Disruption analytics in urban metro systems with large-scale automated data

Nan Zhang

A thesis submitted as fulfilment of the requirements for degree of
Doctor of Philosophy of Imperial College London

Centre for Transport Studies
Department of Civil and Environmental Engineering
Imperial College London, United Kingdom

February 2022

Abstract

Urban metro systems are frequently affected by disruptions such as infrastructure malfunctions, rolling stock breakdowns and accidents. Such disruptions give rise to delays, congestion and inconvenience for public transport users, which in turn, lead to a wider range of negative impacts on the social economy and wellbeing. This PhD thesis aims to improve our understanding of disruption impacts and improve the ability of metro operators to detect and manage disruptions by using large-scale automated data.

The crucial precondition of any disruption analytics is to have accurate information about the location, occurrence time, duration and propagation of disruptions. In pursuit of this goal, the thesis develops statistical models to detect disruptions via deviations in trains' headways relative to their regular services. Our method is a unique contribution in the sense that it is based on automated vehicle location data (data-driven) and the probabilistic framework is effective to detect any type of service interruptions, including minor delays that last just a few minutes. As an important research outcome, the thesis delivers novel analyses of the propagation progress of disruptions along metro lines, thus enabling us to distinguish primary and secondary disruptions as well as recovery interventions performed by operators.

The other part of the thesis provides new insights for quantifying disruption impacts and measuring metro vulnerability. One of our key messages is that in metro systems there are factors influencing both the occurrence of disruptions and their outcomes. With such confounding factors, we show that causal inference is a powerful tool to estimate unbiased impacts on passenger demand and journey time, which is also capable of quantifying the spatial-temporal propagation of disruption impacts within metro networks. The causal inference approaches are applied to empirical studies based on the Hong Kong Mass Transit Railway (MTR). Our conclusions can assist researchers and practitioners in two applications: (i) the evaluation of metro performance such as service reliability, system vulnerability and resilience, and (ii) the management of future disruptions.

Acknowledgements

I wish to express my sincerest gratitude to all of those who supported and assisted me in the development of this thesis. First, I would like to thank my supervisor, Professor Daniel J. Graham, for his supportive guidance and invaluable advice throughout my PhD research at Imperial College London. I would also like to thank my co-supervisors, Dr Daniel Hörcher and Dr Prateek Bansal, for their continuous help, patient guidance and encouragement in every step of my studies.

I am grateful to the financial support provided by the Transport Strategy Centre (TSC) at Imperial College London during my doctoral studies. Also, I am thankful to Hong Kong MTR Corporation Ltd for providing the large-scale automated data used in this research work. I would like to thank Sarah Willis from the Department of Civil and Environmental Engineering, who have helped substantially in the administration related issues of this PhD.

Further, I would like to thank Dr Jose M. Carbo, Dr Laila Ait Bihi Ouali and Dr Ramandeep Singh for insightful discussion and useful advice, especially during the early stage of my PhD. I extremely appreciate the help from Dr Ramandeep Singh in respect to the processing of large-scale datasets at the beginning of this PhD research, and her kind assistance throughout my study. My very grateful thanks are also extended to the PhD students and postdocs of our research group and the Centre for Transport Studies, for sharing ideas and thoughts in the research life. In random order, special thanks to Ana, Anupriya, Csaba, Farah, Joris, Keita, Liang, Praj, Saeed, Shane, Lin, Chenyang, Shiming, Heather, LinTong and many others.

I am immensely grateful for the endless love and support of my parents; without them, this journey would have been impossible. Last but not least, I would like to thank my husband for his love, patience, sacrifices and supports all the time.

Declaration

I hereby declare that all materials in this thesis is my own work. Information derived from the published and unpublished work of others has been acknowledged in the text and references are given in the list of sources.

Nan Zhang

Contents

Abstract 2

Acknowledgements and Declaration 3

List of Tables and Figures 8

Nomenclature 11

Chapter 1 Introduction 12

1.1 Background 12

1.2 Aims and objectives 16

1.3 Structure of thesis 18

1.4 Contributions 18

1.5 Publications 20

Chapter 2 Literature review 21

2.1 An introduction to metro disruptions 21

2.1.1 Definition of disruptions 21

2.1.2 Factors affecting disruption occurrence 21

2.2 Large-scale automated data in public transport research 22

2.2.1 Smart card data 22

2.2.2 Automatic vehicle location data 24

2.3 Metro disruption detection 25

2.3.1 Review of anomaly detection algorithms 25

2.3.2 Research on traffic anomaly detection 27

2.3.3 Research on metro disruption detection 27

2.4 Metro vulnerability measurement 29

2.4.1 Definition of vulnerability 29

2.4.2 Methods of vulnerability measurement 30

2.5 Disruption impact estimation 32

2.5.1 Simulation-based research 32

2.5.2 Empirical research 33

2.6 Research gaps 38

Chapter 3 Methodological background: Causal inference methods 40
3.1 The basic framework 41
3.1.1 Potential outcomes 41
3.1.2 The assignment mechanism 42
3.2 Challenges and viable methods for disruption impact analysis 43
3.2.1 Treating confoundedness 43
3.2.2 Treating interference 45

Chapter 4 Detecting metro service disruptions via large-scale vehicle location data 49
4.1 Introduction 49
4.2 Methodology 53
4.2.1 Probabilistic detection with Gaussian mixture models 54
4.2.2 Secondary disruption and recovery intervention identification 60
4.3 Data and case study 62
4.4 Results and discussion 65
4.4.1 Input data check: screening potential disruptions (Type II) 65
4.4.2 Optimal number of GMM clusters and probability threshold 67
4.4.3 GMM detection results 70
4.4.4 Secondary disruptions and recovery interventions 72
4.5 Conclusions and future work 74

Chapter 5 A causal inference approach to measure the vulnerability of urban metro systems 77
5.1 Introduction 77
5.2 Methodology 80
5.2.1 Causal inference method to estimate disruption impacts 81
5.2.2 Constructing vulnerability metrics 87
5.2.3 Imputing missing vulnerability metrics 88
5.3 Data and case study 90
5.4 Results and discussions 98
5.4.1 Propensity score models 98
5.4.2 Matching results 100
5.4.3 Imputation of missing vulnerability metrics 101
5.4.4 The MTR insights 102
5.5 Conclusions and future work 108

Chapter 6 Quantifying the direct and spillover effects of disruptions in urban metro networks 111

6.1 Introduction 111

6.2 Methodology 113

6.2.1 The modified synthetic control method 113

6.2.2 The choice of weights 116

6.3 Data and case study 118

6.4 Results and discussion 119

6.4.1 Modified synthetic control design 119

6.4.2 Synthetic control results of the disrupted station 121

6.4.3 Spillover disruption effects and propagation 125

6.5 Conclusions and future work 128

Chapter 7 Conclusions 130

7.1 Main findings and contributions 130

7.2 Potential applications 132

7.3 Future research 135

References 137

Appendix A Supplementary Material: Chapter 4 157

A.1 Robustness check against percentile of headway deviations in simulation 157

A.2 Sensitivity analysis of detection accuracy 158

Appendix B Supplementary Material: Chapter 5 160

B.1 Balance improvements under different matching methods 160

Appendix C Supplementary Material: Chapter 6 161

C.1 Estimated spillover effects on average travel speed over time 161

List of Tables

Table 1.1: The corresponding metro performances of the disruption and baseline scenarios in this illustration 15

Table 2.1: Summary of important literature on the general anomaly detection algorithms 25

Table 2.2: A comparison of recent research on metro disruption detection 36

Table 2.3: A comparison of recent research on metro vulnerability..... 37

Table 4.1: The two-way table of simulation performance (averages of 1000 runs) and optimal GMM parameters 68

Table 4.2: A sample of selected disruption records with the corresponding category identification results 72

Table 5.1: Available covariates for PSM (stage 1) and vulnerability imputation (stage 3)..... 92

Table 5.2: Estimation results of the propensity score model (logistic regression) 98

Table 5.3: Comparison of prediction accuracy of different imputation methods 101

Table 5.4: Top 5 vulnerable stations based on demand loss and speed loss vulnerability metrics 107

Table 5.5: Top 5 vulnerable stations based on irregularity in flow vulnerability metrics 107

Table 6.1: Potential predictors of metro performance 121

Table 6.2: Mean values of predictors for average speed measures before the disruption occurred 124

Table 6.3: Mean square prediction errors of four outcome measures before the disruption . 124

Table 6.4: Synthetic control weights of the disrupted station (for four outcome measures) . 125

Table A.1: Robustness check against the choice of deviation percentile in simulation 157

Table A.2: Results of sensitivity analysis for two example stations (1000 runs) 158

Table B.1: Means of confounding factors before and after matching 160

List of Figures

Figure 1.1: The distribution of confounding factors for an example disruption and for three different baseline scenarios. Day_of_week and Time_of_day are two dummy variables 14

Figure 2.1: The interrelationship among vulnerability, reliability, risk and resilience [adapted from Faturechi and Miller-Hooks (2014)] 30

Figure 4.1: The relationship between abnormal demand and service disruption – example of a busy station, MTR. Relative changes in entry and exit ridership are derived by comparison with the average ridership under normal services. 51

Figure 4.2: Flowchart of the chapter's methodological framework 54

Figure 4.3: An illustration of the right-most cluster in GMM 57

Figure 4.4: The procedure of applying the GMM-based detections 59

Figure 4.5: The MTR system map 64

Figure 4.6: The histogram of observed headway deviations for different scheduled headways from the given platform-interval of the example station 66

Figure 4.7: Histogram of overall headway deviations observed from the given platform-interval. Observations below the 95th percentile are used to generate the simulation data 67

Figure 4.8: Histogram of a sample synthetic headway deviations for the given platform-interval. The proportion of the disrupted synthetic deviations is 5% 68

Figure 4.9: Changes in the right-most clusters of the estimated GMM, under different number of clusters (M). Each solid line represents the distribution of the right-most cluster for a given cluster number 70

Figure 4.10: Final detection results of the given platform-interval: probabilities of belonging to the disrupted cluster and the optimal threshold. The purple dots represent the identified disruptions 71

Figure 4.11: Spatial-temporal train movement diagram with detected disruptions and their categories. The propagation process of two primary disruptions 73

Figure 5.1: Flowchart of the chapter's methodological framework 81

Figure 5.2: The four urban lines that we study in the MTR network (highlighted in colour) . 95

Figure 5.3: Station affected areas that are used to calculate the supplementary factors surrounding metro stations. The radius of each circle is 500 meters.....96

Figure 5.4: Examples of the distribution of land use and transport facilities in Hong Kong ..97

Figure 5.5: Histogram of the normalised propensity scores (common support check). Red and green colour represent the control group and disruptions respectively 100

Figure 5.6: Spatial distribution of station-level vulnerability metrics in the MTR (four urban lines). Each dot represents a metro station.....106

Figure 6.1: Schematic overview of the modified synthetic control method for metro disruptions. The donor pool consists of observations from non-disrupted days. ao represent any other station in the network, it can be upstream, downstream or a surrounding station to the disruption. 116

Figure 6.2: Results of synthetic control estimation and causal effects on the disrupted station – with comparison of other impact quantification methods.....123

Figure 6.3: Spillover effects on average travel speed at different time periods. The star symbol indicates the location of the example disruption.....127

Figure C.1: The spillover effects of the example disruption on other 48 stations in the MTR 169

Nomenclature

ATET	Average treatment effect of treated units
AUC	Area under curve
AVL	Automated vehicle location
CBD	Central business districts
CIA	Conditional independence assumption
COV	Coefficient of variation
ED	Euclidean distance
EM	Expectation maximisation
GAM	Generalised additive model
GMM	Gaussian mixture model
HD	Hellinger distance
HK	Hong Kong
IPW	Inverse probability weighted
KL	Kullback–Leibler divergence
MAE	Mean absolute error
MSPE	Mean squared prediction error
MTR	Mass Transit Railway
PSM	Propensity score matching
RAE	Relative absolute error
RCM	Rubin causal model
RSE	Relative squared error
SCD	Smart card data
SUTVA	Stable unit treatment value assumption
XGBoost	Extreme gradient boosting

Chapter 1
Introduction

1.1 Background

With urbanisation on the rise, there is an ever-growing need for members of city populations to use public transport. In serving this growing demand for urban travel, public transport provides people with affordable and sustainable access to essential activities such as employment, education, health care, shopping and recreation. Metros, also known as subways or as rapid transit, have become a vital component of public transport due to their large capacity and high-frequency services. In 2017, 178 metro systems worldwide carried a total of 53,768 million trips (International Union of Public Transport, 2018). One of the main challenges of metro systems are frequently occurred disruptions, especially for those have been operated for over a century or without adequate maintenance. These disruptions are often caused by unpredicted infrastructure malfunctions (e.g., signal failures and track blockage), rolling stock breakdowns and accidents, planned maintenance work, and temporal dispatching adjustments (Jespersen-Groth et al., 2009). Disruptive incidents can cause service delays, crowding and safety issues, which may decrease passenger satisfaction and lead to significant loss of social welfare. For instance, the London Underground encountered 7,973 service disrupting incidents of above 2 minutes duration between April 2016 and April 2017, causing a total loss of around 34 million customer hours (Transport for London, 2017; Transport for London, 2019). Thus, understanding the dynamics of metro disruptions and their corresponding impacts is an important area of research.

Operators need to monitor disruption occurrences closely in order to reduce their detrimental effects. With accurate information on the location, time, duration, and propagation process of disruptions, they can assess the reliability and resilience of metro systems comprehensively. Thus, the detection of service disruptions is a prerequisite of any further research on disruption management. Also, operators may consider investing in new technologies to improve metro facilities and mitigate the effect of incidents. For instance, the New York City Subway was in a state of emergency in June 2017 after a series of derailments,

track fires and overcrowding incidents. The Metropolitan Transportation Authority invested over $8 billion to stabilise and modernise the incident-plagued metro system (Metropolitan Transportation Authority, 2019). It is apparent that metros are willing to invest in their infrastructure systems, but it is often not known how those investments compare in achieving improvements. To facilitate project selection, metros are increasingly relying on disaggregate performance metrics that reveal the most vulnerable parts of the network. Moreover, effective recovery strategies depend on detailed disruption information, such as the affected ridership, delayed time and crowding level in stations or trains. For passengers, knowledge of historical disruption impacts can also help them reschedule their travel plans. Therefore, a comprehensive understanding of disruptions requires us to master the following questions:

i) How to detect when and where disruptions happened and their durations?

ii) How to measure the vulnerability of metro systems under disruptions?

