LEVERAGING DATA ANALYTICS AND PREDICTIVE MODELING TO INFORM STRATEGIC ENHANCEMENTS IN URBAN METRO SYSTEMS

LI RUI LIN

SULTAN IDRIS EDUCATION UNIVERSITY

2025

LEVERAGING DATA ANALYTICS AND PREDICTIVE MODELING TO INFORM STRATEGIC ENHANCEMENTS IN URBAN METRO SYSTEMS

LI RUI LIN

THESIS PRESENTED TO QUALIFY FOR A DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTING AND META-TECHNOLYGY
SULTAN IDRIS EDUCATION UNIVERSITY

2025

UPSI/IPS-3/BO 32
Pind : 00 m/s: 1/1

UNIVERSITI PENDIDIKAN SULTAN IDRIS
SULTAN IDRIS EDUCATION UNIVERSITY

Please tick (√)	
Project Paper	
Masters by Research	
Master by Mixed Mode	
PhD	√

INSTITUTE OF GRADUATE STUDIES

DECLARATION OF ORIGINAL WORK

This declaration is made on the 3th November 2025

i. Student's Declaration:

I, Li Ruilin, P20211002409, Faculty of Computing and Meta-technolygy (PLEASE INDICATE STUDENT'S NAME, MATRIC NO. AND FACULTY) hereby declare that the work entitled Leveraging Data Analytics and Predictive Modeling to Inform Strategic Enhancements in Urban Metro Systems is my original work. I have not copied from any other students' work or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for me by another person.

Li Rui lin

Signature of the student

ii. Supervisor's Declaration:

I Wong Yoke Seng (SUPERVISOR'S NAME) hereby certifies that the work entitled Leveraging Data Analytics and Predictive Modeling to Inform Strategic Enhancements in Urban Metro Systems (TITLE) was prepared by the above named student, and was submitted to the Institute of Graduate Studies as a * partial/full fulfillment for the conferment of Doctor of Philosophy (PhD) (PLEASE INDICATE THE DEGREE), and the afore mentioned work, to the best of my knowledge, is the said student's work.

27,08,2025

Date

Signature of the Supervisor

UPSI/IPS-03/BO31 iii
Pind:02 m/s: 1/1

UNIVERSITI PENDIDIKAN SULTAN IDRIS
SULTAN IDRIS EDUCATION UNIVERSITY

INSTITUT PENGAJIAN SISWAZAH / *INSTITUTE OF GRADUATE STUDIES*

BORANG PENGESAHAN PENYERAHAN TESIS/DISERTASI/LAPORAN KERTAS PROJEK
DECLARATION OF THESIS/DISSERTATION/PROJECT PAPER FORM

Tajuk / *Title*: Leveraging Data Analytics and Predictive Modeling to Inform Strategic Enhancements in Urban Metro Systems

No. Matrik /: P20211002409
Matric's No.:

Saya / *I am*: Li Ruilin
(Nama Pelajar / *Student's Name*)

mengaku membenarkan Tesis/Disertasi/Laporan Kertas Projek (Kedoktoran/Sarjana) ini disimpan di Universiti Pendidikan Sultan Idris (Perpustakaan Tuanku Bainun) dengan syarat-syarat kegunaan seperti berikut: -
acknowledged that Universiti Pendidikan Sultan Idris (Tuanku Bainun Library) reserves the right as follows: -

1. Tesis/Disertasi/Laporan Kertas Projek ini adalah hak milik Universiti Pendidikan Sultan Idris.
The Thesis/Dissertation/Project Paper is the property of Universiti Pendidikan Sultan Idris.
2. Perpustakaan Tuanku Bainun dibenarkan membuat Salinan untuk tujuan rujukan dan pendidikan.
Tuanku Bainun Library has the right to make copies for the purpose of reference and education.
3. Perpustakaan dibenarkan membuat salinan Tesis/Disertasi/Laporan Kertas Projek ini sebagai bahan pertukaran antara Institusi Pengajian Tinggi.
The library has the right to make copies of the thesis/Dissertation/Project Paper for academic exchange.
4. Sila tandakan (√) bagi pilihan kategori di bawah / *Please tick (√) for category below: -*

√	**AKSES TERBUKA** ***OPEN ACCESS***	Saya bersetuju bahawa Tesis/Disertasi/Kertas Projek boleh diterbitkan sebagai salinan fizikal atau akses terbuka dalam talian. *I agree that the Thesis/Dissertation/Project Paper may be published as a physical copy or open access online.*
	TERHAD** ****LIMITED	*Mengandungi maklumat terhad seperti yang dinyatakan oleh organisasi/institusi di mana penyelidikan dijalankan. (Perpustakaan akan menyekat akses sehingga ____ tahun). Contains restricted information as specified by the organization/institution where the research was conducted. (Library will restrict access until ____ years).*
	SULIT** ****CONFIDENTIAL	*Mengandungi maklumat sulit di bawah Akta Rahsia Rasmi 1972. Contains confidential information under the Official Secrets Act 1972.*
	EMBARGO	Tesis/Disertasi/Kertas Projek ini tertakluk kepada tempoh embargo selama 2 tahun atas sebab-sebab berikut: *This Thesis/Dissertation/Project Paper is subject to an embargo period of 2 years for the following reasons:*

*Nota / *Notes:
i. Jika tiada pilihan yang ditandai, pilihan pertama akan dilaksanakan secara automatik.
If no option is marked, the first option will be automatically implemented.
ii. Jika Tesis/Disertasi/Kertas Projek ini dikategorikan sebagai **SULIT** atau **TERHAD**, **sila lampirkan surat dari pihak berkuasa/organisasi** yang berkaitan dengan **menyatakan alasan dan tempoh tesis ini perlu diklasifikasikan sebagai TERHAD atau SULIT**. / *If the Thesis/Dissertation/Project Paper is categorized as* ***CONFIDENTIAL*** *or* ***LIMITED, kindly attach a letter from the relevant authority/organization stating the reason and duration for the thesis should be classified as LIMITED or CONFIDENTIAL.***

LiRuilin

(Tandatangan Pelajar/ Signature)

Tarikh:27,08,25

TS. DR. WONG YOKE SENG
Pensyarah Kanan
Jabatan Komputeran
Fakulti Seni, Komputeran & Industri Kreatif
Universiti Pendidikan Sultan Idris
35900 Tanjong Malim Perak

(Tandatangan Penyelia / Signature of Supervisor)
& (Nama & Cop Rasmi / Name & Official Stamp)

