near *Stadium-Armory, Metro Center*, and *Rosslyn* with net outflow) and terminal stations (e.g., nodes near *Branch Ave* and East Fall Church). This phenomenon suggests that tracks near terminals and transfer stations tend to exhibit stronger homogeneous dependencies on other nodes within the network. And the homogeneity within the clusters possibly reflects the local delay propagation (e.g., nodes near *Stadium-Armory*, *East Fall Church*, and between *Shady Grove* and *Metro center*). Moreover, the direction (positive or negative) of these clusters' net flows remains consistent during the entire morning peak but may vary in magnitude. For instance, nodes around Stadium-Armory consistently show a net outflow throughout the morning peak, but at 9:00 am, more nodes near Stadium-Armory exhibit a deep blue color, indicating a peak in their contribution values at that specific time.

Additionally, some of the red directional edges show an accumulatively unified direction, representing strong interdependencies among nodes. Some interdependencies can last for the entire morning peak, while others demonstrate clear dynamic changes during specific periods. For example, the transfer node *Rosslyn* shows a consistent extension of impact on nodes near *King St-Old Town*. At the same time, nodes near *Stadium-Armory* may exhibit a significant increase in impact on nodes near *L'Enfant Plaza, Gallery Palace*, and *East Fall Church* before 9:00 am, with the impact diminishing considerably after that time.

Figure 19 also shows that most of the phenomena that appear during morning peak hours still apply to evening peak hours. For instance, the trends of aggregated homogeneous dependencies among nodes near terminal and transfer stations persist. However, the trends of interdependencies among specific pairs of clusters changed. For instance, during the evening peak, nodes near *Stadium-Armory* are more likely to be represented in red, indicating a significant impact on the delay occurrence of other node clusters, such as *East Fall Church, Rosslyn*, and *Branch Ave*.

This phenomenon is opposite to what was observed during the morning peak hours, as mentioned earlier.

Moreover, some correlation trends that appear in the morning peak hours may also become less significant or disappear during the evening peak hours. For example, nodes on the red line between the terminal *Shady Grove* and the transfer station *Metro center* show a much less significant impact on the other nodes than they appear during the morning peak hours. This implies a clear tidal dynamic phenomenon in node delay correlations.

Summarizing the observations derived from the link graph in Figure 19, we can draw some general regulations regarding the delay interdependencies among network segments within the context of the Washington Metro network:

- Nodes in clusters between or near terminal and transfer stations exhibit homogenous, stronger, and longer-lasting dependencies on other nodes, especially those located in the clusters also between terminal and transfer stations. This indicates that the node clusters at the specific locations tend to appear the similar vulnerability spontaneously.
- The notable correlation between node clusters might indicate distinct tidal dynamics or persistent trends during morning and evening peaks. This refers that the vulnerability and delay propagation among nodes at various locations show the different dynamic shift over time.
- While most nodes located in the corridors connecting transfer stations tend to exert a relatively lesser impact on other nodes, there exists a smaller subset among them that is susceptible to significant influence from other nodes, displaying vulnerability.

6. Conclusion

In this paper, we employed a post-hoc explainability method to probe a data-driven black box model, aiming to uncover the correlation of delay occurrences in the metro network. Through this approach, we significantly improved the explainability of the data-driven model, allowing us to extract and solidify the domain knowledge on delay occurrence interdependencies dynamics acquired by the model.

The case study on the Washington metro network allows the implementation of the purposed methodology. By analyzing the explanation variables obtained through the post-hoc method, we uncovered distinct and significant patterns in the spatial-temporal features of the input data. Moreover, we unveiled the interdependency among different segments in the metro network's delay predictions. The mappings observed were "many-to-one," representing the relationship between input historical data and prediction results for individual nodes, and "many-to-many," revealing the interdependencies among various tracks within the network.

6.1 Key findings

We identified representative patterns that shed light on how historical delay data impacts current delay occurrences at different locations. Additionally, we observed consistent regulations in these dependencies over varying periods. By unearthing these findings, we gained valuable insights into the complex interactions and dependencies within the metro network. This knowledge contributes to a better understanding of delay occurrences and facilitates the development of effective strategies for managing and mitigating delays within the system. To be specific, the main research question is solved by answering the following sub-questions.