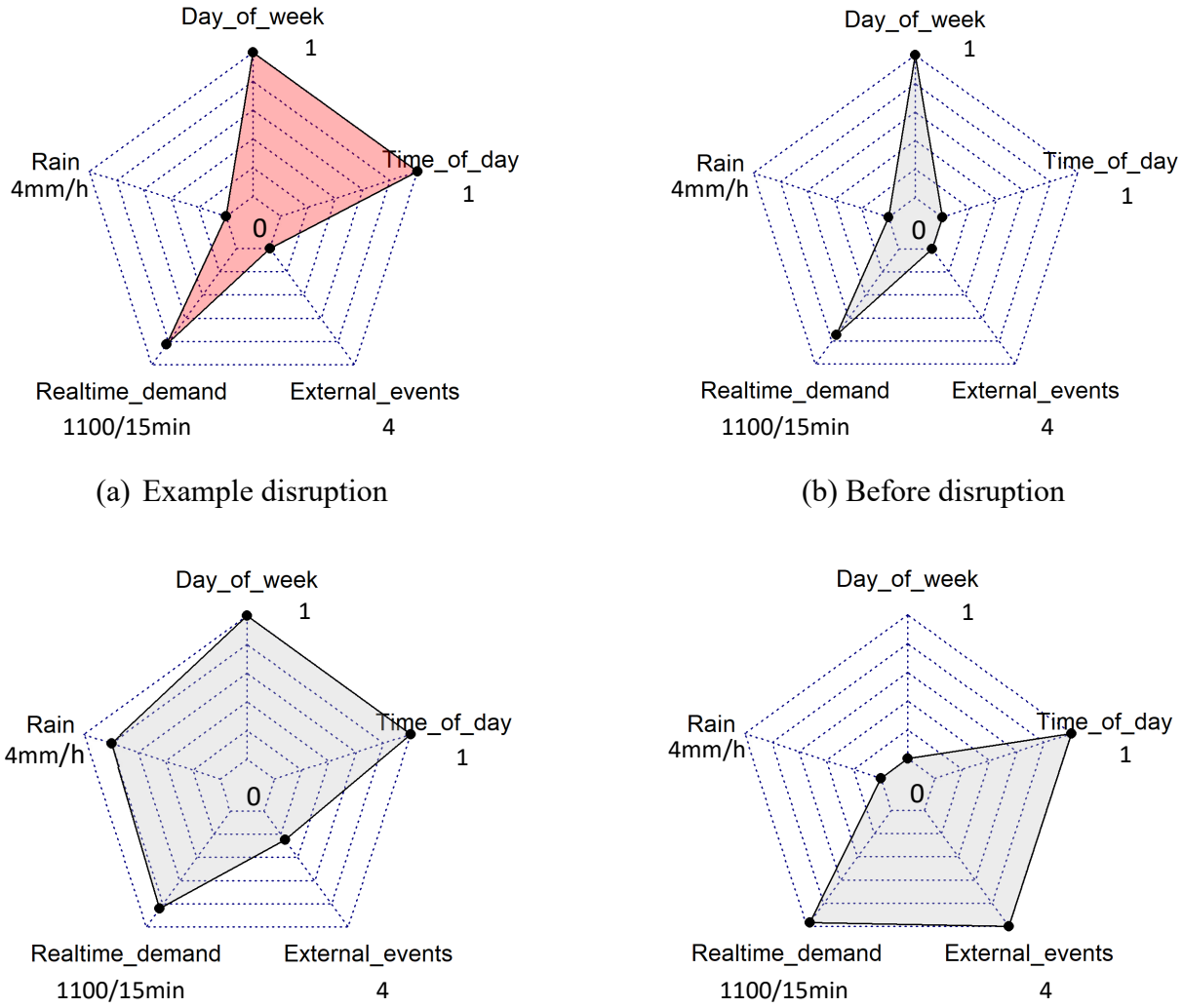
iii) How to quantify the disruption impact on passengers affected?

Meanwhile, how does the disruption impact spread along the metro network spatially and temporally?

It is worth noting that, we specifically focus on service disruptions which are defined as events that interrupt normal train operations for a specific period of time. To distinguish from the broader term "incidents" or "anomalies", the disruptions in this thesis do not include the events unrelated to the interruption of train services. For example, the escalator failure or corridor congestion in metro stations.

In recent years, practices in metro disruption detection have been undertaken, using manual inspections, social media data or smart card data (Sun et al., 2016; Ji et al., 2018; Tonnelier et al., 2018). However, these detection methods either suffer from human errors, limited monitoring ability or inaccurate indication of service interruptions. There is a need to carry out more reliable and comprehensive detections based on additional sources of information about train operations. In metro systems, the growing availability of large-scale datasets provides new possibilities to achieve progress in this area. The large-scale automated vehicle location (AVL) data can act as an ideal source of new information, which captures detailed trajectory data over time for each train on each platform. The train headway extracted from AVL data is a straightforward indicator of service provision, and detection via such indicator can make up for the limitations of previous studies (Tirachini et al., 2021).

In terms of disruption impact and vulnerability measurement, researchers have paid close attention to simulation-based analysis that relies on hypothetical disruption scenarios. Again, the availability of large-scale datasets provides new possibilities for empirical-based analysis. The AVL data and smart card data enable researchers to observe passengers' behaviour under real disruptions, helping avoid unrealistic or oversimplified assumptions. Most of the current empirical analysis applies before-after control-impact methods (Silva et al., 2015; Sun et al., 2016; Yap and Cats, 2020). That is, comparing the disrupted metro system with a baseline scenario without disruptions. However, besides the exposure to disruptions, other factors may also affect the outcomes of service interruptions, such as weather conditions, real-time passenger demand, external mega-events. These factors may even influence the occurrence of disruptions at the same time. We illustrate this phenomenon with the following example.



(a) Example disruption

(b) Before disruption

(c) Non-disrupted scenario 1

(d) Non-disrupted scenario 2

Figure 1.1: The distribution of confounding factors for an example disruption and for three different baseline scenarios. Day_of_week and Time_of_day are two dummy variables

Table 1.1: The corresponding metro performances of the disruption and baseline scenarios in this illustration

Performance measure	Example disruption	Baselines		
		Before disruption	Non-disrupted scenario 1	Non-disrupted scenario 2
Entry ridership /15min	1049	745	1016	1101
Exit ridership /15min	746	553	921	754
Average journey time (minute)	36.71	25.66	26.76	28.07
Average travel speed (km/h)	16.98	22.76	23.18	22.01

Given the example disruption,[1] we compare it with two types of baseline scenarios that are commonly used in before-after control-impact methods: (i) before the disruption occurrence; (ii) the same time and location on a non-disrupted day. In Figure 1.1 and Table 1.1, the potential confounding factors and the corresponding metro performances are displayed separately for each scenario. Figure 1.1 (a) represents the disruption, the remaining plots represent baseline scenarios. Figures 1.1(b) to 1.1 (d) show that although under well-designed baselines, the difference in the distribution of the aforementioned confounding factors can be huge. These differences are then revealed by the heterogeneous performance measures in Table 1.1. Thus, the conventional before-after studies failed to fully control such confounding issues, which can lead to biased impact estimations.

Causal inference[2] is a powerful tool to address the above concerns of confoundedness. This approach aims to quantify the causality between treatments (or interventions) and outcomes of interest. It has been widely studied in the statistics and econometrics literature, and has been widely applied in social and biomedical sciences (Imbens and Rubin, 2015). A fundamental notion underlying causal inference is the potential outcomes: pairs of outcomes defined for the same study unit given different exposure to the treatment, with only one being possibly observed. Rubin (1973a, b, 1974, 1977, 1990) proposed the interpretation of causal statements as comparisons of potential outcomes. This comparison is feasible when the

[1] The example disruption is a real case observed on the Island Line in HK MTR.

[2] In this thesis, causal inference is referred to the Rubin causal model (Rubin, 1973a, b, 1974, 1977, 1990), which is the dominant approach in modern causal analysis, rather than the graphical-based approaches (Pearl, 2000).

counterfactual potential outcome of the treated units can be approximated by the observed outcomes of control units. In observational studies with confounding problems, the biggest attraction of causal inference is that, following some form of assumptions, it can adjust treatment and control groups for differences in observed covariates, or pre-treatment variables, thus remove all biases in comparisons between treated and control units. Despite the superiority of causal modelling, however, such techniques have not yet been used in metro vulnerability analysis, especially the evaluation of disruption impacts.

Therefore, in the thesis, with the growing large-scale datasets and continued advancements in statistical/econometric methodologies, we advance the existing literature and substantially improve the comprehension of the occurrence and impacts of metro disruptions.

1.2 Aims and objectives

The overall aim of this PhD research is to achieve a comprehensive understanding of metro disruptions in their occurrence, severity and impact, and to apply this understanding to support the evaluation of metro performances. Within this overall aim, there are three main objectives corresponding to the research questions respectively.

i) Detect metro service disruptions via a data-driven probabilistic unsupervised learning approach, with identifying disruption propagation and operator's intervention.

- Based on AVL data, we generate the headway series of train services for each platform of a given line within a given station.
- We detect service disruptions from abnormal (overlong) headways, by adopting probabilistic Gaussian mixture model methods.
- We design semi-synthetic simulations to support (finding optimal parameters) and validate the GMM-based detection.
- We identify primary and secondary disruptions, together with the recovery interventions from metro operators. The relationships between primary and secondary disruptions reveal the spatiotemporal propagation of disruption status.
- Based on the detection results, we build a reliable database to facilitate the following disruption research.

ii) Measure the vulnerability of urban metros based on empirical causal effects of disruptions.

- We relax the tacit assumption of random disruptions, and consider the potential confounding factors in impact quantification, which includes real-time demand, station characteristics and weather conditions.
- We introduce the causal inference framework into the context of metro disruptions, which utilises the benefits of large-scale automated data. We intend to apply propensity score matching to estimate the unbiased causal impacts of disruptions on station performance.
- We construct vulnerability measures for metro systems, based on empirical disruption impacts.

iii) Quantify the direct and spillover causal effects of disruptions, and analyse the impact propagation in urban metros.

- For metro networks, we emphasise (i) the presence of interference among connected or neighbouring stations; (ii) the threat of confounding caused by non-random disruptions.
- Under the causal inference framework, we relax the tacit assumption that metro stations are independent during disruptions and are not affected by disruptions at other stations.
- We apply a novel modified synthetic control method, which fits perfectly with the patterns of metro operation and leverages the benefits of large-scale automated data. This causal inference approach enables the quantification the spillover effects on all non-interrupted stations in the network (indirect disruption impacts).
- We analyse the spatial and temporal propagation of spillover disruption impacts within metro networks.

1.3 Structure of thesis

The rest of the thesis is organised as follows.

Chapter 2 presents a comprehensive review of the data and existing work in the area of metro disruption analysis, including automated large-scale data in public transport, disruption detection techniques, metro vulnerability measurement and impact quantification methods.

Chapter 3 provides a methodological overview of the causal inference literature. We discuss the challenges related to confounding and interference, with a brief review of two viable causal inference methods.

Chapter 4 focuses on disruption detection in metro systems. The proposed GMM-based detection method is tested in the case study of the Mass Transit Railway (MTR), Hong Kong.

Chapter 5 develops measures of metro vulnerability based on empirical disruption impacts. We introduce a causal inference approach for impact estimation. Based on the disruptions detected in Chapter 4, an empirical case study is undertaken on the same metro system.

Chapter 6 quantifies the direct and spillover effects of metro disruptions, via a novel modified synthetic control framework. With the disruption data from Chapter 4, a case study is also carried out to demonstrate the propagation of disruption impacts within the selected metro network.

Chapter 7 summarises the main findings from the above studies, along with recommendations for potential future research.

1.4 Contributions

The main contributions of this thesis can be summarised as follows:

i) **New research questions**

Our research contributes to developing the understanding of new research questions about metro disruption. In the detection chapter, we broaden the research scope to identify secondary disruptions (line-level disruption propagation), and to detect the recovery interventions of metro operators. For the impact estimation, we raise attention to the spillover effects for metro disruptions, which is of the same importance as direct impacts but has not been explored empirically.

ii) **Application of advanced econometric and machine learning methods**

For disruption detection, we apply the probabilistic mixture models (unsupervised clustering) instead of deterministic detection algorithms. Wherever applicable, we dynamically learn the parameters of the detection model from semi-synthetic simulations, rather than subjectively determining parameters. The data-driven detection method is effective for any type of disruptions including minor interruptions of few minutes.

For disruption impact estimation, non-randomness of disruption occurrence and interference among stations have been ignored in past studies. Before and after comparisons fail to control for confounding from non-random disruptions and account for the spillover causal effects. To adjust for these potential sources of bias, we use advanced econometric methods under the Rubin causal framework, which are introduced in the literature of metro disruption analysis for the first time.

iii) **Analysis of unique and new sources of data**

For the detection of metro disruptions, we use high-quality AVL data provided by the Hong Kong MTR. To the best of our knowledge, this type of automated data has not been used in previous disruption detections in metro systems.[3] For the metro vulnerability measurement, along with the use of smart card data, we collect the unique land-use, demographic and transport facility information of Planning Scheme zones in Hong Kong and hourly weather data, to provide new sources for confounding factors, and to support the missing metrics imputation. For the research of direct and spillover disruption impacts, additionally, we include a unique citywide mega-events data, which has not been included in previous impact analysis. Such data enables the comparison of the external environment between disrupted and control units, leading to more accurate impact estimates than previously possible.

iv) **Integrated research design**

[3] In other sectors of public transport, such as bus and tram, the AVL data have been used to detect service delays, disruptions and abnormal routes.

The thesis is centred on metro disruptions, with three main research focuses on disruption detection, disruption impacts and their role in metro vulnerability, respectively. The outputs of the detections are ideal input for impact quantifications. The vulnerability measurement can then be regarded as an application of the estimated disruption impacts. This integrated research design enables a more holistic and in-depth understanding of disruptions in urban metro systems.

1.5 Publications

The core methodology elaborated in Chapters 5 is published as part of the following journal article. Early results of this research have been shared with the wider research community through the following conference papers and presentations.

1) Zhang, N., Graham, D.J., Hörcher, D. and Bansal, P., 2021. A causal inference approach to measure the vulnerability of urban metro systems. *Transportation*, pp.1-32.

2) Zhang, N., Graham, D.J. and Carbo, J.M., 2018. Using smart card data to analyse the disruption impacts on urban metro systems. *7th Symposium of the European Association for Research in Transportation,* Athens, Greece.

3) Zhang, N., Graham, D.J. and Carbo, J.M., 2019. Using smart card data to analyse the disruption impacts on urban metro systems. *Transportation Research Board 98th Annual Meeting*, No. 19-03762, Washington DC.