Tarikh: 27,08,25

ACKNOWLEDGEMENTS

Time has flown by, and my four years of doctoral study are drawing to a close. Looking back, I am filled with gratitude for the supervisors, friends, and fellow students who have supported and encouraged me throughout this journey, allowing me to progress continually in my academic pursuits and grow steadily through exploration. First and foremost, I would like to extend my heartfelt thanks to Dr Wong Yoke Seng, who has provided patient guidance in every stage of my work—from topic selection and research design to thesis writing. His rigorous academic approach and broad scholarly perspective have enabled me to examine problems from multiple angles. I am equally grateful to Dr Suliana binti Sulaiman, Dr Shir Li Wang, Dr Mad Helmi Bin Ab. Majid, and Dr Lee Hoiyeh for their professional input and inspiration, which have broadened my research horizons. In addition, I owe a debt of gratitude to my leadership team and colleagues for their understanding and support, enabling me to allocate my time and energy more effectively. My fellow doctoral students have not only sparked new ideas in conceptualising my thesis and designing experiments but have also been a source of strength as we advanced together. Each discussion and debate have been invaluable, furthering my growth both academically and personally. I am profoundly grateful to my parents and parents-in-law for their unwavering care and encouragement, which have fortified my confidence to overcome numerous challenges. My younger brother's simple words— "Study well, and if you run out of money, you can count on me"— have lit up the many late nights I spent persevering. I especially wish to thank my wife, Ms Li Honghong, whose understanding and patience have provided a powerful driving force as I juggled demanding coursework and personal commitments. Her warmth and support have allowed me to regain my energy in times of exhaustion and convinced me that by persevering, I would ultimately reach my goal. Lastly, I would like to offer my sincere gratitude to everyone who has assisted and encouraged me. Your kindness and contributions have enabled me to experience a genuine sense of integration and warmth in both academic work and life. May this thesis serve not only as the culmination of my scholarly endeavours but also as a tribute to the profound support I have received from all of you.

ABSTRACT

With the acceleration of urbanization, metro passenger flow patterns in megacities have become increasingly complex. Shanghai, with the world's largest metro network, poses significant challenges to accurate demand forecasting. Traditional approaches often neglect passenger behavioral heterogeneity and service satisfaction, resulting in limited prediction performance. This study integrates travel behavior mining and satisfaction perception into forecasting by proposing a Context-Aware Graph Transformer (CaGT) model that fuses spatio-temporal features with human-centered factors.The dataset comprises smart card transaction records (5-minute intervals), network topology of more than 500 stations, and 2660 valid passenger satisfaction surveys based on the SERVQUAL framework. Analysis revealed that peak-hour demand is highly concentrated, with transfer bottlenecks at People's Square and Century Avenue, and five major commuting corridors were identified. Survey results showed that crowding, convenience transfer, and in-train comfort significantly affect satisfaction and loyalty.The CaGT model integrates Graph Attention Networks (GAT) with Transformer encoders, embedding contextual factors such as satisfaction, crowding index, and temporal preferences into dynamic graph weights. Experimental evaluations demonstrated that CaGT significantly outperformed benchmark models (MLP, LSTM, STGCN, ASTGCN), achieving RMSE = 10.21, MAE = 6.48, and MAPE = 8.7%, with superior accuracy particularly during peak hours.This study contributes theoretically by bridging passenger behavior with graph-based forecasting, and practically by offering actionable insights for metro scheduling, transfer management, and passenger service improvement. The framework holds strong potential for cross-city transferability and multimodal transit integration in the future.

MEMANFAATKAN ANALITIK DATA DAN PEMODELAN RAMALAN UNTUK MEMAKLUMKAN PENAMBAHBAIKAN STRATEGIK DALAM SISTEM METRO BANDAR

ABSTRAK

Dengan kepesatan urbanisasi, corak aliran penumpang dalam sistem metro di bandar mega menjadi semakin kompleks. Shanghai, dengan rangkaian metro terbesar di dunia, menghadapi cabaran besar dalam penghasilan ramalan permintaan yang tepat. Kaedah tradisional sering mengabaikan kepelbagaian tingkah laku penumpang dan tahap kepuasan perkhidmatan, lantas mengakibatkan prestasi unjuran yang terhad. Kajian ini menggabungkan perlombongan tingkah laku perjalanan dan persepsi kepuasan dalam ramalan dengan membangunkan Model Pengubah Graf Berkesedaran Konteks (CaGT) yang menggabungkan ciri ruang-masa dengan faktor berpusatkan manusia.Set data kajian ini mengandungi rekod transaksi kad pintar metro (selang 5 minit), topologi rangkaian lebih 500 buah stesen dan 2660 soal selidik kepuasan penumpang yang sah berdasarkan rangka kerja SERVQUAL. Analisis menunjukkan bahawa permintaan waktu puncak tertumpu pada stesen teras seperti People's Square dan Century Avenue, dengan lima koridor utama perjalanan ulang-alik dikenal pasti. Hasil soal selidik membuktikan bahawa tahap kesesakan, kemudahan pertukaran dan keselesaan dalam tren memberikan kesan yang signifikan terhadap kepuasan dan kesetiaan penumpang.Model CaGT ini menggabungkan Rangkaian Perhatian Graf (GAT) dengan pengekod Pengubah, serta memasukkan faktor kontekstual seperti kepuasan, indeks kesesakan dan keutamaan masa dalam pemberat graf dinamik. Penilaian eksperimen menunjukkan bahawa CaGT mengatasi model penanda asas (MLP, LSTM, STGCN, ASTGCN) dengan ketara, mencapai RMSE = 10.21, MAE = 6.48 dan MAPE = 8.7%, dengan ketepatan yang lebih baik khususnya pada waktu puncak.Sumbangan kajian ini adalah pada aspek teori, iaitu menjambatani tingkah laku penumpang dengan ramalan berasaskan graf, dan pada aspek praktikal, iaitu menyediakan pandangan yang boleh dilaksanakan untuk penjadualan metro, pengurusan pertukaran serta peningkatan perkhidmatan penumpang. Rangka kerja ini juga berpotensi kukuh untuk pemindahan silang bandar dan integrasi pelbagai mod transit pada masa hadapan.

CONTENTS

CHAPTER 3 METHODOLOGY

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

3DQN	Doppio Duello Deep Q-Network
A3T-GCN	Attention Temporal Graph Convolutional Network
ABC	ArtificialBee Colony
ABi-GRU	Attention-based Bidirectional Gated Recurrent Unit
AdaGrad	Adaptive Gradient Algorithm
AFC	Automatic Fare Collection
AHP	Analytic Hierarchy Process
AI	Artificiale Intelligence
AMVs	AutonomousModularVehicles
APC	Automatic Passenger Counting xivxiv
APGCN	Adaptive Position Graph Convolution ()
AR	Augmented Reality
AST3DRNet	Attention-based Spatio-Temporal 3D Residual Neural Network
ASTGCN	Attention-based Spatio-Temporal Graph Convolutional Network
AVL	Automatic Vehicle Location
BiGRU	Bidirectional Gated Recurrent Units
BLIP	Bus Lanes with Intermittent Priority
BRT	Bus Rapid Transit
C3D	Convoluzionale 3D
CaGT	Context-aware Graph Transformer
CAS-CNN	Channel Attention Separable Convolutional Neural Network
CCD	Cross-Correlation Distance