- **What are the spatial-temporal explanatory variables of network passenger delay occurrence prediction?**

The spatial-temporal explanatory variables for each delay prediction outcome are represented as a matrix, which serves as a mask to assess historical estimated delays’ importance using quantified unitless scores. Within the input time window range at a particular location, the importance score in the matrix corresponds to the significance of an individual estimated delay on historical time in shaping predictions made by the STGCN model. Each prediction outcome is associated with an explanatory variables matrix, offering insights into the contribution of delays with varying spatial-temporal characteristics.

- **How does the spatial-temporal feature of historical delay data contribute to the model prediction?**

This study uncovered three distinct patterns that elucidate the influence of historical estimated delay data across different spatial locations and temporal features on the prediction outcome. Notably, these patterns exhibit significant imbalances in their distribution across both time and space scales within the network. With regard to temporal features, it was observed that delays during the morning and evening peaks tend to exhibit dependencies on the historical delays of key nodes from a comparatively distant past than most of the other nodes' historical delays. During the hours that have lower passenger demand, a larger percentage of delay occurrences is only relative to the most recent delay. To describe the underlying universal patterns in the temporal correlation between the historical estimated delay and delay prediction outcomes, we have labeled these identified patterns as "Recent Dominance," "Partial Long Lasting," and "Temporal Global Relevant," based on their predominant characteristics. In the context of spatial feature importance, three patterns are revealed that reflect the dependencies between the prediction outcome of specific nodes and historical delay in different locations. According to their significant spatial distribution features of importance scores, these patterns are named "Key Nodes Dominate", "Spatial Imbalance", and "Spatial Global Relevant" respectively. Among these patterns with distinctive features, we found the common regulation that the terminal section and the segment near the transfer stations always show significant contributions to most of the delay prediction outcomes.

- **How does the spatial-temporal correlation of different metro network segments vary throughout the day?**

The importance scores within the explanatory variable matrix are transformed into unitless contribution values, providing insights into the directional dependencies among different segments within the network. More precisely, these explanatory variables, initially representing a "many-to-one" relationship between input historical data and individual prediction outcomes, are then transformed into a "many-to-many" mapping, reflecting correlations among various segments. This investigation has yielded two significant findings. Firstly, distinct tidal trends emerge in certain pairs of dependencies, and their significant contributions vary. The direction of these dependencies can appear, dismiss, or even reverse during morning and evening peak hours within a day. Moreover, the number of significant correlations changes over time, usually peaking during the hours of highest demand. The second finding pertains to the potential regulations of delay propagation. By exploring into the dynamics of delay dependencies, a clustering effect among network segments becomes evident during specific time intervals. Notably, node clusters proximate to terminals and transfer stations exhibit coherent dependencies among similar clusters in other branches of metro lines. Concurrently, directional delay propagation is observed along particular corridors, particularly during peak hours between the terminal stations and transfer stations.

The insights obtained from the results of this study can be validated and aligned with previous research. This study provides compelling evidence of the existence of spatial-temporal correlation patterns in delay occurrence among metro links. Similar phenomena have been demonstrated in other traffic networks as well (Ermagun et al., 2017; Lopez et al., 2017), supporting the validity and generalizability of the findings. Moreover, the dynamic correlation of delay occurrence points to the significance of node clusters at the core of the metro network, which can exert substantial influence on neighboring nodes, leading to delays that eventually spread throughout the network. This finding aligns with the rationale put forth in previous research that employed a minimal network flow model (Petri et al., 2009). Building on the insights from the work of Petri et al., a compelling approach to mitigating delay propagation would involve isolating and quarantining congested nodes within the network. Such proactive measures can effectively curtail the spread of delays, enhancing the overall reliability and efficiency of the metro system.