The remaining chapters of this thesis are planned to be published as well. Based on Chapter 4, the paper named *'Detecting metro service disruptions via large-scale vehicle location data'* has been submitted to Transportation Research Part C. Meanwhile, based on Chapter 6, the paper named *'Quantifying the direct and spillover effects of disruptions in urban metro networks'* is under development, and it will be submitted in the near future.

Chapter 2
Literature review

2.1 An introduction to metro disruptions

2.1.1 Definition of disruptions

The concept of disruption has been defined in a number of different ways depending on the context. Generally, disruption refers to an interruption in the usual way that a system, process, or event works.

In urban metros, disruptions are commonly accepted as events that interrupt the normal service of the system. The notion of 'service' covers the usage of station facilities to the provision of train services. Therefore, disruptions can originate from a wide range of factors such as access to station facilities, signal failures, rolling stock blockage, screen doors, passenger or driver related behaviours, weather conditions, emergencies (fire or malicious attack) and engineering works (Zhang et al., 2016; Yap and Cats., 2020).

In practice, the scope of metro disruptions can range from single stations, multiple stations, line segments and to the entire network. The duration of disruptions ranges from a few minutes to several days. The presence of disruption threatens the reliability and robustness of metro systems and therefore need to be identified or detected promptly and accurately.

2.1.2 Factors affecting disruption occurrence

Melo et al. (2011) first analysed the factors that influence the number of incidents across metro lines in different areas. They built Poisson regression and negative binomial regression models to represent the relationship between the expected number of incidents and the possible determinants, finding that factors such as passenger demand, signalling type, line or station age, train operation type and rolling stock characteristics can significantly influence the likelihood of having incidents, which provides us with the basis for selecting potential confounders. Yap and Cats (2020) adopted a supervised learning approach to predict the exposure of disruptions. The predictors include general factors such as season, day of the week

and time of day, which also contain station-related characteristics and historical disruption frequency.

Other determinants related to human errors and adverse weather conditions have also been discussed. Wan et al. (2015) classified the behaviours of metro users and explored their effects on metro operation, through questionnaires of passengers and staff. They established the importance hierarchy of different types of behaviours in relation to incident involvement. In addition, incidents caused by improper driver performance were analysed by Rjabovs and Palacin (2017). They used bivariate correlation analysis to assess the inter-dependency of system design-related factors and driver-related incidents.

With global climate change, severe weather events tend to be increasingly frequent. As a result, some open metro systems may experience more weather-related disruptions. Also, it is important to understand how temperature, rainfall and wind speed affect incident occurrence. Brazil et al. (2017) presented an analysis of the role of weather events, temporal effects and their resulting interactions. For instance, for the Rapid Transit rail system in Dublin, rain is found to be the main cause of disruptions. Significant interactions are found between different weather conditions such as rainfall with wind speed.

2.2 Large-scale automated data in public transport research

2.2.1 Smart card data

The concept of smart card data (SCD) refers to the information collected in automatic fare collection systems that have been widely applied in major urban public transport systems. Electronic ticketing gained popularity because of its obvious advantages versus paper tickets; it is a faster, cheaper, safer and more convenient mean of fare enforcement (Hörcher, 2017). As a side product, the smart card data record the time and location of tap-in/tap-out transactions, trip fares and the related characteristics of card owners, all of which have been regarded as an important information source of passenger's travel behaviour (Bagchi and White, 2005). Therefore, a considerable number of SCD-based studies have been published in the public transport field, which mainly focuses on three areas: strategic level for long-term planning, tactical level for service adjustment, and operational level for network performance assessment

(Pelletier et al., 2011). In this subsection we review SCD studies based on a different typology: system perspective and passenger travel behaviour.

In terms of system planning, smart card data are useful due to the outstanding representativeness of passenger demand across networks. Origin-destination matrices, journey time reliability metrics and transport capacity rate are mined from smart card data to support system planning issues (Chan, 2007; Zhang et al., 2010; Munizaga and Palma, 2012). From the perspective of system operation, Bertini and El-Geneidy (2003) demonstrated that archived stop-level data can be converted into valuable transit performance measures. The performance indicators of system supplies and the statistics on service level can be calculated with smart card data spatiotemporally, such as travel speed and time, schedule adherence, vehicle-kilometres or person-kilometres for every individual run, route, and day (Morency et al., 2007; Trépanier et al., 2009). For example, Park et al. (2008) as well as Jang (2010) investigated the travel time, transfers, and time distribution of trips for various modes in Seoul. The spatial and temporal variability of transit use for various types of cards was also analysed by Morency et al. (2007). Utsunomiya et al. (2006) also used smart card data from the Chicago Transit Authority to extract information on passengers' transit usage and access distance.

The research on passenger behaviour has focused on travel patterns. Macroscopically, the movement patterns of passengers across specific regions in urban areas have been uncovered (Srinivasan and Ferreira, 2002; Bagchi and White, 2005). By using an agglomerative clustering method, Kim et al. (2014) discussed both zones and movement patterns. For individual traveller behaviours, mainly regarding regularity and daily patterns, Lee and Hickman (2011) defined regular transit users as those making two or more trips during typical weekdays and found that travel patterns varied with card type. Ma et al. (2013) developed an efficient data mining method to demonstrate the temporal travel patterns and the pattern regularity for transit riders in Beijing, which allows transit authorities to evaluate the policy performance in public transit systems.

To conclude, smart card data deliver several advantages in metro system studies compared to the traditional survey data. First, it eliminates the sampling error by recording the movements of full proportion of passengers, while this completeness can be limited if smart cards are used in parallel with other payment methods. Second, it records continuously over a long period of time, allowing us to observe temporal fluctuations in travel demand and journey time and easily construct panel datasets for specific analysis.

2.2.2 Automatic vehicle location data

Automatic vehicle location (AVL) data is the information collected from the AVL system, which continuously monitors the geographic location and status of vehicles operating in an urban environment (Riter and McCoy, 1977). In public transport, AVL systems play an increasingly important role in bus and train operations.

In the bus sector, AVL data are mainly used to support real-time fleet management and operational control (Riter and McCoy, 1977; Ma et al., 2014). Providing huge amounts of accurate, continuous, disaggregated data on bus departure and/or arrival times at bus stops, the AVL data enable us to analyse when and where services are not operating as planned (Barabino et al., 2015), which shed the light on bus operational performance evaluation. Measures including percentile travel times, coefficient of variation (COV) of travel times, and average commercial speed and travel time distribution have been proposed to evaluate the reliability of bus routes (Yan et al., 2016). Based on AVL data, Ma et al., (2014) created a reliable buffer time to measure the vulnerability passengers perceived at bus stations. Barabino et al. (2015) integrated the AVL data and passenger patterns to construct a new punctuality measure, reflecting the fraction of passengers who will be served promptly after arriving at a bus station. From another perspective, AVL data can also be applied to improve the real-time bus information system and bus priority at signalised junctions (Horbury, 1999a). Horbury (1999b) also illustrated how historical AVL data are used to identify the segments of bus routes and estimate passenger arrival rates at stops.

In the public transport systems that consist of bus and tram, Marra and Corman (2020) applied the AVL data to cluster vehicle delays and inversely inferred the existence of disruptions. Similarly, in the rail-based transit sector, with scheduled and realised train departure and arrival times, AVL data have also been applied to service performance measurement and evaluation. Mesbah et al., (2012) carried out the reliability analysis for Melbourne tram network. Their outputs include mean and the standard deviation of travel times across the network, comparison between actual and scheduled travel times, and coefficients of variation.

In the urban metro systems, the AVL data are usually integrated with smart card data, which allows researchers to pair passengers' trips with train movement trajectories, thus enabling passenger to train assignment analysis. The most probably routes used by passengers can be inferred from the integrated data. These possible routes are then used to assign

passengers to trains. The results of such assignment will help obtain the decomposed journey time, transferring choice, and most importantly, the dynamic passenger flow in stations or trains, which cannot be achieved with only SCD or AVL data (Gordon et al., 2013; Hörcher et al., 2017; Yap et al., 2017; Zhu et al., 2017b). Such analysis on the one hand supports the crowding estimation in metro systems (Hörcher et al., 2017). On the other hand, the assignments can be combined with disruption impact analysis (Yap et al., 2017).

2.3 Metro disruption detection

In this section, we first review the common algorithms for anomaly detection. Then, in the field of transport, the literature on anomaly detection in road traffic has been briefly reviewed. Finally, we focus on the previous research on metro disruption detection.

2.3.1 Review of anomaly detection algorithms

According to Ahmed et al. (2016), the anomaly detection algorithms can generally be classified into four types: statistical-based, classification-based, clustering-based and information theory. Classification and clustering algorithms are two main branches in machine learning techniques, with the latter not relying on pre-labelled data. Statistical methods are usually based on underlying distribution, non-parametric model or some statistics, while information theory tries to distinguish the anomaly from the normal by understanding their underlying characteristics and mechanisms (Zhu, 2019). A summary of typical detection algorithms, their characteristics and classical representatives are shown in Table 2.1.

Table 2.1: Summary of important literature on the general anomaly detection algorithms

Category	Typical algorithms	Representatives	Characteristics
Statistical-based detection	Probability distributions model (Normal/Gamma…)	Vic Barnett and Tolewis (1994)	i). Assume that observed data are conformed to a distribution. ii). Outliers are identified outside the confidence interval.
	Histogram-based model	Kind et al. (2009)	i). No need to assume the distribution of data.

			ii). Model histogram patterns and detect deviations from the histogram models.
	Regression model	Vic Barnett and Tolewis (1994) Rousseeuw and Leroy (2005)	i). Assume the data follows a specific model. ii). Outliers are detected based on the deviation from estimated regression models.
	Mixture model	Eskin (2000)	i). Unsupervised with no label data. ii). A combination of machine learned probability distribution and statistical test detection.
Classification-based detection	Bayesian networks	Kruegel et al. (2003) Patcha and Park (2007)	i). A graphical model that encodes probabilistic relationships among variables of interest. ii). Similar results with threshold-based algorithms, but need higher computational requirements.
	Neural networks	Bishop (1994) Augusteijn and Folkert (2002)	i). Supervised learning ii). No priori assumptions on the properties of data. iii). Very small number of parameters need to be optimised for training networks.
	Rule-based/tree-based algorithms	Wong et al. (2002)	i). Rely on accurate labels ii). Proper configuration of rules requires precise, laborious and time-consuming analysis.
	Support vector machines	Davy and Godsill (2002)	i). Determining optimal hyperplanes for separating data from different classes.
Clustering-based detection	K-means clustering	Smith et al. (2002) Attar et al. (2014)	i). Unsupervised learning; does not require a priori labelled training data. ii). The anomaly score (degree of being outlier) is determined based on the clustering results.
	Nearest neighbour clustering	Eskin et al. (2002) Liao and Vemuri (2002)	i). Unsupervised learning. ii). Assume that outliers lie in sparse neighbourhoods, and they are distant from their nearest neighbours.
Information theoretic detection	Entropy-based	Bereziński et al. (2015)	i). Able to be used as unsupervised learning. ii). No assumption about the distribution for the sample data.

2.3.2 Research on traffic anomaly detection

In the road traffic sector, anomaly detection has been widely analysed. The concept of anomaly includes abnormal traffic conditions such as accidents, congestions (disruption of road services) or specific road-related events. The detection methods have evolved from constant human observance through CCTV monitors to automatic detection based on sensors and algorithms (Mahmassani et al., 1999). The exploited sensor data include inductive loop detectors data (Rossi et al., 2015; Zhu et al., 2018), social media data (Gu et al., 2016; Zhang et al., 2018), GPS trajectory data (D'Andrea and Marcelloni, 2017; Yu et al., 2020; Zhang et al., 2021), camera data (Cano et al., 2009; Sodemann et al., 2012; Riveiro et al., 2017; Santhosh et al., 2020) and mobile phone data (Steenbruggen et al., 2016; Bolla and Davoli, 2020), among others.

The above practices have shown that the detection of traffic anomalies is a data-specific task. Specifically, the choice of detection method depends on the structure and characteristics of the data used. For similar detection problems, those experiences in the road sector may guide the application of new sensor technologies in urban metros or shed light on developing suitable detection algorithms.

2.3.3 Research on metro disruption detection

In the urban rail sectors, huge efforts have been made to automatically detect faults in railway track circuits. Using circuit sensor signals and video images, researchers have developed diagnostic algorithms based on neural network, deep learning and Bayesian network to automatically detect faults in railway track circuits (Chen et al., 2008; Zhao, Wu and Ran, 2012; De Bruin et al., 2017; James et al., 2018; Welankiwar et al. 2018; Wei et al., 2019).

However, most urban metro systems still rely on manual incident detection methods, which rely on reports of manual inspections from metro operators and complaints from passengers (Ji et al., 2018). For example, when a disruption occurs in the London Underground, the staff involved are required to complete an Incident Reporting Form (IRF). After verification by an operational manager, the IRF is entered into the service data system called CuPID, which finally generates incident logs (London Datastore, 2018). These traditional detection methods suffer from human errors and manpower constraints, leading to missing observations due to limited monitoring range in space and time and incorrect records (Ji et al., 2018). Such shortcomings of traditional methods have encouraged researchers to explore automatic

disruption detection methods in urban metro systems by leveraging the emergence of large-scale datasets.

The first type of new data source is social media. Metro-related social media data include reviews or comments made by passengers about metro services. For instance, Ji et al. (2018) used Twitter data to detect service disruptions in the Washington Metro. The authors first filtered tweets with keywords of metro lines and stations during a given period. Subsequently, by mining common complaint vocabulary in these tweets (fail, disrupted, interrupted, and injury, among others), they predicted if there was a delay on a specific metro line. Similarly, Zulfiqar et al. (2020) used real-time Twitter data to develop an open-source system for the early detection of emergencies or criminal events within rail-based transit systems. By tracking the emerging information about each particular incident, they are able to track the chronological development of threatening events during the day. Compared with conventional incident logs data, the social media data can capture passengers' feedback and complaints promptly and cleverly monitor train services throughout the entire network. However, the detection of disruptions based on social media data cannot avoid the limitations of human inspections. For example, it might be the cases that not all disruptions are mentioned on social media, line/station information might be missing, and the posts may contain wrong or fake disruption information. Thus, the detection accuracy largely depends on the representativeness and quality of metro-related social media data.