CEEMDAN	Complete Ensemble Empirical Mode Decomposition with Adaptive Noise
CMR-GCN	Rete di convoluzionegraficarelazionale in modalitàincrociata
CNN	Convolutional Neural Networks
C-V2X	Cellular Vehicle-to-Everything
DBL	Dynamic Bus Lane
DBN	Deep Belief Network
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCM	Differential Compensation Method
DCRNN	Diffusion Convolutional RNN
DFRGCN	Deep Full Residual Graph Convolution Network
DGCNs	Dynamic Graph Convolutional Networks
DL	Deep Learning
DNN-BTF	Deep Neural Network-based Traffic Flow
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
DSTAGNN	Dynamic Spatiotemporal Awareness Graph Neural Network
DSTGAT	Dynamic Spatio-Temporal Graph Attention Network
DTW	Dynamic Time Warping
EMD	Empirical Mode Decomposition
ETL	Exit State Traffic Light
FSVR	RegressioneVettoriale di SupportoFuzzy
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GAT	Graph Attention Network
GATCN	Graph Attention Temporal Convolutional Network

GBRT	Gradient Boosting Regression Trees
GCAM	Graph Convolution Attention Module
GCN	Graph Convolutional Networks
GE-AGCN	Graph Embedding Adaptive Graph Convolution Network
GNN	Graph Neural Networks
GNN-MTLF	Graph Neural Network-based Multi-Task Learning Framework
GRANP	Graph Recurrent Attention Neural Process
GRU	Gated Recurrent Unit
GSTGAT	Gated Spatiotemporal Graph Attention Network
HOV	High-Occupancy Vehicle
HPPPO	HybridProximalPolicyOptimisation
IQR	Interquartile Range
ITS	Intelligent Transport Systems
KNN	K-NearestNeighbors
LCM	LatentClass Model
LGC	Learnable Graph Convolution
LGSTGCN	Local-Global Spatio-Temporal Graph Convolution Network
LSSVM	Least SquaresSupportVectorMachine
LSTM	Long Short-Term Memory Networks
MAE	ErroreAssolutoMedio
MAPE	ErroreMedio AssolutoPercentuale
MAPE	Mean Absolute Percentage Error
MASTGN	Multi-Stage Attention Spatio-Temporal Graph Network
MD-GCN	Multi-Scale Temporal Dual Graph Convolution Network

MHSTA-GCN	Multi-Head Spatio-Temporal Attention Graph Convolution Network
MILP	Mixed-IntegerLinearProgramming
MINLP	Mixed-IntegerNonlinearProgramming
MLP	Multilayer Perceptron
MPC	Modello di ControlloPredittivo
MSSTRN	Multi-Scale Spatio-Temporal Recurrent Network
M-STGAT	MultidimensionaleSpazio-temporale Graph Attention Network
MSTNet	Multi-Head Attention Spatio-Temporal Graph Neural Network
MTA	AutoritàMetropolitanadeiTrasporti
MTDT	Multi-Task Deep Learning Digital Twin
MTNDP	Multi-modal TransportNetworkDesignProblem
NLP	Natural Language Processing
OD	Origin-Destination
ODMTS	On-Demand Multimodal Transport System
PACF	PartialAutocorrelationFunction
PDVRPTW	pickup and delivery vehicle routing problem with time windows
POI	Point of Interest
PSO	Particle Swarm Optimisation
RGNN	Retineuronaligrafichericorrenti
RL	Reinforcement Learning
RMSE	Root Mean Square Error
RMSprop	Root Mean Square Propagation
RNN	Retineuronaliricorrenti
RNNs	Recurrent Neural Networks

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

RSMT	Rectilinear Steiner Minimal Tree
SAC	Soft Actor-Critic
SARIMA	Seasonal Autoregressive Integrated Moving Average
SDC	Subsequence-Based Dilated Convolution
SEM	structural equation model
SGD	Stochastic Gradient Descent
SPTC	Shanghai Public Transportation Cards
SS-GRU	Single-Step GRU
SST-GNN	Simplified Spatiotemporal Traffic Prediction Graph Neural Network
STANN	Spatio-Temporal Attention Neural Network
STCGCN	Spatio-Temporal Correlation Graph Convolution Network
ST-DGDE	Spatiotemporal Dynamic Graph Differential Equation Network
STFGNN	Spatial-Temporal Fusion Graph Neural Network
STGAT	Spatial-Temporal Graph Attention Networks
STGCN	Spatio-Temporal Graph Convolutional Networks
ST-GDN	Spatio-Temporal Graph Diffusion Network
STGGAT	Spatio-Temporal Gated Graph Attention Network
STGNNs	Spatiotemporal Graph Neural Networks
STIDGCN	Spatio-Temporal Interactive Dynamic Graph Convolution Network
ST-MGAT	Spatio-Temporal Multi-Graph Head Attention Network
STRGCN	Spatio-Temporal Residual Graph Convolution Networks
STSAtt	AttenzioneSincronaSpaziotemporale
STSGCN	Spatio-Temporal Synchronous Graph Convolutional Network
SVR	SupportVectorRegression
T-GAT	Temporal Graph Attention Networks

TOD	Transit-Oriented Development
TSCND	Time Subsequence-Based Convolutional Network

CHAPTER 1

INTRODUCTION

1.1 Overview

This study focusing on the Shanghai metro system, this research provides insights into how big data and predictive modelling can be used to address the complex challenges faced by metro systems in modern urbanisation. As cities continue to expand and population densities increase, public transport systems, especially metro systems, are carrying increasingly heavy loads. In this context, how to enhance passenger experience while ensuring efficient operation has become a key issue for city managers and researchers. By systematically collecting, processing and analysing multi-source data from the Shanghai metro, this study constructs an accurate prediction model, which not It only improves the accuracy of passenger flow prediction, but also puts forward a series of data-driven strategy improvement suggestions, providing strong support for the optimisation of the Shanghai metro system.

This study began with data collection and processing, and details how to integrate multi-source data to ensure the reliability and accuracy of the analysis. The metro system involves a wide variety of data, including historical passenger swipe data, single-trip ticket data, mobile phone swipe data, and so on. In order to obtain representative conclusions, it is necessary rigorously cleaned and processed these data, eliminated incomplete or abnormal data, and integrated data from different sources into a unified analysis platform through multiple data fusion techniques. The successful completion of this process lays a solid foundation for the subsequent prediction model construction. In the model construction phase, the context-aware graph transformer model (CaGT), which utilises a hybrid architecture combining a transformer encoder, a graph attention network (GAT) had been proposed, and a transformer decoder to capture the temporal and spatial dependencies inherent in the city-area usage data Studies have shown that this model significantly improves the accuracy and adaptability of the predictions, far surpassing the traditional prediction methods. This analysis is based on the data, passenger flow prediction model and questionnaire results. However, this is not to propose a definitive solution, this analysis aims to identify key areas where strategic interventions could improve the efficiency, reliability and overall user experience of the metro system. By utilising the insights gained from both quantitative and qualitative data, we can outline a range of optimisation strategies to address the most pressing challenges facing the metro network.