6.2 Contributions

Compared with existing research on explainable metro delay prediction, the significant advantage of this study is that it only requires basic passenger travel data and schedule information, while effectively capturing the complex dynamics of delay occurrences and dependencies among network segments through the data-driven model. By deciphering the model, we explored the underlying delay occurrence dependencies in metro operation and passenger travel data, which might be obscured due to unexpected events, external factors, and fluctuations in passenger demand. Although the correlations we uncovered do not imply causation of delay occurrences, these insights into the spatial-temporal correlations and dependencies enhance our understanding of the metro network's behavior and guide improved network performance and management. The methodology proposed in this study, along with the corresponding results, can assist operators in comprehending the interior dependencies of delay occurrences. The identified correlations should be supplemented with analytical and simulation models for metro operations to examine the relations between the (un)reliability of system components.

The insights gained from this study offer valuable support to metro operators in comprehending the intricate dynamics of delay occurrences within the network. This understanding provides a holistic perspective on how delays propagate and affect different parts of the system, enabling more informed decision-making. Operators can identify crucial nodes that play a significant role in delay propagation and proactively implement congestion management measures, schedule optimizations, and bottleneck node isolation during peak periods. Furthermore, these findings on delay correlations can aid in providing passengers with timely updates about possible delays and alternative routes, as well as assessing the effectiveness of new delay mitigation strategies. The identified patterns also serve as valuable guidance for developing targeted approaches to minimize delays and enhance network reliability. By integrating these insights into operational planning, operators can efficiently manage the metro system, leading to improved passenger satisfaction and system-wide performance.

The scientific contribution of this study lies in its innovative use of an XAI approach to uncover the dynamics of network segment correlations and phenomena, delve into the mechanics of delay propagation, and assess network vulnerability, surpassing the scope of existing research. Given the accurate prediction outcome of the data-driven model, the obtained insight not only increase the model’s explainability but also demonstrate applicability for decision-making support. This approach could be used to validate other studies on network analysis. Besides, the revealed insights like the clustering effect of homogeneous dependencies among nodes uncover mechanisms of delay propagation from a novel aspect. With additional data sources, further exploration could delve into the underlying reasons behind correlations appearing during specific periods and at specific locations.

Furthermore, the potential applications offer benefits not solely to present network operators but also extend to other transportation systems equipped with the necessary input data. The methodology, as showcased in this study, is versatile and can be easily tailored to a wide range of datasets, aiming to investigate correlations among diverse attributes of network segments, including but not limited to delay occurrences, passenger counts, and real-time status. Its applicability extends beyond metro and public transportation networks, encompassing multiple transport systems like road networks and shipping networks.

6.3 Limitations & Future research

This study's advantage lies in its ability to capture complex delay dynamics only using passenger travel and vehicle scheduling data through a data-driven model. However, the simplicity of the model training and parameter tuning for the black-box prediction model restricts its ability to capture delay occurrences with a step higher accuracy, even with the use of deeper neural network layers and primarily data transformation techniques, although the trend of delay occurrence is captured. The trustfulness of the explanatory variable comes from the prediction performance. The higher accuracy of prediction refers to the more reliable explanation in most cases. (Faber et al., 2021). This limitation arises due to the data distribution in the estimated delay dataset. Most nodes in the network graph experience delays lower than 5 seconds, while heavy delays are rare in some nodes. Such an imbalanced distribution can negatively impact the neural network's performance, which might otherwise perform well with datasets having a more even distribution. Addressing this issue in future research is crucial, as considering the input data distribution can enhance the performance of the black-box prediction model. This enhancement would result in more reliable and convincing explanations extracted from the black-box model.

Moreover, when the image pattern recognition technique is implemented to cluster the feature importance matrix image, we select the image that is closest to the midpoint to represent the pattern of all the other images in the same cluster. This is efficient but also may neglect the distinction among these images, which means some of the potential correlation patterns of historical data importance might be overlooked and represented by the identified dominant patterns.

There are several directions for future research based on this study. Conducting cross-comparisons across different metro networks would offer valuable insights into the consistency and generalizability of the identified phenomena across diverse systems. Exploring the intricate interplay between correlations among node clusters within the metro network graph and the size of each cluster could provide a deeper understanding of how network structure influences delay propagation dynamics. Delving into temporal variations, such as tidal trends, over extended periods would uncover deeper insights into the evolution of delay correlations and their relationship to time-dependent factors.