The second type of new data source is automated fare collection or smart card data (SCD). SCD provides information about the origin and destination of passengers with timestamps, and thus, reveals the journey time and travel behaviour of passengers. Sun et al. (2016) used SCD to obtain passenger flows and regarded abnormal changes in passenger flow as a sign of disruption occurrence. They assumed that the passenger arrival rate during a specific period follows a normal distribution and estimated the distributional parameters via Bayesian inference. Subsequently, they considered all observations beyond three standard deviations of the mean passenger flow as the disruption indicator. In another study, Tonnelier et al. (2018) proposed four approaches for anomaly detection using SCD. The first three approaches inferred a daily temporal prototype (that is, a specific pattern of passenger behaviour depending on the day of the week) according to entry logs using three different methods: the average, the normalised average and a discrete probability density function obtained from the nonnegative matrix factorisation algorithm. Next, they obtained anomaly scores by determining the

difference between the inferred prototypes of station flow and the real observations of a particular day. The fourth approach is a user-based model in which they computed the log probability distribution of the entry frequencies at different stations for each passenger. Abnormal entry behaviour was detected at the passenger level and aggregated in spatiotemporal dimensions. Briand et al. (2019) used the unexpected increases or decreases in passenger demand to detect atypical events in the French transit network of the city of Rennes. After clustering observations with similar ridership activities, they conducted outlier detection based on the boxplot method. Jasperse (2020) also used SCD, but he relied on abnormality in journey time rather than demand patterns to identify irregular metro operations. The author attempted to detect the propagation of passenger delays in both spatial (that is, through nearby stations) and temporal dimensions. However, this study focuses on passenger delays rather than train service delays. In metro systems, service delays would generally result in passenger delays. But on the contrary, service delays cannot be directly inferred from passenger delays, since passenger delays may originate from other factors, such as overcrowding. This distinction is crucial.

In summary, SCD-based methods might omit some service disruptions because abnormal passenger behaviour is not highly correlated with the abnormal headways. Such weaker correlations can be attributed to the fact that anomalous travel patterns can be caused by many other factors apart from service disruptions: (i) inherent fluctuations in the passenger demand itself, (ii) weather conditions, (iii) mega-events near metro stations, and (iv) temporary demand control measures. Thus, SCD could be useful in detecting ridership related incidents, but the detection of service disruptions requires a new data source and method, unless the above causes of anomalous travel patterns can be fully controlled. Table 2.2 shows a comparison of recent research on metro disruption detection.

2.4 Metro vulnerability measurement

2.4.1 Definition of vulnerability

Vulnerability is generally understood as the quality of being weak and easily influenced under disruptive events. Since the 1990s, this concept has been widely used to characterise the performance of transport systems (Mattsson and Jenelius, 2015; Reggiani et al., 2015), which is often defined as a measure of susceptibility of the transport system to incidents (Berdica,

2002; Jenelius et al., 2006; O'Kelly, 2015). Due to different interpretations of susceptibility, the concept of vulnerability is sometimes confused with resilience, reliability and risk. To avoid such confusion, Faturechi and Miller-Hooks (2014) have summarised the most agreed interrelationship among these concepts (see Figure 2.1). Although there is no clear boundary to distinguish each concept, in general reliability tends to emphasise the probability of encountering disturbances, while vulnerability tends to emphasise the consequences of disturbances. Resilience, on the other hand tends to emphasise the recoverability of transport systems.

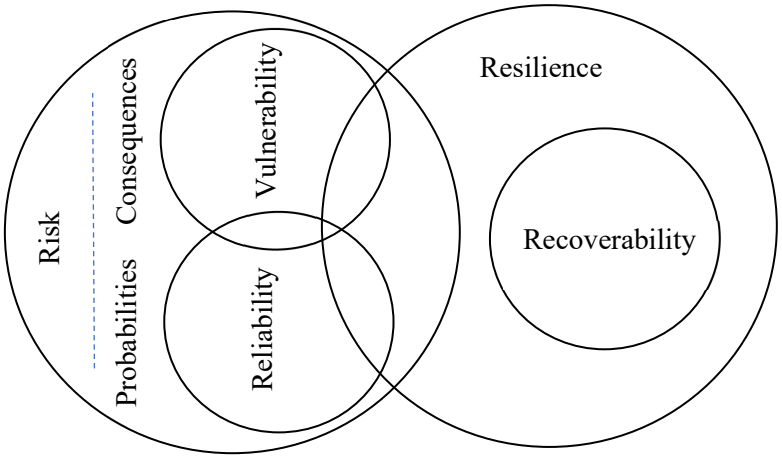


Figure 2.1: The interrelationship among vulnerability, reliability, risk and resilience [adapted from Faturechi and Miller-Hooks (2014)]

In this PhD, the vulnerability of metro systems refers to the extent of degradation in the level of service due to service disruptions. Vulnerability metrics are used to measure the consequences of service interruptions in the form of performance outputs such as train kilometres, passenger volumes or the quality of travel. For operators, such metrics have important implications in identifying weak stations or links in metro systems and efficiently allocating resources to the most affected areas (Sun et al, 2015; Chopra et al, 2016). Given the rising interest in utilising vulnerability metrics in disruption prevention and management, obtaining an accurate measure of such metrics is crucial.

2.4.2 Methods of vulnerability measurement

Traditionally, there are two types of method used to build vulnerability indicators for metro systems: topology-based methods and system-performance-based methods.

Topological methods rely on complex network theory to convert the metro network into a scale-free graph, in which nodes represent metro stations, edges represent links between directly connected stations and the weight associated with each edge is computed based on travel time or distance (Derrible and Kennedy, 2010; Zhang et al., 2011; Mattsson and Jenelius, 2015). The changes in the system's connectivity are reflected in graphs by removing nodes or links, and vulnerability is entirely governed by the topological structure. For instance, the location importance of metro stations or links is indicated by the number of edges connected to a specific node and the fraction of shortest paths passing through the given node/edge (Yang et al., 2015; Sun and Guan, 2016; Sun et al., 2018; Zhang et al., 2018b). Network-level efficiency is indicated by the average of reciprocal shortest path length between any origin-destination (OD) pair. Such global indicators capture the overall reachability as well as the service size of a metro system (Sun et al., 2015; Yang et al., 2015).

System-performance-based analyses not only consider the network topology but also incorporate real data on metro operations (e.g., ridership distribution) into vulnerability measurement (M'cleod et al., 2017; Mattsson and Jenelius, 2015). For instance, Cats and Jenelius (2014) introduced a dynamic-stochastic setting to extend the topological measures of betweenness centrality and link importance. Sun et al. (2018) use a ridership-based indicator – a sum of flows in edges connected with the given node – to complement the topological measures by integrating passengers' travel preferences. Other studies use passenger delay and demand loss as vulnerability indicators (Rodríguez-Núñez and García-Palomares, 2014; Adjetey-Bahun et al., 2016; M'cleod et al., 2017; Cats and Jenelius; 2018; Nian et al., 2019). Specifically, the network-level passenger delay can be calculated based on changes in the weighted average of travel time between all OD pairs due to disruptions where weights are station-level passenger loads. Jiang et al. (2018) suggest integrating land use characteristics around stations into vulnerability measurement because metro systems interact with the external environment during incidents.

To quantify vulnerability based on the aforementioned indicators of the system's performance, almost all previous studies adopt simulation-based approaches and assume hypothetical disruption scenarios. The simplest disruption scenario involves a single station or link closure, assuming one node or edge in the graph is out of service. This incident affects the topology structure and passengers' route choice and the differences in the corresponding performance indicators under normal and disrupted scenarios are quantified to measure

vulnerability (Sun et al., 2015). More complex disruption scenarios include the closure of two or more non-adjacent stations, failure of an entire line, and sequential closure of stations until the network crashes (Adjetey-Bahun et al., 2016; Chopra et al., 2016; Sun and Guan, 2016; Zhang et al., 2018a; Zhang et al., 2018b). Ye and Kim (2019) discussed the case of partial station closure. Cats and Jenelius (2018) also moved beyond the complete failures and investigated the partial capacity degradation at line level and link level.

Simulation-based studies gained popularity because they do not require incident data and can flexibly control simulation settings to imitate a wider range of possible situations. However, researchers have to make many assumptions to infer passengers' response to virtual disruptions. Without observing passengers' movements during real incidents, the validity of the simulation assumptions is questionable. For example, while quantifying passenger delay indicators, Rodríguez-Núñez and García-Palomares (2014) and Adjetey-Bahun et al. (2016) assume that all passengers have the same travel speed and they do not change their destinations under disruptions unless there is no available route. However, in reality, passengers can travel at different speeds, leave the metro system, change their destinations, or reroute during disruptions. As a result, especially for system-based analyses, vulnerability metrics obtained from simulation-based studies may not reflect the true changes in the level of service due to disruptions. There is, therefore, scope to improve vulnerability measurement by empirically estimating the impact of disruptions. The advantage of empirical-based methods is that the aforementioned assumptions are no longer needed, and the estimated impacts of disruptions are more reliable. However, a drawback of empirical studies is that they require high-quality and sufficient data. Table 2.3 shows a comparison of recent vulnerability studies and illustrates the contribution of this research.

2.5 Disruption impact estimation

2.5.1 Simulation-based research

The simulation-based studies quantify the impacts of hypothetical disruption scenarios, with no need for real disruption data. Generally, for metro systems, this part of the literature is overlapped with that of simulation-based vulnerability measurement because vulnerability is defined as the susceptibility to incidents (see Section 2.4.2).

In recent years, the simulation-based research of metro vulnerability has evolved from pure complex network theory to system-performance based analysis. Therefore, the latest studies in this area started to the model the response of individual passengers under simulated disruptions. Shelat and Cats (2017) proposed an indicator of local link criticality to quantify the spatial propagation effects of link disruptions. The public transport network assignment model based on stochastic user equilibrium was developed to calculate such criticality. Since the last ten years, a mesoscopic public transport operations and assignment model called BusMezzo has become a powerful tool to dynamically simulate individual travel decisions, especially for the route choice (Toledo et al., 2010; Cats, 2013). Multiple studies on disruption impacts have been conducted based on BusMezzo. For example, in short-horizon and unplanned disruption scenarios, Cats and Jenelius (2014) simulated the disruption impacts as changes in passengers' welfare and operational costs of rolling stock, based on stochastic and dynamic trip assignments. Malandri et al. (2018) analysed the spill-over effects of public transport disruptions via the passenger volume over capacity ratio (VOC), which intended to indicate the crowding level throughout the network. They used simulated disruption scenarios to measure the change of VOC at individual trip level. Based on the transfer inference algorithm and hypothetic disruptions, Yap et al. (2021) conducted experiments to evaluate the impacts of different train rescheduling strategies on the disruption propagation from the regional train network to other public transport modes (Yap et al., 2017). In addition, Leng et al. (2018) and Paulsen et al. (2018) applied another agent-based simulation software (MATsim) to analyse rail disruption impacts on passenger delays in the metropolitan areas of Zürich and Copenhagen, respectively.

However, without empirical observation of real disruptions, the great number of assumptions on passengers and virtual interruptions can be the main problem for simulation-based studies.

2.5.2 Empirical research

In an urban rail transit context, early attempts to analyse disruption impact relied on survey data. Rubin et al. (2005) conducted a stated preference survey to understand the psychological and behavioural reactions of travellers to the bombing incident that occurred in London in July 2005. They considered passengers' reduced intention of travelling via the London Underground after the attack as the key indicator. Since stated willingness may not reflect real

travel behaviour, Zhu et al. (2017a) performed a revealed preference survey to investigate travellers' reactions to transit service disruptions in the Washington D.C. Metro. By comparing their actual travel choices before and during the metro shutdown, they found a 20% reduction in demand. Results from such surveys are usually presented as the percentage change in passengers' preferences for travel modes, departure time, and destinations. Although this information is useful, we still need detailed information about delays or demand losses to quantify true disruption impacts. Furthermore, there are inherent limitations of survey-based studies. For instance, repeated observations of a respondent are difficult to collect over a long period because of constraints associated with cost, manpower, recording accuracy, and privacy protection of respondents (Kusakabe and Asakura, 2014). Moreover, a survey sample cannot cover all passengers, which may lead to biased estimates of disruption impact if the sample is not representative of the population.

With the wide use of automated fare collection facilities in metro systems, smart card data have become a powerful tool for research related to transit operations and travel behaviour (Pelletier et al., 2011). Compared to survey data, the key advantages of smart card data are accurate, cost-effective, and continuous observations for each passenger within the system for a long time (Kusakabe and Asakura, 2014). Therefore, researchers have started using smart card data to analyse disruption impacts. For instance, Sun et al. (2016) conducted passenger assignment based on smart card data. They estimated the disruption impact as the differences between the assignment results under real incidents and normal conditions, in terms of ridership distribution and journey time across all OD pairs. This study does not require extra assumptions about passengers' reaction because their actual locations and movements have been revealed from smart card data. However, they conventionally assume that metro disruptions occur randomly. In reality, factors such as travel demand, signalling type, passenger behaviour, operating years, rolling stock characteristics and weather conditions can have a significant influence on the likelihood of metro failures (Melo et al., 2011; Wan et al., 2015; Brazil et al., 2017). These confounding factors may also affect the corresponding impact of disruptions (Imbens and Rubin, 2015). This is a particularly important consideration because, under non-random disruptions, the impact estimated from before-after comparison will be biased.

Some researchers also adopt prediction-based approaches to quantify disruption impact via smart card data. Silva et al. (2015) used past disruptions to predict the exit ridership and passenger behaviour for unseen scenarios, such as station closure and line segment closure.

Similarly, Yap and Cats (2020) applied supervised learning approaches to predict the passenger delay caused by incidents. These prediction-based studies still cannot disentangle the causal effect of disruptions and can result into biased estimates due to the existence of confounding factors.

In a very recent empirical study, the above research gaps have been bridged by Zhang et al. (2021a). Based on large-scale smart card data, they proposed to use propensity score matching methods to quantify the causal effects on service performance indicators, which allows for the non-random occurrence of disruptions and adjusts for potential bias caused by confounding factors. Nevertheless, the design of their causal inference framework still suffers from some limitations. First, metro stations are assumed to be independent and there is no interference between different stations. In other words, the disruption impacts are restricted to the station where it occurred, so other parts of the network are considered not to be affected. This assumption oversimplifies the connections between stations in metro networks. Adjacent stations are linked by metro lines and successive train movements; thus, disruption impact can actually spread along metro lines and influence the entire network. Previous empirical-based studies have concentrated on the direct impacts of disruptions, while the propagation of indirect impacts (spillover effects) has not yet been explored empirically. Second, the outputs of propensity score matching methods are the average causal impacts of all disruptions observed during the given study period. The proposed framework is not suitable for estimating the causal effects of individual disruption.