Next, the chapter sets out the research questions based on the context of the study and carefully defines the research objectives to ensure relevance and practicality of the study. The chapter also presents the scope of the study and explains the significance of the study to help readers better understand and engage with the study.

In addition, the chapter also concludes with a detailed description of the research structure of this paper, with each part of the study carefully designed from the research background to the research methodology to the expected outcomes, with a view to providing scientific and practical suggestions for the strategic improvement of the rail transport system in China's major cities.

1.2 Research Background

Since the first ground bus line opened in Shanghai in 1914, China's urban public transport system has shown profound historical heritage and development potential. This year marked the beginning of public transport in China and heralded a new era.

Over the following decades, China's public transport system gradually evolved into a multi-tiered and diversified framework. From initial trams to subsequent ground transport tools, and now to coordinated ground and underground systems, these modes of transport have continuously been updated and have played important roles in their respective fields.

The completion and opening of Beijing Subway Line 1 in 1969 was a landmark event in the history of urban rail transit in China. It marked the beginning of rail transit in China and laid a solid foundation for the country's public transport system (QS Theory (Ed.) 2021). Subsequently, Chinese cities entered a period of rapid urbanisation, with large populations and expanding metropolitan areas posing significant challenges to urban transport. By the end of 2023, 47 cities worldwide had populations exceeding

10 million, 17 of which were in China, reflecting the prosperity of China's urban development (China Statistical Yearbook, 2023). Among these cities, Chongqing and Shanghai ranked first and second in population in China. However, this is not always a cause for celebration. While a city's population size can indicate economic strength, economic development and increased private car ownership have led to traffic congestion and related problems in China's large cities, sometimes even more severely than in other major cities worldwide.

Take Beijing as an example. Between 1993 and 2003, the traffic situation in the capital was a prolonged struggle. Vehicle speeds decreased from 45 kilometers per hour in 1994 to 12 kilometers per hour in 2003, with more than 80 major roads nearing saturation and experiencing excessive congestion (Window to the Capital, 2018). A similar situation was evident in Shanghai. Despite the lack of comprehensive statistical data, citizens reported severe congestion at major intersections during peak hours, sometimes taking over half an hour to pass through tunnels (Window to the Capital, 2018). Even in this context, public transport systems in Shanghai and Beijing faced immense pressure, with less than half of all trips completed using public transport (Municipal State-Owned Assets Supervision and Administration Commission, 2022).

Table 1.1

Severe Traffic Congestion Section of Beijing and Shanghai

Year	Quantity	
	Beijing	Shanghai
1993	27	29
1994	36	42
1995	55	54
1999	99	96
2003	87	90

Since 2000, both the central and local governments of China have re-evaluated public transport development and committed to promoting it as the dominant mode of urban transport (State Council of the People's Republic of China, 2020). Before 2010, rail transit construction was mainly led by the state and developed rapidly. However, starting in 2012, the government's attitude towards public transport shifted significantly, emphasising that urban systems should primarily focus on surface transport (State Council of the People's Republic of China, 2019). By 2018, the government's public transport plans became more defined, with the central government intensifying the approval and construction of urban rail transit. It also required major cities to enhance the integration of urban rail and surface transport, moving towards a more coordinated development approach (State Council of the People's Republic of China, 2019).

For example, in Shanghai, the period from 2000 to 2010 saw rapid development in urban rail transit, with operating mileage increasing from 62 kilometres to 452 kilometres. Ground bus services also underwent updates, and infrastructure construction accelerated. Despite the rapid growth of rail transit, public transport remained the core mode of urban passenger transport, accounting for more than half of the total (Xue and Liu, 2019). By the end of 2021, the Shanghai Metro's operating scale reached 831 kilometres, with an average daily passenger volume exceeding 10 million. The metro system included 7,226 vehicles. At the same time, the Shanghai government increased investment in surface transport. From 2011 to 2020, government subsidies for surface transport grew from 4.22 billion yuan to 8.57 billion yuan, with surface transport investment accounting for more than 70% of total public transport subsidies (Yu, 2022).

Table 1.2

China's Urban Public Transport Evolution

Period/Year	Event & Development	Implication
1914	Shanghai opened the first ground bus line	Beginning of China's modern public transport
1969	Beijing Subway Line 1 opened	Start of China's rail transit era
1990S	Rapid urbanisation, severe congestion in mega-cities	Highlighted need for efficient public transport
2000-2010	Rail transit expanded rapidly (e.g., Shanghai metro mileage grew from 62 km to 452 km)	Metro became backbone of urban mobility
2012-2018	Policy shift to integrate rail and surface transport	Coordinated, multi-level transport development
2020S	Metro + bus integration in major cities (Beijing, Shanghai, Guangzhou, Chongqing)	Towards smart, sustainable mobility systems

To better summarise the evolution of China's urban public transport system, Table 1.2 provides a concise overview of key milestones and their implications.

Currently, all major cities in China are actively pursuing the joint development of rail transit and public transport, aiming to achieve synergy between these two key transport networks. In Beijing, the coordination between surface transport lines and the metro network has been improved by adjusting routes on major urban roads to enhance transport efficiency. Guangzhou has made two large-scale adjustments to its surface transport network to optimise existing bus routes in response to the continued growth of metro capacity. Shanghai has adopted a strategy of zoning and functionally optimising its surface bus network, establishing a multi-level operational structure from the inner city to the outer areas. Chongqing has added several intermediate bus stops, limiting the distance between 69% of surface bus stops and metro stations to 500 metres (Zhou, Li, Zhao, Feng, and Yao, 2022).

These measures undoubtedly help to alleviate urban traffic pressure, enhance travel convenience, and promote the efficient operation of urban transport networks. These efforts are crucial not only for mitigating traffic issues but also for providing residents with more convenient and efficient travel options. With advancements in technology and updates in management concepts, the future of public transport in China is expected to witness broader development opportunities, contributing to the construction of harmonious and liveable modern cities.

Table 1.3 shows the rankings of major Chinese cities by population and subway line length (Transport Release, 2021).