Besides, when analyzing the dynamic correlation among nodes in the link graph, only the significant correlations that surpass a threshold level are visualized in the figure, and the analysis is conducted qualitatively rather than quantitatively. This approach could limit the potential to compare the dynamic of correlation among various metro networks, or to accurately quantify the variations in correlations within a specific network. As a result, there is a need for a more quantitative method to characterize the dynamic correlations among nodes based on the derived explanatory variables in this study. Expanding the scope of input data could also yield valuable improvements. Incorporating additional datasets that are widely available to most of the circumstances, such as passenger count data from smart cards, into the black-box model as input variables could significantly enhance its accuracy and provide more compelling explanations. This augmentation could contribute to refining the model's predictive capabilities, ultimately leading to more reliable and informative insights. Finally, beyond the overarching understanding of delay propagation as conducted in this study, there's potential for future research to offer more direct and actionable recommendations for scheduling management and operator decision-making. Future research can explore how to improve the network robustness according to the identified delay occurrence characteristic, and to what extent the identified correlation among network segment could help with public transport delay recovery. Such research could provide a bridge between the identified delay propagation phenomena and their translation into concrete strategies for real-world application, enhancing the practical impact of this study's findings.

Acknowledgements

In this report, GPT-3.5 large language model is employed to assist in grammar checking and fine-tuning of phrases.

Bibliography

Adadi, A., Berrada, M., 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. https://doi.org/10.48550/arXiv.1910.10045

Bešinović, N., 2020. Resilience in railway transport systems: a literature review and research agenda. Transport Reviews 40, 457–478. https://doi.org/10.1080/01441647.2020.1728419

Bui, K.-H.N., Yi, H., Cho, J., 2021. UVDS: A New Dataset for Traffic Forecasting with Spatial-Temporal Correlation, in: Nguyen, N.T., Chittayasothorn, S., Niyato, D., Trawiński, B. (Eds.), Intelligent Information and Database Systems, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 66–77. https://doi.org/10.1007/978-3-030-73280-6_6

Cats, O., Hijner, A.M., 2021. Quantifying the cascading effects of passenger delays. Reliability Engineering & System Safety 212, 107629. https://doi.org/10.1016/j.ress.2021.107629

Dai, R., Xu, S., Gu, Q., Ji, C., Liu, K., 2020. Hybrid Spatio-Temporal Graph Convolutional Network: Improving Traffic Prediction with Navigation Data.

Dalmau, R., Ballerini, F., Naessens, H., Belkoura, S., Wangnick, S., 2021. An explainable machine learning approach to improve take-off time predictions. Journal of Air Transport Management 95, 102090. https://doi.org/10.1016/j.jairtraman.2021.102090

Defferrard, M., Bresson, X., Vandergheynst, P., 2017. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. https://doi.org/10.48550/arXiv.1606.09375

Ermagun, A., Chatterjee, S., Levinson, D., 2017. Using temporal detrending to observe the spatial correlation of traffic. PLOS ONE 12, e0176853. https://doi.org/10.1371/journal.pone.0176853

Faber, L., K. Moghaddam, A., Wattenhofer, R., 2021. When Comparing to Ground Truth is Wrong: On Evaluating GNN Explanation Methods, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Presented at the KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, Virtual Event Singapore, pp. 332–341. https://doi.org/10.1145/3447548.3467283

Faroqi, H., Mesbah, M., Kim, J., 2017. Spatial-Temporal Similarity Correlation between Public Transit Passengers Using Smart Card Data. Journal of Advanced Transportation 2017, e1318945. https://doi.org/10.1155/2017/1318945

Fey, M., Lenssen, J.E., 2019. Fast Graph Representation Learning with PyTorch Geometric. https://doi.org/10.48550/arXiv.1903.02428

Fontoura, W.B., Ribeiro, G.M., Chaves, G.D.L.D., 2020. A framework for evaluating the dynamic impacts of the Brazilian Urban Mobility Policy for transportation socioeconomic systems: A case study in Rio de Janeiro. Journal of Simulation 14, 316–331. https://doi.org/10.1080/17477778.2019.1701392

Gordon, J.B., Koutsopoulos, H.N., Wilson, N.H.M., Attanucci, J.P., 2013. Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data. Transportation Research Record 2343, 17–24. https://doi.org/10.3141/2343-03

Guérin, J., Boots, B., 2018. Improving Image Clustering With Multiple Pretrained CNN Feature Extractors.