Table 2.2: A comparison of recent research on metro disruption detection

Research	Source data	Detection indicator	Detection method	Detection accuracy		Disruption propagation	Recovery intervention
				No human errors	True service disruption		
Sun et al., 2016	Smart card data	Boarding ridership	Three-standard-deviation rule with Bayesian inference	√			
Ji et al., 2018	Twitter data	Complain/delay vocabulary in tweets	Multitask supervised learning		√		
Tonnelier et al., 2018	Smart card data	Entry logs	Anomaly scores compared to baseline	√			
Briand et al., 2019	Smart card data	Passenger demand	Boxplot	√			
Jasperse, 2020	Smart card data	Passenger delay	Hierarchical clustering and probabilistic classification	√		*[4]	
Zulfiqar et al., 2020	Twitter data	Crime/emergency vocabulary in tweets	Based on keywords and dynamic query expansion				
Our approach	Vehicle location data	Service headway	Probabilistic Gaussian mixture model	√	√	√	√

[4] Jasperse (2020) analysed the spread of passenger delays, rather than the propagation of train service delays.

Table 2.3: A comparison of recent research on metro vulnerability

Research	Vulnerability metrics or disruption impacts		Analysis approach		Smart card or OD data	Land use	Non-random disruptions
	Topology-based	System performance-based	Simulation-based	Empirical (real incidents)			
Derrible and Kennedy, 2010	√		√				
Zhang et al., 2011	√		√				
Yang et al., 2015	√		√				
Chopra et al, 2016	√		√				
Zhang et al., 2018a	√		√				
Zhang et al., 2018b	√		√				
Ye and Kim, 2019	√		√				
Rodríguez-Núñez and García-Palomares, 2014		√	√		√		
Adjetey-Bahun et al., 2016		√	√		√		
M'cleod et al., 2017		√	√		√		
Cats and Jenelius, 2018		√	√				
Cats and Jenelius, 2014	√	√	√		√		
Sun et al., 2015	√	√	√		√		
Sun and Guan, 2016	√	√	√		√		
Sun et al., 2018	√	√	√		√		
Lu, 2018	√	√	√		√		
Jiang et al., 2018		√	√		√	√	
Sun et al., 2016		√		√	√		
Zhang et al., 2021a		√		√	√	√	√

2.6 Research gaps

We conclude this section with a summary of the potential research gaps identified in the literature.

Disruption detection:

i). For the detection of service (operational) disruption in metro systems, methods based on incident logs and social media data can be unreliable due to inevitable human errors and missing observations. Recent SCD-based methods capture abnormal passenger behaviour as an indicator of disruption occurrence, but they might not detect service disruptions due to the lack of one-to-one association between service delays and abnormality in passenger behaviour.

ii). Previous detection methods rarely pay attention to the propagation of disruption across space (that is, along metro lines) and time. Neither the SCD nor the social media data contain effective information to identify the disruption propagation. SCD can capture passenger delay propagation, but it does not translate into the measure of service delay propagation.

Metro vulnerability measurement:

iii). Previous studies on vulnerability metrics of transit systems are largely based on simulation approaches. These studies do not account for the actual behaviour of passengers under disruptions. Basing analyses on empirical data, rather than simulations, obviates the need for making potentially unrealistic assumptions about passengers' movement.

Disruption impact quantification:

iv). In urban metro systems, disruption occurrences can be non-random. Therefore, empirical studies on quantifying disruption impacts should account for this non-randomness to eliminate confounding biases in estimation.

v). Besides the interrupted location, disruption impacts can spread to other functioning stations in the metro network. Therefore, under the causal inference framework, empirical studies on quantifying disruption impacts should account for such

interference among stations. That is, the assumption on independent stations needs to be relaxed.

vi). Empirical-based impact analyses published so far have not explored the propagation of indirect impacts (spillover effects) throughout metro networks.

In this PhD research, gaps (i) and (ii) serve as a basis for Chapter 4 in the literature. Chapter 5 shows that both improvements (iii) and (iv) can be made by adopting causal inference methods and using the empirical disruption impacts to generate vulnerability measures. Finally, in Chapter 6 we implement the research ideas (iv) and (v) under the causal inference framework, to explore the spillover effects of disruptions and their spatiotemporal propagation within metro networks.

Chapter 3
Methodological background: Causal inference methods

According to the review in Chapter 2, there are two key gaps in the literature of empirical vulnerability measurement and disruption impact estimation. The first is confounding issues caused by non-random disruption occurrence, and the second is the interference among stations in metro networks. To fill these gaps, an important contribution of the thesis is to apply novel causal models via large-scale automated data to improve the evaluation of disruption impacts. This chapter provides the fundamental knowledge of causal inference framework and a review of the literature on techniques for the aforementioned research gaps. The objectives of Chapter 3 are to highlight the importance of causal inference methods and to facilitate the understanding of specific techniques that are used for vulnerability measurement in Chapter 5 and impact quantification in Chapter 6.

In recent decades, causal inference has been widely studied in social and biomedical sciences, contributing to the discovery of how actions or interventions (commonly referred to as treatments) affect outcomes of interest (Imbens & Rubin, 2015, Bojinov et al., 2020). Statisticians and econometricians leverage this powerful tool to determine causality in many fields, such as urban economics, public health, and a wide range of public policy decisions. However, such techniques have not been applied in metro disruption analysis. Therefore, in Section 3.1 we first provide a general overview of the literature on the causal inference framework.

Meanwhile, the two research gaps identified above also imply methodological challenges in causal inference. First, the issue of confounding is a common concern in observational studies. With appropriate assumptions and assignment mechanisms, the bias from confoundedness can be eliminated. The second is the interference issue which is a more challenging problem since it violates a basic assumption for the majority of causal inference methods: the stable unit treatment value assumption (SUTVA). In Section 3.2, we discuss the

viable solutions to these two challenges in detail. Further contextual discussion of the proposed solutions in this chapter is presented in Chapter 5 and Chapter 6.

3.1 The basic framework

3.1.1 Potential outcomes

The fundamental notion underlying the Rubin causal model (RCM) is that causality behind a $treatment$ (or intervention) is applied to a $unit$ (Imbens & Rubin, 2015). The unit i can be a physical object, a person or a collection of objects/persons, at a particular point in time. For unit i, let W_i indicate the treatment enrolment, and Y_i denote the outcomes (results or effects of the treatment on a response variable) of interest. The potential outcomes for a binary treatment are defined as (Imbens and Wooldridge, 2009):

$$Y_i(W_i) = Y_i(0) \times (1 - W_i) + Y_i(1) \times W_i, \tag{3.1}$$

$$Y_i = \begin{cases} Y_i(0) & if\ W_i = 0 \\ Y_i(1) & if\ W_i = 1 \end{cases}.$$

$Y_i(1)$ denotes the outcomes that unit i would attain if it is exposed to the treatment ($W_i = 1$); conversely $Y_i(0)$ denotes the outcomes that would be attained if unit i did not receive the treatment ($W_i = 0$). Since the unit i can be either treated or not, the two potential outcomes are counterfactual and only one will be ultimately observed. Y_i denotes the observed outcome. The causal effect of a treatment involves the comparison of these two corresponding potential outcomes at the unit level, which is defined as $Y_i(1) - Y_i(0)$.

Although the definition of causal effects does not require more than one unit, learning about causal effects typically requires multiple units. In order to exploit the presence of multiple units, the stable unit treatment value assumption (SUTVA) is introduced by Rubin (1978). First, the potential outcomes for any unit does not vary with the treatments assigned to other units. Second, for each unit there is no hidden variation of treatment that leads to different potential outcomes. Thus, under SUTVA the average treatment effect on the treated units (ATET) is

$$\tau_{ATET} = E[Y_i(1) - Y_i(0)|W_i = 1]. \tag{3.2}$$

'The fundamental problem of causal inference' (Holland, 1986) is therefore a missing data problem that the missing potential outcomes associated with the unrealised treatment need to be imputed.

3.1.2 The assignment mechanism

The second ingredient of the RCM is the assignment mechanism, which is defined as the conditional probability of receiving the treatment, and formulated as a function of potential outcomes and unit-specific background attributes, satisfying

$$0 < \Pr(W_i|X_i, Y_i(0), Y_i(1)) < 1.$$

The attributes for unit i, also referred to as pre-treatment covariates, are denoted by vector X_i with k-component row. Such covariates can explain some of the variation in outcomes and the key characteristic is that they are unaffected by the treatment assignment.

Generally, based on different assumptions, there are three classes of assignment mechanism: randomized experiments, unconfounded assignment and other assignment mechanisms. The first class is randomized experiments, where the probability of assignment to treatment does not vary with potential outcomes:

$$W_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big).$$

This assignment mechanism has a known function of covariates controlled by the researcher. The statistical analysis of such experiments is straightforward, but experimental evaluations have traditionally been rare in economics and transport fields (Imbens and Wooldridge, 2009).

The second class of assignment mechanics is under the unconfounded assumption, where the assignment probabilities are conditional independent to the potential outcomes, given covariates X_i (Rosenbaum and Rubin, 1983),

$$W_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) \,\big|\, X_i.$$

Unconfounded assignments have been widely used in observational studies, incorporated in approaches such as outcome regressions, propensity score matching and weighting (Heckman et al., 1998). One of the most popular approaches is propensity score matching, which is discussed in the next section.

The final class of assignment mechanisms consists of all remaining mechanisms with a certain amount of dependence on potential outcomes. Relaxing the unconfounded mechanism

may lead to biased causal estimates. The representative solutions for this problem include differences-in-differences (Ashenfelter,1978; Ashenfelter and Card; 1985), instrumental variables (Angrist et al., 1996) and regression discontinuity (Thistlewaite and Campbell, 1960; Cook, 2008). However, in this thesis, those methods are not applicable in the context of metro disruptions.

3.2 Challenges and viable methods for disruption impact analysis

In this PhD, we aim to introduce the causal inference framework to urban metro systems. Therefore, the study units are metro stations at a specific time of day, and the treatments are defined according to whether they receive disruptions. As mentioned at the beginning of this Chapter, there are two major concerns in estimating disruption impacts: (i) non-random occurrence of disruptions, and, (ii) metro stations are connected and therefore the outcomes interact with each other. The non-randomness of disruptions implies that some station covariates can influence both the disruption occurrence and the outcomes, ignoring which may lead to biased causal estimates. Fortunately, this challenge can be resolved by applying unconfounded assignment: propensity score matching in case of the present thesis. Section 3.2.1 discusses the rationale behind and briefly reviews this method. The second challenge originated from the presence of interference among station-level outcomes, which violates the SUTVA. In Section 3.2.2, we summarise possible solutions that incorporate the interactions among units, specifically discuss the synthetic control methods.

3.2.1 Treating confoundedness

Under confoundedness, by adjusting for differences in pre-treatment covariates and outcomes, the bias in comparisons between treated and control units can be obviated. That is the main idea of unconfounded assignment, conditional on the covariates, the treatment assignments are independent of the potential outcomes. Matching is one of the prominent approaches based on unconfounded assignment, which pairs each treated unit with an untreated unit with the same value on observed attributes. The causal effects are then estimated from the comparison within the matched pairs. Matching estimators have been widely applied in settings where (i) the interest is in the average treatment effect for the treated, and (ii) there is a large reservoir of

potential controls (Rubin, 1973a, b; Rosenbaum, 1989; Rubin and Thomas, 2000; Heckman et al., 1998).

In practice, considering the difficulty in comparing high-dimensional covariates, a single index (propensity score or also referred to as balancing score) is proposed to represent all the covariates. Mathematically, the propensity score is described as a function of the covariates (Imbens, 2000)

$$W_i \perp\!\!\!\perp \left(Y_i(0), Y_i(1)\right) \mid e(X_i). \tag{3.3}$$

When the conditional independence assumption, SUTVA and overlap (common support) assumption, $0 < Pr(W_i \mid X_i) < 1$, are all held, the average treatment effect of the treated can be derived as

$$\tau_{ATET} = E[Y_i(1) - Y_i(0) | W_i = 1] = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{Y}_i(1) - \hat{Y}_i(0)\right), \tag{3.4}$$

$$\hat{Y}_i(0) = \frac{1}{M}\sum_{j \in J_M(i)} Y_j,$$

where $\hat{Y}_i(0)$ denotes the missing potential outcome for unit i. $J_M(i)$ represents a collection of indices of the closest M control units matched for the current treated unit. The composition of J_M is determined by matching algorithms. Five commonly used matching algorithms are outlined below. In general, the performance of different algorithms largely depends on the structure of the data used (Stuart, 2010). A contextual discussion of the propensity score matching methods (Imbens and Wooldridge, 2009) is further presented in Chapters 5.

Nearest neighbour matching:

Units from the comparison group with the closest propensity score are chosen as matches for given treated units. This is one of the most straightforward matching methods (Rubin, 1973a).

Optimal matching:

Optimal matching, one complication of nearest neighbour matching, takes into account the overall set of matches when choosing individual matches, minimising the total distance within matched pairs (Gu and Rosenbaum, 1993; Rosenbaum, 2002).

Caliper and radius matching:

This is a generalised form of nearest neighbour matching. Given a tolerance on the maximum propensity score distance, this matching scheme uses not only the nearest neighbour but all comparison members within the tolerance level (Smith, 1997; Rubin and Thomas, 2000).

Subclassification and interval matching:

The idea is to partition the common support of the propensity score into a set of strata. Within each stratum, the treatment effects are evaluated as the difference between the outcome of treated and all comparison individuals (Rosenbaum and Rubin, 1985).

Kernel and local linear matching:

These are nonparametric matching estimators using a weighted average of all untreated individuals to construct the counterfactual outcome of each treated individual. The closer to the propensity score of treated units, the greater the weight is (Heckman et al., 1998).

3.2.2 Treating interference

In the classical potential outcomes framework, the SUTVA plays an important role. It assumes that there is no interference between units. However, in many experimental and observational studies, where units interact with each other physically or socially, interference is likely to be present and the SUTVA is no longer plausible (Kuang et al., 2020). With such interference, the assigned treatment on one unit can have direct effects on its own and spillover effects on the potential outcomes of other units. For example, in the context of metro networks, adjacent stations are connected by tracks and continuous train services. When one disruption occurs in a station, the adverse impacts such as delays and crowding can spread to the entire network via metro lines. The presence of interference among stations implies that an individual disruption may affect the performances of all stations in the system. In practice, both direct and spillover effects are of great interest for metro operators. Thus, estimating causal effects in the presence of interference becomes an inevitable challenge.

To address interference issues under the causal inference framework, one general solution is to redefine the unit of interest and try to eliminate the interactions among the newly defined units, e.g., by aggregation (Imbens and Wooldridge, 2009). Out of the ordinary in this area, a series of novel studies view interference or interactions as the primary object of interest, rather than as a nuisance. These possible solutions rely on specifying the interactive

relationships among units and possibly modelling the interference mechanism, which can be summarised into two directions (Kuang et al., 2020).