Table 1.3

Ranking of Large Cities, Population and Length of Rail Lines in China

	City	Population (10,000)	Urban Resident Population (10,000)	Rail Transit	Length of line network (KM)
1	Shanghai	2487	1987	20	831
2	Beijing	2189	1775	27	783
3	Shenzhen	1749	1744	12	289
4	Chongqing	3205	1634	10	264
5	Guangzhou	1868	1488	16	303
6	Chengdu	2094	1334	13	518.3
7	Tianjin	1387	1093	8	269.7
8	Wuhan	1245	995	11	282
9	Dongguan	1047	956	1	37.8
10	Xi'an	1218	928	8	258
11	Hangzhou	1194	874	12	516
12	Foshan	950	854	2	70.3
13	Nanjing	931	791	11	193
14	Shenyang	907	707	4	116.5
15	Qingdao	1007	601	6	279.4
16	Jinan	920	588	3	84.23
17	Changsha	1005	555	6	209
18	Harbin	1001	550	3	79.4
19	Zhengzhou	1260	534	8	232.5
20	Kunming	846	534	6	184
21	Dalian	745	521	5	215.6

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

It is evident that, like many major cities globally, China is experiencing traffic congestion as it undergoes urbanisation. This phase encompasses the construction and development of rail transit along with the coordinated development of rail and surface transport. These efforts have bolstered the development of public transport systems in major cities, providing greater convenience for citizens. However, surface transport still faces several challenges, such as capacity and integration with rail transit, which remain unresolved. In past empirical efforts, the following challenges were observed.

1.2.1 Overemphasis on Prioritising Surface Transport

On Jingshi Road, a main thoroughfare in Jinan, Shandong Province, due to congestion, it takes about half an hour to travel a distance of approximately 100 metres during the morning rush hour. In 2003, this road was widened to five lanes in each direction but still could not cope with the rapidly increasing number of private cars in the area. In 2016, the local government decided to allocate two lanes in each direction exclusively for buses (Sina Shandong (Ed.), 2016). However, this decision did not optimise public transport until the promotion of the local subway, which significantly alleviated congestion on Jingshi Road (Sina Finance (Ed.), 2019). Some city policymakers have overly focused on surface transport, neglecting the overall development of the local transport network. The excessive emphasis on urban surface transport at the expense of rail transit has turned surface transport into a major cause of traffic congestion. A continuous decline in surface transport's passenger volume is evident. Since 2014, the average daily passenger volume of rail transit has surpassed that of surface transport, becoming the main mode of urban passenger transport. As of August 2023, rail transit accounted for 74.65% of Shanghai’s public passenger transport, while surface buses accounted for 20.13% (Shanghai Statistical Yearbook, 2023). This is not an isolated example in China’s urbanisation process but a common phenomenon.

Table 1.4

Top 10 Ranking of Public Transport Passenger Volume in Chinese Provinces and Cities from January to August 2023

Rank	Province or City	Total passenger volume from January to August (10000 people)	Proportion of rail transit (%)	The proportion of public buses and trams (%)	The proportion of cruising taxis (%)	Region
1	Guangdong	657403	60.95	31.65	7.22	South China
2	Sichuan	424591	36.03	41.37	22.60	Southwest China
3	Beijing	372980	63.38	32.77	3.85	North China
4	Shanghai	326108	74.65	20.13	4.57	East China
5	Liaoning	305595	16.76	42.43	40.82	Northeast China
6	Jiangsu	304230	44.61	42.44	12.87	East China
7	Zhejiang	281706	42.99	44.93	12.07	East China
8	Hubei	269428	33.55	43.80	22.42	Central China

Rank	Province or City	Total passenger volume from January to August (10000 people)	Proportion of rail transit (%)	The proportion of public buses and trams (%)	The proportion of cruising taxis (%)	Region
9	Chongqing	265589	34.44	45.68	19.84	Southwest China
10	Shandong	262081	16.36	65.12	18.51	East China

In Shenzhen, the average daily surface transport passenger volume decreased by 15% in 2023, and in Beijing, the average daily surface bus passenger volume in 2019 dropped by 33.5% compared to 2010. With the further advancement of rail transit construction in China, the proportion of surface transport is bound to decrease further. Currently, the function of surface transport is no longer just simple travel but also needs to connect with rail transit (Yu, 2022).

From the table1.4 show, above ranking of passenger volume in various inland provinces and cities, it is evident that rail transit systems play a crucial role in many major and megacities. In economically developed regions such as Shanghai, Beijing, and Guangdong, more than 60% of passenger volume is handled by rail transit. Particularly in Shanghai, this figure reaches an impressive 74.65%, highlighting the core importance and significance of rail transit in these megacities.

In contrast, the situation in Shandong Province is notably different. The province’s passenger volume mainly relies on public electric buses, accounting for 65.12%, while rail transit accounts for only 16.36%. This phenomenon reflects Shandong Province’s differentiated transport needs, which may be due to geographical factors or varying levels of economic development. Similarly, the situation in Liaoning Province is also noteworthy. The proportion of passenger volume from rail transit is only 16.76%, which might be related to its unique geographical location and economic structure. However, it is important to note that in Liaoning Province, public electric buses and taxis have become the main travel choices, accounting for 42.53% and 40.83%, respectively.

These figures reveal that nearly half of the city's transport capacity is taken up by taxis, which not only exacerbate urban traffic congestion but also contribute to environmental pollution to some extent. This imbalanced transport model needs to be adjusted and optimised to achieve more sustainable transport development goals.

1.2.2. The Integration Level between Surface Transport and Rail Transit

The integration level between surface transport and rail transit is relatively low. As rail transit systems are constructed in major cities, new challenges have arisen for surface buses. Since rail transit stations are typically situated along key public transport routes, passenger volumes for surface transport on these routes drop significantly once rail transit operations begin. Therefore, the integration level between surface transport and rail transit has become a significant concern. Although both systems aim to alleviate traffic pressure, they differ in their functional roles and target demographics.

In some major cities, as the rail transit network continues to expand and improve, surface bus operations encounter new challenges. Firstly, rail transit stations are often situated in areas with high public transport demand and dense routes, necessitating a reduction in surface bus frequency on these routes. Consequently, during peak times, passengers may need to transfer to other modes of transport or wait longer to reach their destinations. Secondly, the introduction of rail transit places higher demands on the planning and management of the surface transport system. Ensuring the efficient operation of rail transit while maintaining the competitiveness and service quality of surface buses becomes a crucial issue for urban planners to address.

Furthermore, as metro lines extend, land use planning must be adjusted to accommodate new transport demand patterns, inevitably increasing the complexity of urban construction and management. Therefore, although rail transit brings convenience and efficiency to urban transport, effectively integrating it with surface transport remains a complex and evolving process. True transport integration can only be achieved through continuous optimisation of coordination mechanisms between the two, providing citizens with more convenient and efficient travel options.

Taking Nanjing as an example, following the operation of three metro lines, public transport along the metro corridors experienced a 12.2% decline within a year. Less than 15% of public transport routes had passenger volumes exceeding 50%, yet no significant adjustments were made to the public transport routes. This led to overlaps between metro and surface routes, which, while allowing the public freedom to choose their mode of travel, represents a waste of public resources from a city management perspective. Additionally, only 13.6% of surface transport routes in Nanjing connect to metro stations, causing inconvenience for passengers who need to transfer.