Hammond, D.K., Vandergheynst, P., Gribonval, R., 2011. Wavelets on graphs via spectral graph theory. Applied and Computational Harmonic Analysis 30, 129–150. https://doi.org/10.1016/j.acha.2010.04.005

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Heglund, J.S.W., Taleongpong, P., Hu, S., Tran, H.T., 2020. Railway Delay Prediction with Spatial-Temporal Graph Convolutional Networks, in: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). Presented at the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), pp. 1–6. https://doi.org/10.1109/ITSC45102.2020.9294742

Kipf, T.N., Welling, M., 2017. Semi-Supervised Classification with Graph Convolutional Networks. https://doi.org/10.48550/arXiv.1609.02907

Krishnakumari, P., Cats, O., van Lint, H., 2020. Estimation of metro network passenger delay from individual trajectories. Transportation Research Part C: Emerging Technologies 117, 102704. https://doi.org/10.1016/j.trc.2020.102704

Kumar, I.E., Venkatasubramanian, S., Scheidegger, C., Friedler, S.A., 2020. Problems with Shapley-value-based explanations as feature importance measures. Presented at the International Conference on Machine Learning.

Lee, W.-H., Yen, L.-H., Chou, C.-M., 2016. A delay root cause discovery and timetable adjustment model for enhancing the punctuality of railway services. Transportation Research Part C: Emerging Technologies 73, 49–64. https://doi.org/10.1016/j.trc.2016.10.009

Li, M., Zhu, Z., 2021. Spatial-Temporal Fusion Graph Neural Networks for Traffic Flow Forecasting. https://doi.org/10.48550/arXiv.2012.09641

Lopez, C., Leclercq, L., Krishnakumari, P., Chiabaut, N., van Lint, H., 2017. Revealing the day-to-day regularity of urban congestion patterns with 3D speed maps. Sci Rep 7, 14029. https://doi.org/10.1038/s41598-017-14237-8

Louie, J., Shalaby, A., Habib, K.N., 2017. Modelling the impact of causal and non-causal factors on disruption duration for Toronto's subway system: An exploratory investigation using hazard modelling. Accident Analysis & Prevention 98, 232–240. https://doi.org/10.1016/j.aap.2016.10.008

Lu, Q.-C., Lin, S., 2019. Vulnerability Analysis of Urban Rail Transit Network within Multi-Modal Public Transport Networks. Sustainability 11, 2109. https://doi.org/10.3390/su11072109

Lundberg, S.M., Erion, G.G., Lee, S.-I., 2019. Consistent Individualized Feature Attribution for Tree Ensembles. https://doi.org/10.48550/arXiv.1802.03888

Menno Yap, Oded Cats, 2021. Predicting disruptions and their passenger delay impacts for public transport stops. Transportation 48, 1703–1731. https://doi.org/10.1007/s11116-020-10109-9

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., 2016. WaveNet: A Generative Model for Raw Audio. https://doi.org/10.48550/arXiv.1609.03499

Petri, G., Jensen, H.J., Polak, J.W., 2009. Global and local information in traffic congestion. EPL 88, 20010. https://doi.org/10.1209/0295-5075/88/20010

Rößler, D., Reisch, J., Hauck, F., Kliewer, N., 2021. Discerning Primary and Secondary Delays in Railway Networks using Explainable AI. Transportation Research Procedia, 23rd EURO Working Group on Transportation Meeting, EWGT 2020, 16-18 September 2020, Paphos, Cyprus 52, 171–178. https://doi.org/10.1016/j.trpro.2021.01.018

Sánchez-Martínez, G.E., 2017. Inference of Public Transportation Trip Destinations by Using Fare Transaction and Vehicle Location Data: Dynamic Programming Approach. Transportation Research Record 2652, 1–7. https://doi.org/10.3141/2652-01