The first direction targets the partial interference assumption (Sobel, 2006): when study units are partitioned into non-overlapping clusters or groups, interference exists only within each cluster/group and there is no interference between clusters/groups (Hudgens and Halloran, 2008; Forastiere et al., 2016; Kang and Imbens; 2016; Rigdon and Hudgens, 2015; Grossi et al., 2020). In the other direction, researchers have considered relaxing the partial interference assumption to account for a more general structure of interference (van der Laan, 2014; Forastiere et al., 2016; Liu et al., 2016; Aronow and Samii, 2017; Forastiere et al., 2021). New designs of the interference structure have been proposed, for example Verbitsky-Savitz and Raudenbush (2012) allowed the potential outcomes of one unit to depend on a function of the treatment assignments of all other units. Aronow and Samii (2017) and van der Laan (2014) limited the interference in immediate neighbours and ruled out the influence from other units. An (2018) developed a treatment diffusion network to measure the treatment interference between treated and control units.

The causal inference methods involved in the above literature include randomised experiments (Rosenbaum, 2007; Aronow, 2012; Basse and Feller, 2018), inverse probability weighted (IPW) (Hudgens and Halloran, 2008; VanderWeele et al., 2012; Liu et al., 2016), matching (An, 2018) and synthetic control (Cao and Dowd, 2019; Grossi et al., 2020). Next, we specifically discuss synthetic control methods, which can properly relax the non-interference assumption with simple modifications.

Synthetic control methods were originally proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010). The main idea of this method is that a combination of unaffected units often provides a more appropriate comparison than any single unaffected unit alone. Therefore, to quantify causality, a treated unit is compared with a synthetic control unit, which is the weighted average of the unaffected units from the corresponding "donor pool". The donor pool is a group of unaffected units that have similar pre-treatment characteristics as the treated units. Formally, let j denote $J+1$ units $j = 1, 2, \dots, J+1$, the first unit ($j = 1$) is set to be the treated unit and the donor pool ($j = 2, \dots, J+1$) is a collection of untreated units, not affected by the treatment. For time span T periods, given the first T_0 periods are before the treatment. The effect of the treatment on the affected unit in period t is defined as (Abadie, 2021)

$$\tau_{1t} = Y_{1t}^{I} - Y_{1t}^{N} \quad t > T_0, \tag{3.5}$$

$$\hat{Y}_{1t}^{N} = \sum_{j=2}^{J+1} c_j \, Y_{jt}$$

where for each unaffected unit j, at time t, Y_{1t}^{N} denotes the potential outcomes without treatment. Y_{1t}^{I} defines the potential outcomes under the treatment. A synthetic control (defined as a weighted average of the units in the donor pool) is represented by a $J \times 1$ vector of weights $\boldsymbol{C} = (c_2, \dots, c_{J+1})'$. Given a set of non-negative weights $\boldsymbol{V} = (v_1, \dots, v_k)'$, optimal synthetic control $\boldsymbol{C}^* = ({c_2}^*, \dots, {c_{J+1}}^*)'$ is obtained from the following minimization problem:

$$\min_{\boldsymbol{C}} \|\boldsymbol{X_1} - \boldsymbol{X_0}\boldsymbol{C}\| = \sqrt{\sum_{h=1}^{k} v_h \left(X_{h1} - c_2 X_{h2} - \cdots - c_{J+1} X_{hJ+1}\right)^2}, \tag{3.6}$$

$$such\ that \sum_{j=1}^{J+1} c_j = 1, \quad c_j > 0,$$

where $X_{1j}, \dots, X_{kj}$ denote a set of k predictors of the outcomes for unit j. Vector $\boldsymbol{X_1}$, denotes the predictors for the treated unit. $\boldsymbol{X_0} = [\boldsymbol{X_2}, \dots, \boldsymbol{X_{J+1}}]$ represent the corresponding predictors for the J untreated units. Given, the synthetic control weights $\boldsymbol{C}(\boldsymbol{V})$, the choice of $\boldsymbol{V}$ can be determined by minimising the following mean squared prediction error:

$$\min_{\boldsymbol{V}} \sum_{t=t_0+1}^{T_0} \left(Y_{1t} - \widetilde{w}_2(V) Y_{2t} - \cdots - \widetilde{w}_{J+1}(V) Y_{J+1t}\right)^2, \tag{3.7}$$

where t_0 is the length of the initial training periods within the pre-treatment periods T_0. The remaining periods belong to a subsequent validation period. Both weights $\boldsymbol{V}$ and $\boldsymbol{C}$ are selected by best fitting the outcomes and predictors of the treated unit using pre-treatment data (training and validation periods).

It is worth noting that generally the synthetic control methods follow the SUTVA, where untreated units are not affected by the treatment. This is because, under normal circumstances, the weighted average of post-treatment control units (in the donor pool) is used to predict the counterfactual outcomes of the treated units. In the presence of interference, post-treatment controls will be contaminated by the spillover effects, resulting in a biased estimator of the counterfactual potential outcomes, which implies a biased estimate of causal effects. However, with simple improvements to the design of the donor pool or the specification of the interference mechanism, the modified synthetic control methods can be unbiased and thus relax

the SUTVA. For instance, Cao and Dowd (2019) proposed to assume that the treatment effects and the spillover effects are linear in some unknown parameters. With the known structure of interference, they obtained asymptotically unbiased estimators for the treatment and spillover effects. Grossi et al. (2020) generalised synthetic control group methods under the partial interference assumption.

In this thesis, we introduce a novel and intuitive modification of the donor pool design. With large-scale automated data on both disrupted and normal days, all control units in the donor pool are selected from the days without any disruption. That is, control units in the donor pool would not be affected by any treatment, which therefore results in unbiased direct and indirect causal estimates under interference. A contextual discussion of our modified synthetic control method is further presented in Chapter 6.

Chapter 4

Detecting metro service disruptions via large-scale vehicle location data

Urban metro systems are often affected by disruptions such as infrastructure malfunctions, rolling stock breakdowns and accidents. The crucial prerequisite of any disruption analytics is to have accurate information about the location, occurrence time, duration and propagation of disruptions. To pursue this goal, in the present chapter we detect the abnormal deviations in trains' headways relative to their regular services by using Gaussian mixture models (GMM). Our method is a unique contribution in the sense that it proposes a novel, probabilistic, unsupervised clustering framework and it can effectively detect any type of service interruptions, including minor delays of just a few minutes. In contrast to traditional manual inspections and other detection methods based on social media data or smart card data, which suffer from human errors, limited monitoring coverage, and potential bias, our approach uses information on train trajectories derived from automated vehicle location (train movement) data. As an important research output, this chapter delivers innovative analyses of the propagation progress of disruptions along metro lines, which enables us to distinguish primary and secondary disruptions as well as recovery interventions performed by operators.

4.1 Introduction

With high-frequency services and large capacity, metros (also known as subways or rapid transit) play a vital role in transporting the urban population. However, large-scale urban metro systems are vulnerable to service disruptions, which cause passenger delays, crowding concerns and can negatively affect passenger satisfaction with metro operations. These disruptions are often caused by unpredicted infrastructure malfunctions (e.g., signal failures and track blockages), rolling stock breakdowns and accidents, planned maintenance work, and temporal dispatching adjustments (Jespersen-Groth et al., 2009). To quantify and improve a system's reliability, metro operators need accurate information on the disruption location,

occurrence time, duration and network propagation. This information can also help operators prepare an effective recovery plan, an essential input to disruption management and future maintenance planning. Providing service disruption information to metro users in real-time is also an integral component of the advanced passenger information system. With the latest updates regarding the detected disruptions, delayed status and expected recovery time, passengers are better able to reschedule their trips under unexpected disruptions. Considering that the reliability of metro services and the information regarding disruptions are critical for both metro operators and passengers, this research develops a data-driven method to detect abnormal deviations in trains' headways relative to their regular services due to both sudden disruptions and planned interventions.

Traditionally, to detect disruptions urban metro operators rely on reports from manual inspections and complaints from passengers. Such detection results usually suffer from human errors and are restricted to a limited monitoring range in both space and time due to resource constraints (Ji et al., 2018). Therefore, recent studies have used two new data sources to identify disruptions. First, Ji et al. (2018) and Zulfiqar et al. (2020) leverage social media data such as tweets with the keywords of metro lines, stations and common complaint vocabulary to predict disruptions. Although social media data can capture a significant amount of passengers' feedback and cleverly monitor metro disruptions in spatiotemporal dimensions, human errors cannot be circumvented in this approach. Second, a few studies have mined automated fare collection or smart card data (SCD) to capture abnormal passenger behaviour and assume that uncommon travel patterns such as anomalous change of station ridership and extra journey time are good indicators of incident occurrence (Sun et al., 2016; Tonnelier et al., 2018; Briand et al., 2019; Jasperse, 2020). However, such indicators may not be ideal for detecting service disruptions because, instead of train interruptions, other factors can also significantly affect passenger behaviour and corresponding demand measures. For instance, adverse weather conditions and external mega-events (e.g., concerts or sports matches) may cause demand fluctuations.

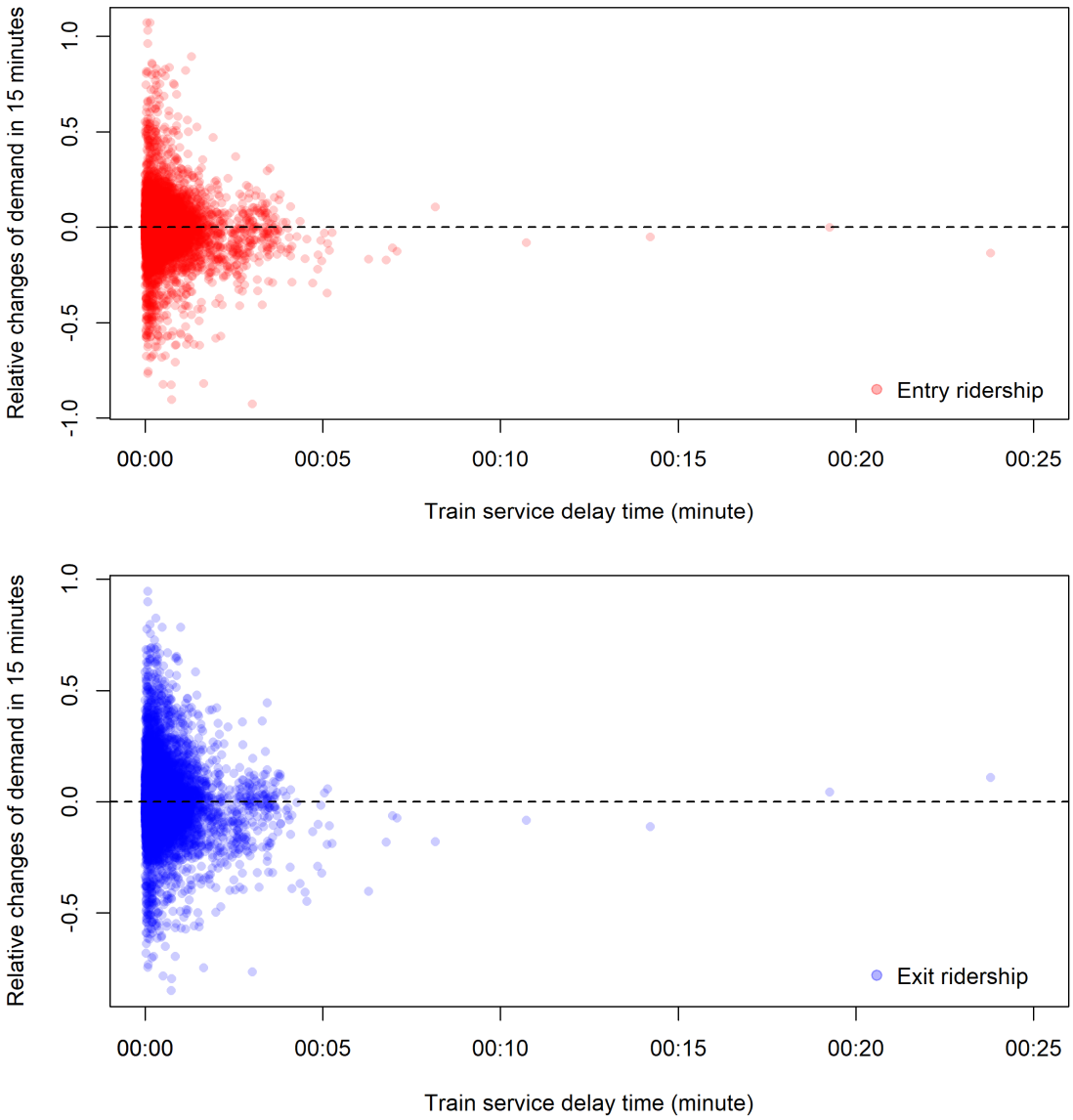


Figure 4.1: The relationship between abnormal demand and service disruption – example of a busy station, MTR. Relative changes in entry and exit ridership are derived by comparison with the average ridership under normal services.

To validate this hypothesis with data, we plot the changes in entry and exit ridership relative to the mean ridership against service delay time (that is, deviation from the scheduled headway) in Figure 4.1 for a busy station on a densely used line of the Hong Kong Mass Rapid Transit (MRT). The figure shows that both entry and exit ridership may change substantially even if there is no service delay. This trend implies that demand fluctuations might be caused by factors other than service disruptions. Conversely, the passenger demand appears to remain normal (with minor variations) even under ten-minute-long train disruptions. We also observe

a similar trend if we replace service delay time with journey time deviation. Although SCD are useful in detecting demand-related incidents in the system, an aberrant change in demand or journey time is not necessarily a sign of service disruption. Therefore, there is a need to explore the potential of other emerging datasets and methods to detect service disruptions. We focus on service (train) delays because information on the propagation of such delays is more useful in informing the corresponding control measure as it can be directly integrated with schedule optimisation models. Historical delay data have been applied to optimise timetables, disturbance-recovery strategies, and energy consumption of metro systems (Yang et al., 2019; Li et al., 2020).

This research proposes a novel approach to detect service disruptions using large-scale vehicle location data. The deviation in headway relative to the scheduled headway is used as the indicator of disruption occurrence, which is free of human errors and would enable an analyst to investigate service disruptions across spatial and temporal dimensions. The proposed method involves two steps – (i) split the day into 30-minute intervals and detect whether the platform is disrupted during a specific interval, and (ii) identify the propagation of disruption across the metro line over time.

First, we apply a Gaussian mixture model (GMM) on the headway deviations to identify a cluster of abnormal headways. This approach solves the detection problem within an unsupervised learning framework and obviates the need for subjective definition of outliers. The GMM not only fits the distribution of outliers well, but also provides the probability that a station will be interrupted at a given interval. To convert the disruption probabilities into final detection decisions, the optimal probability threshold (i.e., minimum probability of observation to fall in the abnormal cluster to be called disrupted) is learned from a simulation-based method rather than subjectively determined.