A similar trend has been observed in Shanghai. Between 2000 and 2016, the passenger volumes for both rail transit and surface buses continued to grow. However, starting in 2017, despite the opening of new lines, the passenger volumes for both systems began to decline (Xue and Liu, 2019).

Additionally, there are some shortcomings in the research on metro and surface transport.

Insufficient Emphasis on Real-World Data: Current research does not sufficiently emphasize real-world data. Over the past five years, most doctoral dissertations on rail transit and surface transport found in Google Scholar and CNKI have primarily relied on mathematical model construction and virtual verification as their main methods (Luo, 2020; Huang, 2020). Few researchers have integrated large amounts of public transport data with the optimisation of surface and rail transit predictions. Additionally, existing studies utilizing big data to analyze urban transport systems tend to focus on single transport systems rather than the integration of rail and surface transport (Lu, Liu, Zhou, and Han, 2020).

Lack of Focus on Passenger Travel Experience: Insufficient attention is given to the passenger travel experience in the process of optimizing public transport. When evaluating public transport optimisation results, the passenger travel experience is also important. This experience is influenced by factors such as walking distance and time from home to the metro station and from the metro station to the surface bus, as well as the availability of infrastructure to cope with unexpected weather changes at transfer stations. There is also the issue of waiting time: whether the surface bus system can accurately adjust schedules during peak times to manage the constantly changing crowds at metro stations, and conversely, whether metro stations can predict the corresponding passenger flows.

The above discussion indicates that existing efforts and research on optimizing public transport in major Chinese cities lack comprehensiveness. Therefore, it is necessary to conduct a more empirical study , in this case through a structured questionnaire survey administered to Shanghai metro passengers. The survey approach was selected as it enables the collection of quantitative data from a large sample, ensuring statistical representativeness while complementing passenger flow records.Incorporating big data as a tool in the research is essential to more accurately identify the key issues in the integration of metro and public transport, and to comprehensively evaluate potential optimisation methods based on real-world data. Secondly, this study will use rail transit as a starting point to explore passenger flow predictions along metro lines in major Chinese cities, as well as the setting of surface bus routes and stations. This approach, grounded in the actual conditions of the Chinese public transport system, will facilitate more effective solutions to the current challenges facing public transport.

1.3 Problem Statement

In China's urban development, the metro system plays a crucial role not only as a convenient mode of travel but also as a key driver of urban quality of life and sustainability (Wang & Xia, 2024). With the rapid pace of economic growth and urbanisation, metro systems face increasing pressure to operate efficiently while providing high-quality services to millions of passengers daily. To achieve this, transport authorities and researchers must rely on data-driven approaches that enable a deeper understanding of passenger behaviour, accurate flow prediction, and effective system optimisation.

Firstly, although Shanghai Metro has accumulated massive amounts of passenger data from sources such as smart cards, ticketing systems, and passenger surveys, these datasets are heterogeneous, highly complex, and often fragmented (Yu, Chang, Hu, Yin, & Wu, 2024). Extracting valuable behavioural patterns—such as travel times, number of transfers, trip origins and destinations, and group differences (e.g., super-users vs. occasional riders)—is challenging due to the sheer volume and multidimensional nature of the data. Without advanced analytical approaches, much of this behavioural insight remains underutilised.

Secondly, in the face of increasingly complex urban traffic demand, the ability to accurately predict passenger flows has become essential for maintaining operational stability, particularly during peak hours and unexpected disruptions (Xu, Zhang, Mi, & Wang, 2024). Existing prediction models are often restricted to single-variable or unidimensional analyses and fail to comprehensively incorporate temporal, spatial, and behavioural dimensions. This lack of integration reduces predictive accuracy and limits the practical applicability of such models (Li, Guo, Li, Wang, Gandomi, & Chen, 2023).

Finally, while questionnaire surveys and flow analyses provide insights into passenger satisfaction and service quality (Wang & Gao, 2022), current optimisation strategies remain insufficiently linked to predictive modelling outcomes. There is still a lack of systematic frameworks that translate behavioural insights and predictive results into actionable operational strategies. This gap highlights the need for advanced AI-based approaches that can integrate heterogeneous data, capture multi-dimensional passenger behaviour, and support proactive optimisation strategies tailored to the Shanghai Metro system.

1.4 Research Objective

In order to solve the problem, the following research objectives are formulated in this study.

i. To extract passenger travel behaviour patterns from transport passenger data through data mining techniques.

ii. To develop a prediction model is for predicting future passenger flows.thus helping to formulate more reasonable and effective traffic operation strategies.

iii. To improve the overall operational efficiency and passenger experience of the big city metro system.

1.5 Research Questions

Based on the above research objectives, this paper proposes the main research questions of this study.

i What factors influence the emergence of different passenger travel behaviour patterns in the Shanghai metro system?

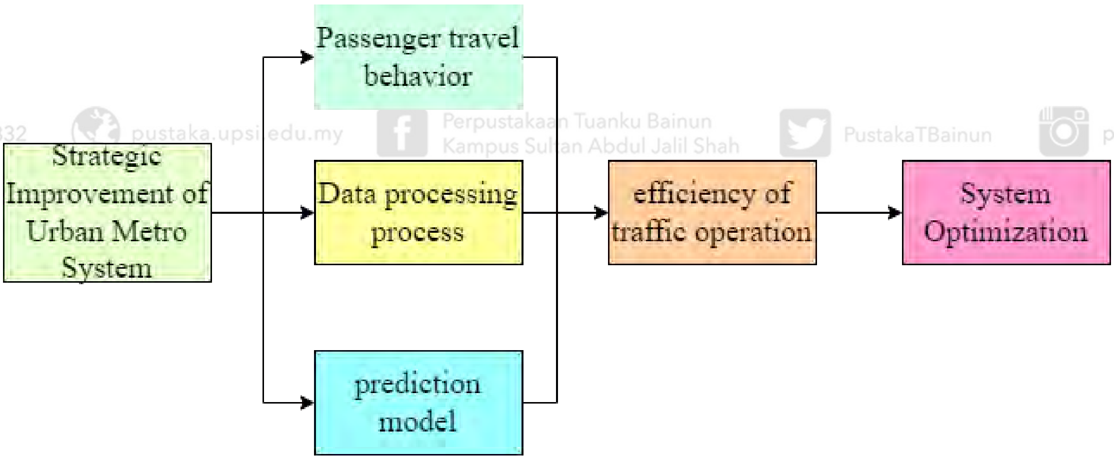
ii *How accurately can the CaGT model predict passenger flows compared with existing baseline models (e.g., LSTM, GAT, Transformer)?*

iii. How can the results of passenger behaviour analysis, prediction models, and questionnaire feedback be utilized to propose feasible optimisation strategies for the Shanghai metro system to improve operational efficiency and passenger experience?