Shao, Z., Zhang, Z., Wang, F., Xu, Y., 2022. Pre-training Enhanced Spatial-temporal Graph Neural Network for Multivariate Time Series Forecasting, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1567–1577. https://doi.org/10.1145/3534678.3539396

Spanninger, T., Trivella, A., Büchel, B., Corman, F., 2022. A review of train delay prediction approaches. Journal of Rail Transport Planning & Management 22, 100312. https://doi.org/10.1016/j.jrtpm.2022.100312

Su, F., Dong, H., Jia, L., Tian, Z., Sun, X., 2017. Space–time correlation analysis of traffic flow on road network. Int. J. Mod. Phys. B 31, 1750027. https://doi.org/10.1142/S0217979217500278

Sun, J. (Daniel), Liu, Q., Peng, Z., 2011. Research and Analysis on Causality and Spatial-Temporal Evolution of Urban Traffic Congestions—A Case Study on Shenzhen of China. Journal of Transportation Systems Engineering and Information Technology 11, 86–93. https://doi.org/10.1016/S1570-6672(10)60143-2

Taleongpong, P., Hu, S., Jiang, Z., Wu, C., Popo-Ola, S., Han, K., 2022. Machine learning techniques to predict reactionary delays and other associated key performance indicators on British railway network. World Transit Research.

Tang, R., De Donato, L., Bešinović, N., Flammini, F., Goverde, R.M.P., Lin, Z., Liu, R., Tang, T., Vittorini, V., Wang, Z., 2022. A literature review of Artificial Intelligence applications in railway systems. Transportation Research Part C: Emerging Technologies 140, 103679. https://doi.org/10.1016/j.trc.2022.103679

Tiong, K.Y., Ma, Z., Palmqvist, C.-W., 2023. A review of data-driven approaches to predict train delays. Transportation Research Part C: Emerging Technologies 148, 104027. https://doi.org/10.1016/j.trc.2023.104027

van Oort, N., n.d. Service Reliability and Urban Public Transport Design.

Wales, J., Marinov, M., 2015. Analysis of delays and delay mitigation on a metropolitan rail network using event based simulation. Simulation Modelling Practice and Theory 52, 52–77. https://doi.org/10.1016/j.simpat.2015.01.002

Wei, D., Liu, H., Qin, Y., 2015. Modeling cascade dynamics of railway networks under inclement weather. Transportation Research Part E: Logistics and Transportation Review 80, 95–122.

https://doi.org/10.1016/j.tre.2015.05.009

Yang, Y., Jia, L., Qin, Y., Han, S., Dong, H., 2017. Understanding structure of urban traffic network based on spatial-temporal correlation analysis. Modern Physics Letters B 31, 1750230–364. https://doi.org/10.1142/S021798491750230X

Ying, R., Bourgeois, D., You, J., Zitnik, M., Leskovec, J., 2019. GNNExplainer: Generating Explanations for Graph Neural Networks.

Yu, B., Yin, H., Zhu, Z., 2018. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. pp. 3634–3640. https://doi.org/10.24963/ijcai.2018/505

Zhang, D., Xu, Y., Peng, Y., Du, C., Wang, N., Tang, M., Lu, L., Liu, J., 2023. An Interpretable Station Delay Prediction Model Based on Graph Community Neural Network and Time-Series Fuzzy Decision Tree. IEEE Transactions on Fuzzy Systems 31, 421–433. https://doi.org/10.1109/TFUZZ.2022.3181453

Zhang, Z., Li, M., Lin, X., Wang, Y., He, F., 2019. Multistep speed prediction on traffic networks: A deep learning approach considering spatio-temporal dependencies. Transportation Research Part C: Emerging Technologies 105, 297–322. https://doi.org/10.1016/j.trc.2019.05.039

Zhu, Y., Koutsopoulos, H.N., Wilson, N.H.M., 2017. A probabilistic Passenger-to-Train Assignment Model based on automated data. Transportation Research Part B: Methodological 104, 522–542. https://doi.org/10.1016/j.trb.2017.04.012