Second, by merging the detection output of the first step with the train trajectory data at the line level, propagation of the disruption across the connected stations is identified. In this way, we can identify the station with the origin of disruption (i.e., primary disruption) and the extent of the spill-over interruption on downstream/upstream platforms (i.e., secondary disruption). Our approach involves a smart screening algorithm, followed by visualisation of disruptions on the space-time diagram of train movement. These diagrams also reveal the recovery interventions performed by metro operators, such as dispatching adjustments or rescheduling. The knowledge of secondary disruptions and recovery interventions is hard to

obtain from traditional inspections or other data-driven detection methods. The former reflects the impact of the primary disruption on the service provision of downstream stations, which is essential to comprehensively evaluate line-level reliability. The latter reflects the ability of operators to restore normal services under disruptions. Quantifying recoverability plays a key role in assessing metro resilience.

In the case study, we apply the proposed method to detect service disruptions in the Hong Kong MTR and display the results of a densely used metro line. Compared to manual incident logs, we have detected all disruptions (over 5 minutes) and 96% of minor incidents (between 2 to 5 minutes). In terms of the validation via simulated detections, across all stations of the studied line under both minor and mixed disruption scenarios, the average detection accuracy is above 0.99. Specifically, the average precision is nearly uncompromised, and the average recall rate is over 0.9. The detection results contain detailed information of historical disruptions, including their occurrence time and location, lasting duration as well as the propagation of service delays along metro lines. Such information is the foundation for further research on disruption impacts and management, with which operators can optimise recovery strategies and dynamic scheduling. Accurate service delay information also improves the evaluation of service reliability.

The rest of the chapter is organised as follows. Section 4.2 presents the probabilistic framework used to detect train service interruptions. Section 4.2.1 demonstrates GMM, followed by Section 4.2.2 which explains the screening algorithm and space-time diagram approach to identify secondary disruptions and recovery interventions. In Section 4.3, we present an empirical analysis to detect disruptions in the Hong Kong MTR. Results are then discussed in Section 4.4. Finally, conclusions and future work are summarised in Section 4.5. The sensitive analysis of detections is presented in Appendix A.

4.2 Methodology

Our detection approach has two stages. First, in Section 4.2.1, we demonstrate how Gaussian mixture models (GMMs) can be applied to probabilistically detect platform-level metro service disruptions. Second, in Section 4.2.2, we analyse the line-level disruption propagation to

identify the primary source and secondary spread of the disruption. Figure 4.2 details all steps of the proposed detection framework.

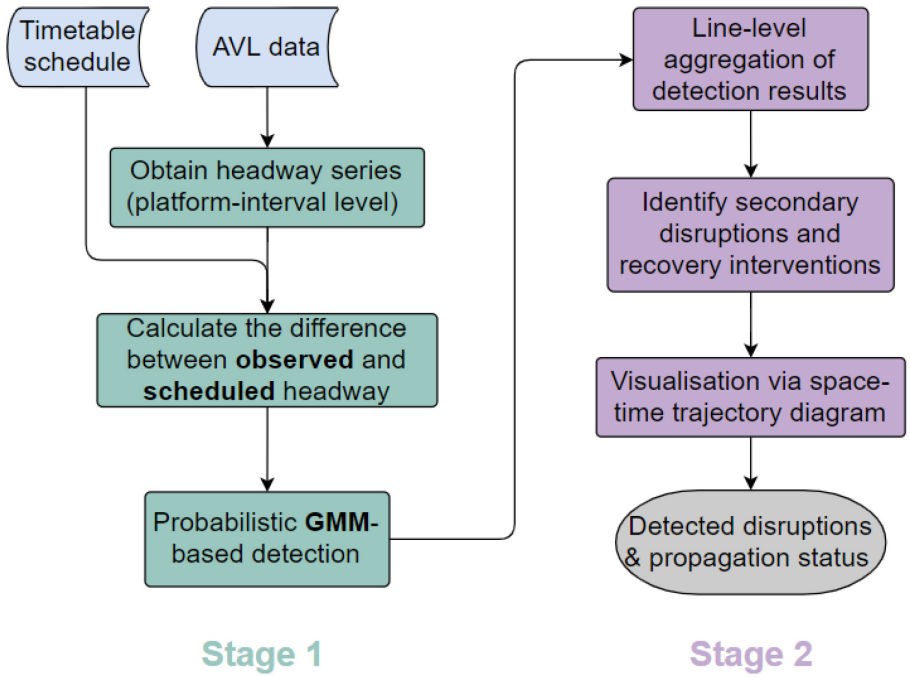


Figure 4.2: Flowchart of the chapter's methodological framework

4.2.1 Probabilistic detection with Gaussian mixture models

This section describes GMMs and motivates their application in detecting abnormal headways. The train service analysis to extract the observed headways from the AVL data and scheduled headways from the timetable is presented in the first subsection. GMM specification, its parameterisation, and the maximum-likelihood estimation are described in the next subsection. The procedure of applying the GMM-based model to detect service disruptions is presented in then the subsequent subsection. Finally, the last subsection details a simulation-based algorithm to derive optimal thresholds of disruption probabilities to designate a station to be disrupted.

Train service analysis

In urban metro systems, train services are planned according to a timetable defined by operators. Headway, the inverse of train frequency or the distance between two successive trains

measured in time or space, is the key measure of service quality.[5] Under regular operating conditions, the observed headway is similar to the scheduled headway with some natural deviation. However, when train services are interrupted, the difference between the observed and scheduled headway is likely to exceed an acceptable level. Thus, abnormal (overlong) headway can be regarded as an indicator of the service disruption occurrence.

As shown in Figure 4.2, the observed headway series (denoted by H) are extracted from the AVL data for each platform. The scheduled headway series (S) are obtained from service timetables. We define the gap between observed and scheduled headways by

$$\boldsymbol{G_{dt}^{alp}} = \boldsymbol{H_{dt}^{alp}} - \boldsymbol{S_{dt}^{alp}}, \tag{4.1}$$

where vector $\boldsymbol{H_{dt}^{alp}}$ denotes the observed train headways on platform p ($p = 1, \dots, P$) of line l ($l = 1, \dots, L$) at station a ($a = 1, \dots, A$) on a given day d ($d = 1, \dots, D$) during time interval t ($t = 1, \dots, T$). The vector $\boldsymbol{H_{dt}^{alp}}$ stacks the headways of trains departing from platform p in time interval t. We use the same indices for $\boldsymbol{S_{dt}^{alp}}$ and $\boldsymbol{G_{dt}^{alp}}$. Considering that headways vary at different stations and lines during a given time interval and platforms can be located at the corresponding station and line, we focus on identifying service disruptions at platform-interval level. We merge multiple days of observations by segmenting the day into multiple predefined time intervals. The length of interval can be determined based on the magnitude and variation in scheduled headway on the given platform. One should specifically ensure that within each time interval, (i) there are adequate headway observations, and (ii) the corresponding scheduled headway remains close in the interval. On these grounds, we set the interval length of 30 minutes in the case study of the Hong Kong MTR (that is, $T = 36$ for 18 service hours). For the entire study period of D days,[6] the platform-interval level headway deviation data is stacked in a vector as follows:

$$\boldsymbol{G_{t}^{alp}} = \{\boldsymbol{G_{1t}^{alp}}, \boldsymbol{G_{2t}^{alp}}, \dots, \boldsymbol{G_{Dt}^{alp}}\}. \tag{4.2}$$

Thus, the detection problem involves the identification of abnormal headway deviations for each platform-interval across all days.

[5] In this research, we define time as the unit of measurement, so headway here represents the tip-to-tip time from the departure of one train to the departure of the next train on a platform.

[6] Please note that we only focus on weekdays in this study, which have similar headways across different days of the week.

GMM and disruption identification

The GMM is a probabilistic model to identify subpopulations or clusters of observations with similar characteristics within a population (Peel and McLachlan, 2000). For example, in the context of this study, GMM can endogenously identify clusters of regular and abnormal headway deviations. There are two motivations behind using GMM to detect abnormal (overlong) headways. First, without true labels (normal and abnormal) on the headway deviation data, this detection problem is an unsupervised learning problem. Moreover, due to the nature of unexpected incidents or failures, the headway data is expected to contain relatively fewer anomalous observations (i.e., small subpopulation with abnormal characteristics). Since higher headway deviations indicate more severe disruption, such monotonicity can assist in naturally grouping even fewer abnormal headways into the right-most cluster (that is, with the highest cluster mean; see Figure 4.3). Thus, GMM can address this unsupervised learning problem by systematically separating abnormal headways from other clusters of regular headways. Second, GMM is probabilistic, and thus we can obtain the probability of each headway observation to belong to the right-most cluster. In other words, the GMM-based detection method provides the probability of a platform being disrupted during a specific interval.

Compared to the deterministic detection methods based on empirical rules (e.g., three standard deviations away from the mean), the GMM-based method does not require the analyst to define subjective thresholds to characterise a headway to be abnormal. However, the threshold on the probability of a headway gap belonging to the right-most cluster is required in the GMM-based model to identify the disrupted headways. Such thresholds can be learned through a semi-synthetic simulation (see **Selecting parameters through simulation** for details). The data-dependent probabilistic thresholds in GMM perform better than the subjective thresholds in deterministic models in detecting minor abnormalities because the former is normalised but the latter suffers from scale of standard deviation. For example, when the standard deviation of the observed headway is longer than 5 minutes, minor or moderate service interruptions (i.e., those under 10 minutes) cannot be identified using a deterministic rule in which headways beyond two standard deviations of the mean are designated as outliers.

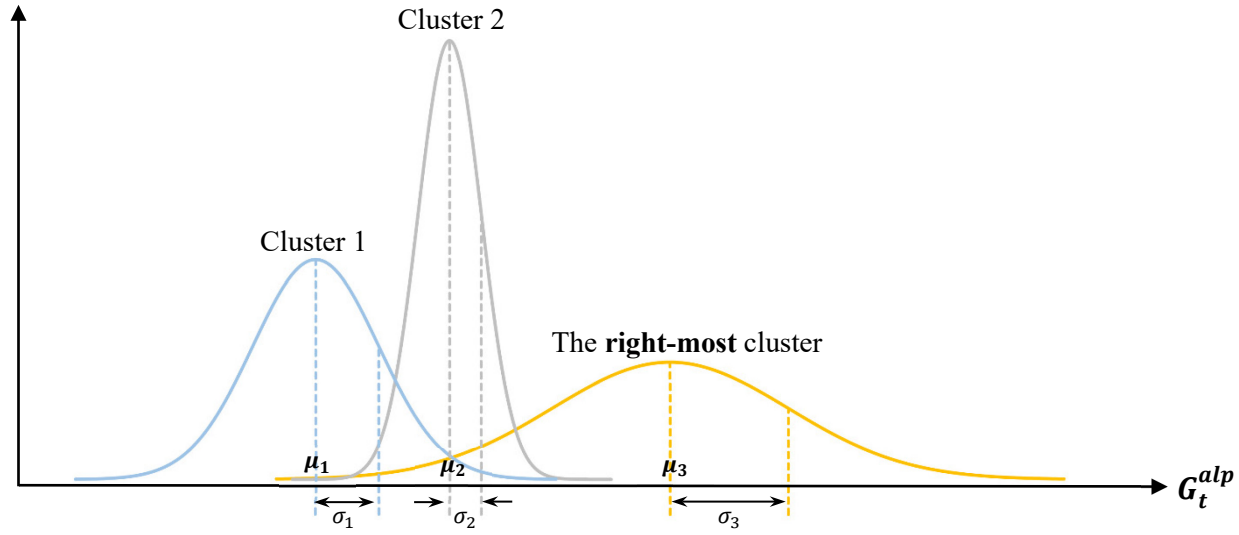


Figure 4.3: An illustration of the right-most cluster in GMM

We succinctly discuss one-dimensional GMM formulation in the context of this study. A Gaussian mixture density of an observation G_{it}^{alp} is a weighted sum of M component densities:

$$p\left(G_{it}^{alp}\right) = \sum_{j=1}^{M} w_j \, p_j\left(G_{it}^{alp}\right), \tag{4.3}$$

where G_{it}^{alp} $(i = 1, \dots, N)$ is an observation of vector $\boldsymbol{G_t^{alp}}$ (the headway deviations belonging to a specific platform-interval across all days), w_j is the mixture weight of the j^{th} component, and $p_j(.)$ is the Gaussian density of the j^{th} component with mean μ_j and variance σ_j^2:

$$p_j(G_{it}^{alp}) = \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{\left(G_{it}^{alp}-\mu_j\right)^2}{2\sigma_j^2}\right). \tag{4.4}$$

The mixture weights satisfy the following conditions:

$$\sum_{j=1}^{M} w_j = 1 \text{ and } 0 \le w_j \le 1. \tag{4.5}$$

The log likelihood function of observation for platform-interval can thus be written as:

$$\log p\left(\boldsymbol{G_t^{alp}}\right) = \sum_{i=1}^{N} \log\left\{\sum_{j=1}^{M} w_j \, p_j\left(G_{it}^{alp}\right)\right\}. \tag{4.6}$$

Here $\left\{\mu_j, \sigma_j, w_j\right\}_{j=1}^{M}$ are the identified parameters in GMM, which are obtained by maximising the loglikelihood presented in Equation (4.6). Since direct maximisation of the loglikelihood is cumbersome, we resort to an expectation-maximisation (EM) algorithm to maximise the loglikelihood (Dempster et al., 1977; Bansal et al., 2018). The probability of the headway

difference $\left(G_{it}^{alp}\right)$ belonging to the j^{th} component is obtained using the estimated parameters and Bayes rule.

The procedure of applying the GMM-based detection model

Figure 4.4 displays the procedure of GMM-based detections. Note that the distribution of the right-most cluster (with the highest mean headway gap) depends on the variation in input headway deviations $\boldsymbol{G_t^{alp}}$. Since disruptions do not occur often, composition of the headway gap data for different platform-intervals can inherently be of two types – (i) all regular observations with the headway deviations close to zero; (ii) both normal and abnormal headway deviations. While training GMM with the first type of data, the mean and standard deviation of the right-most cluster is likely to be small (e.g., zero to one minute). However, since we always focus on the right-most cluster to identify the service disruptions, the right-most cluster with a narrow tail and negligible mean headway deviation can be wrongly identified as a cluster of disrupted instances. GMM is likely to perform well for the second type of data, but we also want to circumvent the false detection of abnormal headways (that is, disruptions) in the first type of data.