1.6 Conceptual Research Framework

Figure 1.1

Conceptual research framework

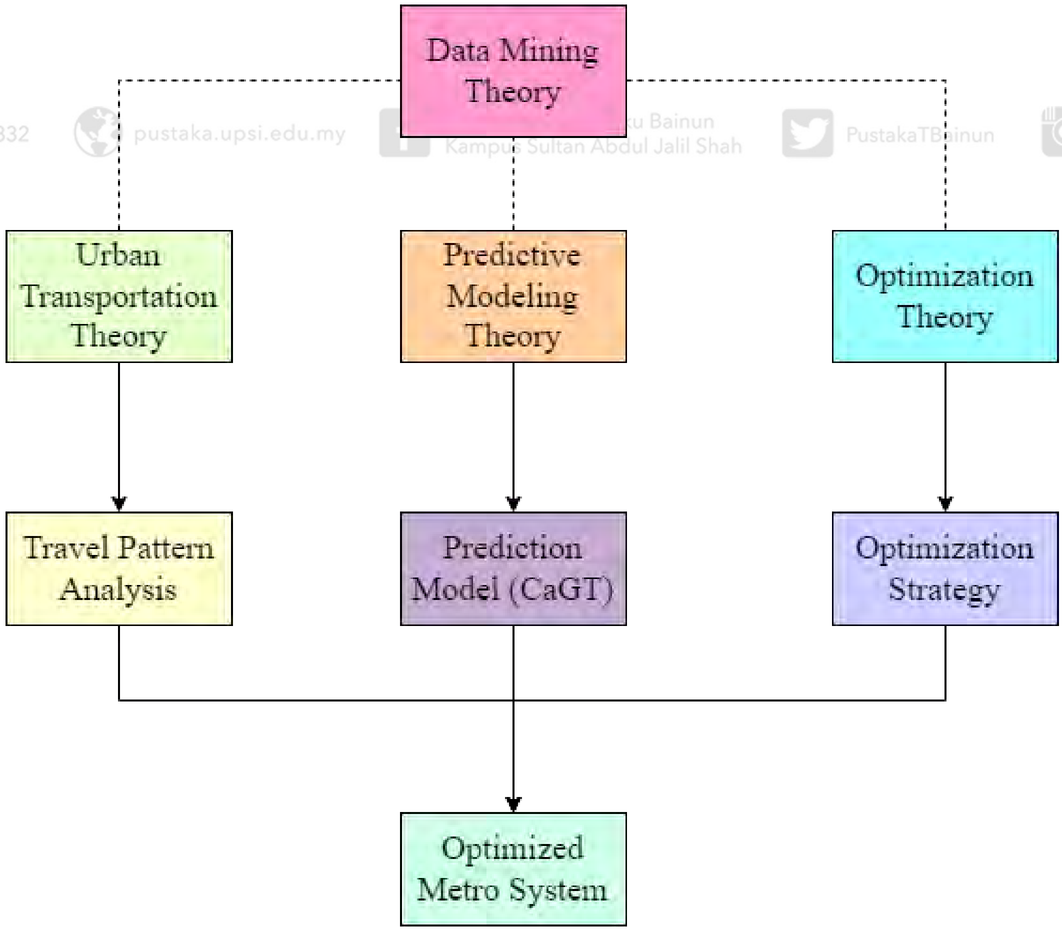


The diagram presents a framework for the strategic improvement of the urban subway system by focusing on interconnected components. It begins with an understanding of passenger travel behavior, which helps identify user patterns and preferences. This information feeds into the data processing process, where collected data is analyzed to generate actionable insights. By utilizing a prediction model, future demands and system performance can be forecasted effectively. The framework also emphasizes improving the efficiency of traffic operation to ensure smoother and more reliable transit services. Ultimately, all these components work together to achieve system optimisation, leading to a well-organized and enhanced subway system that meets the needs of passengers while improving overall operational performance.

1.7 Theoretical research framework

Figure 1.2

Theoretical research framework



The diagram outlines a structured approach to achieving an Optimized Metro System through interconnected theories, analyses, and strategies. It starts with foundational concepts such as Data Mining Theory, Urban Transportation Theory, and Predictive Modeling Theory. These theories support the processes of Travel Pattern Analysis and the development of a Prediction Model (CaGT), which enable an understanding of passenger behaviors and system performance forecasts. The integration of Optimisation Theory and Optimisation Strategy ensures that operational efficiency is improved. By combining these elements, the framework leads to a well-optimized metro system that enhances performance and meets urban transportation needs effectively.

1.8 Scope of the Study

The metro operation data of Shanghai will be selected for analysis in this study. Firstly, the reasons for choosing Shanghai for this study are as follows:

i. Shanghai's urban resident population, which ranks first among all major cities in China (Chongqing urban resident population of 22,874,500) (Chongqing Statistical Yearbook, 2023).

ii. The number and length of surface bus routes operating in Shanghai is the highest in the country, reaching 1,576 routes and 24,479 kilometres by the end of 2023 (Shanghai Statistical Yearbook, 2023).

iii. The number and length of railways in Shanghai is also among the highest in China, reaching 831 kilometres, with 20 lines and 508 stations, and 83 interchanges by 2023 (Shanghai Statistical Yearbook, 2023).

Taking these three points together, the development of Shanghai's public transport system is representative of China's large cities, and optimising the city's public transport is most urgent.

Secondly, there are several reasons for choosing the whole area rather than a specific metro line in Shanghai.

i. Shanghai's metro lines are often very long and pass through several administrative areas, so if the study chooses a specific metro line it will not produce research results that have practical utility.

ii. Each administrative district in Shanghai usually has multiple rail and surface bus lines, and the commercial and residential areas within the districts are well developed, allowing for a more comprehensive analysis of surface bus operations. The results obtained from such a sample will be more credible.

This study will collect the swipe card data of the rail transit on one hand, and on the other hand, it will use the online questionnaire to obtain the passengers' own suggestions and opinions on the interchange of the commonly used stations, the commonly used routes, the ideal distance and the actual distance of travelling, the travelling experience, and the related other suggestions and opinions.

1.9 Significance of the Study

This study focuses on the strategic improvement of the Shanghai metro system, which is an issue that is orientated towards both academic and practical applications. In previous studies and examples, there is not yet a sufficiently rich set of findings to enable a comprehensive data-driven analysis of passenger experience and metro traffic forecasting as well as metro optimisation strategies.

1.9.1 Academic Value

Expanding the theoretical framework of traffic behaviour analysis: by digging deeper into passengers' travel behaviour patterns (e.g. travel time, number of transfers, starting point, etc.), this study will provide a new theoretical perspective in the field of traffic behaviour analysis. The results of the study will help enrich the existing urban traffic research methods and promote the further development of traffic flow prediction and traffic behaviour analysis.

Enhancing the accuracy of data-driven prediction models: this study is dedicated to constructing a context-aware graph transformer prediction model based on historical traffic data to improve the accuracy of future passenger flow prediction. This will help to fill the gaps in existing research where prediction models do not adequately consider complex traffic patterns, and provide more refined and reliable prediction tools for the field of traffic engineering.

Innovation of system optimisation strategies: By combining the results of the behavioural analysis and prediction models as well as the questionnaire, this study will explore suitable optimisation strategies for the Shanghai metro system. This methodological innovation not only provides new research ideas for academics, but also provides a reference framework for similar studies in the future.

1.9.2 Practical Application Value

Enhancing the refinement of urban traffic management: the passenger behaviour patterns mined in the study can provide urban traffic management with more detailed insights into user behaviour. This will help optimise the scheduling and management of the metro system, improve service quality and reduce congestion during peak hours.

Supporting the development of Intelligent Transport Systems: The results of this study can provide important data support and technical guidance for the development of Intelligent Transport Systems (ITS). By improving the accuracy of traffic prediction, ITS can better cope with peak flows and emergencies, thus enhancing the traffic efficiency of the whole city.

Contribute to the sustainable development of the city: The system optimisation strategies proposed by the study will, to a certain extent, improve the energy utilisation efficiency and operational efficiency of the metro system, thus reducing the negative impacts of urban transport on the environment, and contributing to the sustainable development goals of Shanghai and other large cities.

Provide data support for policy formulation: The research results will not only help the transport department to optimise its operation, but also provide data support for the government to formulate more scientific and reasonable transport policies, enhance the implementation effect of the policies, and promote the overall progress of the urban transport system.

While the preceding overview highlights the historical development and current challenges of China's urban public transport, it is particularly important to situate this study within the context of the Shanghai Metro system. As the world's largest metro network in terms of operating mileage and daily ridership, Shanghai faces unique challenges of congestion, transfer bottlenecks, and fluctuating passenger demand. In this context, artificial intelligence (AI) predictive modelling provides a powerful tool to address these issues. By applying advanced methods such as deep learning and graph-based models, AI enables more accurate short-term forecasting of passenger flows, supports proactive scheduling strategies, and improves service reliability. Explicitly linking the challenges of the Shanghai Metro with the potential of AI predictive models establishes both the rationale and urgency for the present study.

1.10 Operational Definitions

To ensure clarity and consistency, the following key terms are operationally defined for this study:

i Passenger Behaviour

Passenger behaviour refers to measurable patterns of metro users' travel activities. In this study, it includes temporal travel patterns (e.g., peak and off-peak travel times), transfer behaviour (e.g., number of transfers, time spent at interchange stations), trip frequency, origin–destination (OD) pairs, and passenger satisfaction derived from survey responses.

ii Passenger Flow

Passenger flow is defined as the number of passengers entering, exiting, or transferring at a metro station within a specified time interval. In this study, passenger flow is operationalised using smart card transaction records, aggregated at 5-minute, 15-minute, and hourly intervals.

iii Super-Users

Super-users are defined as the top 100 passengers with the highest frequency of metro trips during a 30-day observation period, based on smart card data. This group represents the high-frequency travel segment and is analysed separately to explore behavioural heterogeneity.

iv Prediction Accuracy

Prediction accuracy refers to the extent to which the outputs of the predictive models match the observed passenger flow data. In this study, accuracy is evaluated using statistical indicators, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2).

v Context-Aware Graph Transformer (CaGT) Model

The CaGT model is the predictive framework developed in this research. It integrates Graph Attention Networks (GAT) to capture spatial dependencies across metro stations and Transformer structures to model temporal dependencies. The model is designed to provide more accurate and robust station-level passenger flow forecasts.

vi Optimisation Strategy

Optimisation strategy refers to the set of data-driven recommendations derived from predictive modelling results. In this study, optimisation strategies focus on metro scheduling adjustments, resource allocation, transfer efficiency improvement, and measures to enhance passenger experience.

vii Satisfaction

Satisfaction is defined as passengers' perceived service quality when using the metro system. It is measured through a structured questionnaire survey based on the

SERVQUAL framework, including dimensions such as reliability, comfort, safety, and convenience.

viii Data Sources

Data sources in this study include:

Smart card data (recording entry/exit times, station IDs, and trip frequency);

Passenger survey data (capturing perceptions of service quality and satisfaction);

Operational records from the Shanghai Metro system (providing system-level statistics and context).

1.11 Research Organisation

This study is divided into five chapters. These five chapters are introduction, literature review, research methodology, data analysis and experimental research and finally discussion and conclusion.

Chapter 1 introduces the whole research process. The chapter firstly discusses the importance of China's public transport system in cities, especially in big cities, and secondly, discusses the dilemmas encountered by the rail transport system in the current urban development. Then the problems encountered and being solved in the Shanghai metro are introduced from the research background to the problem statement, and the chapter builds on the existing research content to construct the preliminary research objectives and research questions. The importance of the study, the scope of the study and the significance of the study are further emphasised.

The second chapter of the report focuses on the existing literature. It first discusses big data and passenger experience, which are elaborated from the role of big data in public transport, to data extraction and mining, to the role of passenger experience in public transport, then discusses traffic flow prediction in public transport, which is carried out from temporal traffic flow prediction to spatio-temporal traffic flow prediction, and lastly discusses the optimisation of public transport, which is carried out from surface transport to the synergistic optimisation of surface and rail transport. Finally, this chapter links all the concepts to form the research model for this study.

Chapter 3 presents the methodology used to conduct this study. It introduces the research site and the network of the metro system from the data context of the study, then proceeds to the discovery of passenger travel behaviour, which contains the presentation of raw data, the analysis of passenger flows at metro stations and the analysis of passenger travel patterns, then develops a passenger prediction model, elaborates on the model by presenting the various modules of the model, and finally proposes the optimisation challenges and the strategic approach for the metro system.

Chapter 4 presents the data analysis and experimental study. The behavioural results of passenger travel are analysed in terms of travel time and number of trips, transfer analysis and origin analysis. Then the experimental study of the prediction model of passenger flow is carried out, the basic model of the comparison experiment is introduced, as well as the relevant parameters, and then a comparison is carried out to conclude that the model we designed is better than the comparison model to achieve the design purpose. Finally, the metro system questionnaire analysis and optimisationis presented, where the satisfaction questionnaire is analysed and strategic suggestions and methods for optimisation are given.

Finally, Chapter 5 focuses on the results of the study and its discussion. We discuss whether the research objectives stated in Chapter 1 were accomplished or not, and also critically discuss the findings of this study through relevant theories and past studies discussed in the literature review. New innovations or novel ideas are then presented based on the findings, highlighting the limitations of this study and suggestions for future research. The last part of this paper summarises and concludes the study.

The following is the research flowchart of this paper, as shown in Figure 1.3:

Figure 1.3

Flowchart of the research