To avoid the false detection of disruptions, we first check whether the data have potential disrupted observations (type II) or not (type I). If the maximum of the headway gap data is lower than the acceptable headway deviation, we conclude that the platform experiences no disruption during a specific interval (type I) and therefore GMM estimation is not required. It is worth noting that the acceptable headway deviation for each platform during a specific interval depends on the scheduled headway. For instance, if the scheduled headway at a metro platform in peak hours is 2 minutes, then service delays of approximately 1 to 2 minutes are acceptable, but 10-minute delays are not. On the other hand, at another platform, where the scheduled headway is 20 minutes, the 10-minute delay can be treated as an acceptable headway deviation. The acceptable headway deviation also depends on the metro operator's aspirations to provide a reliable service. We discuss the selection of acceptable headway deviations in Section 4.4. GMM is applied only on type II datasets.

In addition to the probability threshold for the right-most cluster, the number of clusters (M) needs to be selected to apply GMM for disruption detection. The number of subpopulations is often selected based on the Silhouette score and Bayesian information criterion (Lord et al.,

2017), but in the context of this research, the choice of the number of clusters should not depend on statistical criteria. Specifically, we should ensure that the right-most cluster only contains the potential anomalies to avoid or minimise false disruption detections (see Figure 4.3). Thus, the number of clusters and the optimal threshold on the probability of a headway deviation belonging to the right-most (abnormal) cluster are selected using semi-synthetic simulations. The simulation design is presented in the next subsection.

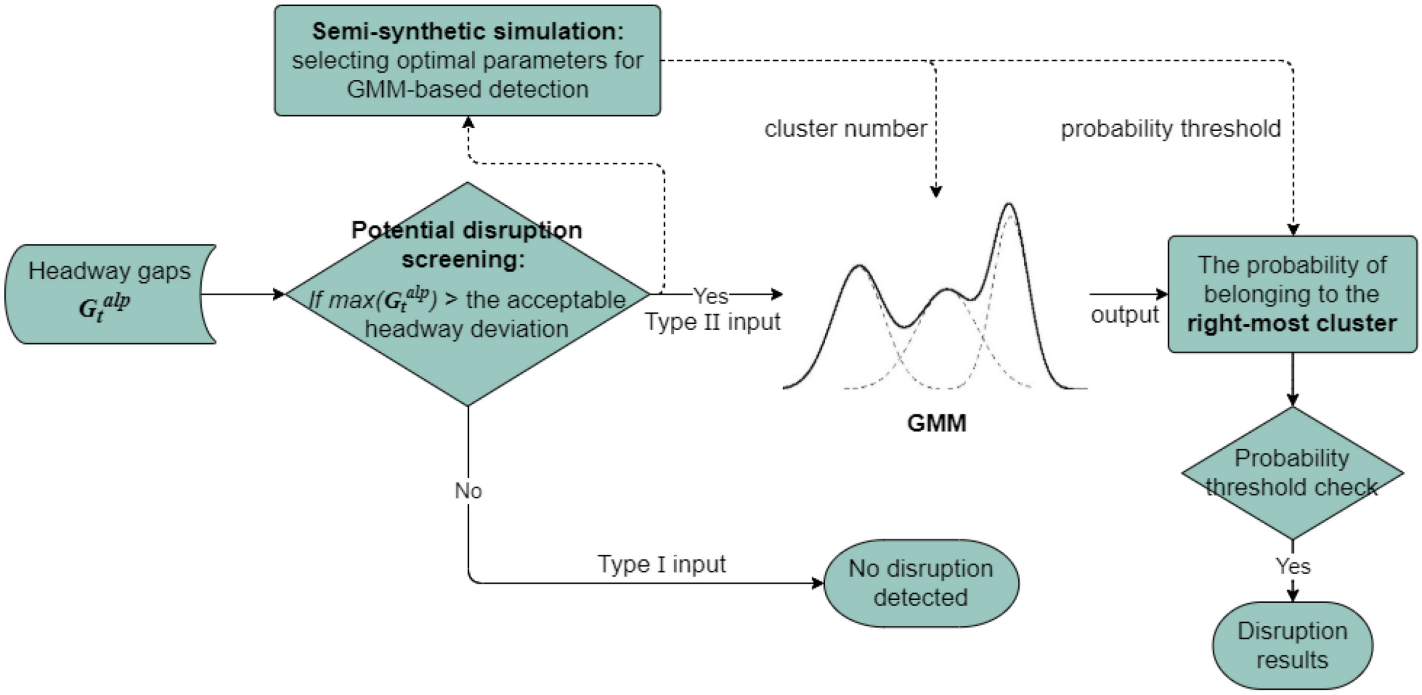


Figure 4.4: The procedure of applying the GMM-based detections

Selecting parameters through simulation

Instead of subjectively selecting the probability threshold and the number of clusters, we adopt a simulation-based grid-search method. The main idea is to use the empirical distribution of the headway deviation data to simulate new data and label a certain proportion (e.g., 1%-5%) of the simulated headway deviation as disrupted observations. The data-generating process can be changed by varying the percentile of empirical deviation data used for the simulation (see Section 4.4.2 and Appendix A). With the labelled disruption data, the problem is translated into a supervised learning problem and the GMM's prediction accuracy can be tested under different combinations of the number of clusters and probability threshold. The proposed model selection method consists of the following steps for each platform-interval:

i). Derive a sample from the empirical cumulative distribution function of the observed headway deviations for a platform-interval, to generate the undisrupted simulation dataset. Calculate the proportion of the potential abnormal deviations (over acceptable level) in the given type II input data, and use this proportion to generate labelled disruptions.

ii). Run GMM-based detection models on the simulated headway deviation data for different number of clusters (e.g., ranging from 2 to 20) and threshold probabilities (e.g., $\{0.99, 0.98, 0.97, \dots, 0.75\}$).

iii). For each combination of the number of clusters and probability threshold, compute performance measures: precision, recall, F1 score and accuracy.[7] Precision is the ratio of correctly detected disruptions to the total detected disruptions. Recall is the ratio of correctly detected disruptions to all the labelled disruptions. F1 score is the weighted average of precision and recall. Accuracy is the ratio of correctly detected observations to the total observations. To mitigate the simulation noise, repeat step (i) and (iii) 1000 times and obtain the average value of performance measures.

iv). Now create a two-way table of performance measures with rows indicating the number of clusters, and columns indicating the optimal threshold probability and the corresponding average value of performance measures. This two-way table is used to identify the optimal number of clusters.

v). Finally, select the best combination of the two parameters (cluster number and threshold probability) obtained in step (iv), and use them to conduct the GMM on the observed deviations data.

4.2.2 Secondary disruption and recovery intervention identification

The GMM-based detection method provides information on the location, time and duration of disruptions. In this section, we use this output to find the linkages between the detected

[7] Formulas of the performance measures:

$Precision = \frac{True\ positives}{True\ positives + False\ positives}$, $Recall = \frac{True\ positives}{True\ positives + False\ negatives}$,

$F1\ score = \frac{2 \times Recall \times Precision}{Recall + Precision}$, $Accuracy = \frac{True\ positives + True\ negatives}{True\ positives + False\ positives + True\ negatives + False\ negatives}$.

disruptions at consecutive platforms along a metro line. We categorise disruptions into two types: primary disruption and secondary disruption. Primary disruption means that the service interruption is *originated* at the given platform during a specific period. In contrast, secondary disruption at a platform is caused by a primary disruption at one of the upstream or downstream[8] platforms along the metro line. As discussed earlier, this is the first novel analysis to apply such categorisations and to identify the spillover of interruption status. Since metro systems reboot every day, we analyse metro line operations on a specific day of disruption(s) using the following steps:

i) Pool the GMM-based detection results on all platforms of a metro line (with the same direction of train services) on the specific day.

ii) Sort all disruption records based on the start time of disruptions. Mark the first record as a *primary disruption*. For the next record, if (i) the platform is downstream to the primary disruption location (follow the train direction); (ii) the start time of the disruption is slightly later; and (iii) the train ID and trip ID are the same, this record is marked as a *secondary disruption*. However, for the next record, if the platform is downstream to a primary disruption location (follow the train direction) and the start time of disruption is slightly later, but the train ID and trip ID are not the same, this record is marked as a *secondary disruption with an intentional dispatching intervention* from the operator. Such interventions aim to restore normal services and reduce the impact of delays on passenger waiting time. Repeat this process until the upcoming record breaks the spatiotemporal continuity of start time and downstream location conditions.

iii) Repeat step (ii) until all disruption records are marked as either primary disruption, secondary disruption, or secondary disruption with intentional dispatching intervention.

iv) Merge the daily disruption records obtained in step (iii) with the corresponding train trajectory data to visualise disruptions and the train movement using the space-time diagram.

[8] Train delays may spread in the opposite direction, upstream to the disrupted station as well, due to queueing or dispatching interventions.

4.3 Data and case study

The Mass Transit Railway (MTR) is the urban and suburban rail operator of Hong Kong, and the member of the Community of Metros facilitated by the Transport Strategy Centre at Imperial College. In 2019, the MTR served 95 stations and 11 lines, connecting the urbanised areas of Hong Kong Island, Kowloon, and the New Territories. The MTR network, as shown in Figure 4.5, is one of the busiest metro systems in the world, by the end of 2019 it had carried over 1.9 billion passengers (Mass Transit Railway, 2019). In this PhD, all the proposed empirical research has been tested on the Hong Kong MTR system.

In this chapter, to illustrate the detailed process of disruption detection, we select a densely used line to carry out the case study. This line has 16 stations with tracks of a total length of 16.9 km. Being the link between major commercial centres of Mong Kok, Tsim Sha Tsui and the Central District, it is constantly busy and crowded. We detect service disruptions that occurred in both upward and downward directions. During the study period (54 weekdays from 01/01/2019 to 31/03/2019, excluding holidays and days of incomplete data), the scheduled headway of the given line ranges from 2 to 10 minutes with an average of around 3 minutes. We choose 30 minutes as the interval to group the headway data of each platform, which also ensures that each platform-interval group has sufficient headway observations. The daily service time of the selected line starts at 6:00 and ends at 24:00, which is thus divided into 36 intervals to conduct GMM-based detections. Therefore, taking account of 16 stations with two platforms and 36 intervals for each platform, our dataset is aggregated into a total of 1152 platform-interval groups. We also identify secondary disruptions and recovery intervention from metro operators. Finally, information on the detected disruptions is collected to support the remaining studies in this PhD.

The following data are used to detect and evaluate the service disruptions. We conducted data processing and analysis using open-source R software (version 4.1.1).

Automated vehicle location (AVL) data

The AVL data from 01/01/2019 to 31/03/2019 are provided by the MTR. Public holidays including the New Year and the Spring Festival are excluded. We consider this duration as our study period. The AVL data contain information on train ID, trip ID, the timestamp of train movements (including precise departure and arrival times), and the location of train movements (including station, line and directions). The resolution of time stamps exacts to one second. By

using the AVL data, we can extract headway series from the consecutive train movements on each platform.

Timetable schedules

The scheduled arrival and departure time of train services on the selected line, provided by the MTR. We utilise this information to extract scheduled headways.

Incident logs

The manual inspection record of incidents, including information such as occurrence time, location, cause and duration of disruptions, provided by the MTR. Incident logs are used to validate our detection results.

Pseudonymised smart card data (SCD)

The SCD contain information on the time and location of tap-in and tap-out transactions throughout the system, recording individual trips. In this research, the role of SCD is constrained to illustrating the limitations of the demand-based disruption detection methods.

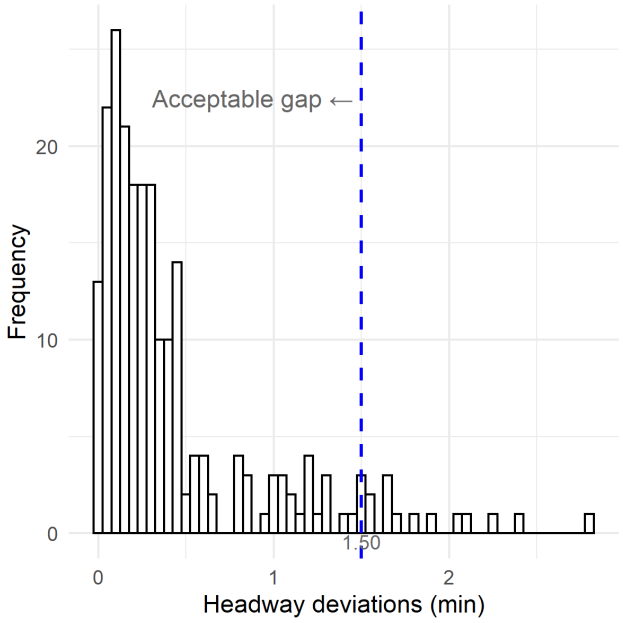
Figure 4.5: The MTR system map

4.4 Results and discussion

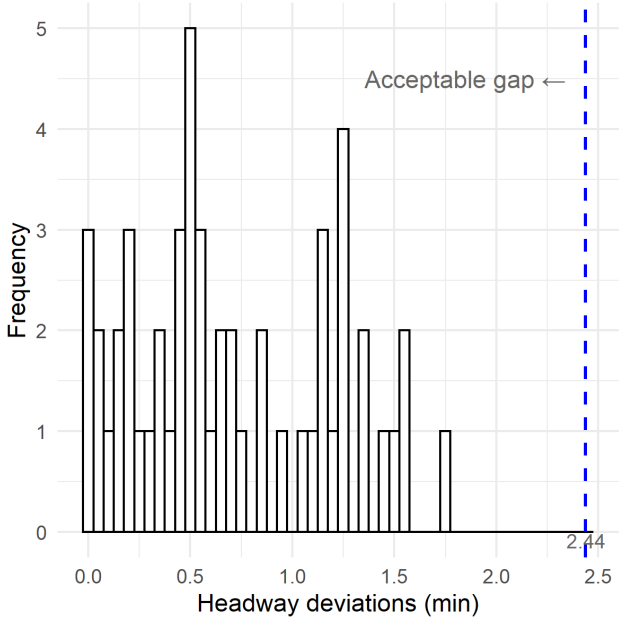
The results are presented in four steps. First, we illustrate the method of screening input data and how to identify platform-intervals where the possibility of disruptions cannot be excluded. Second, we showcase the simulation approach to the optimisation of key hyper-parameters of the proposed GMM method. The third subsection presents the final outputs of the GMM disruption detections. The results in the first three steps are presented through a sample dataset from a randomly selected platform of the studied metro line, in the initial stage of the morning peak period. Finally, Section 4.4.4 demonstrates disruption propagation through the identification of secondary disruptions and dispatching interventions along the entire line.

4.4.1 Input data check: screening potential disruptions (Type II)

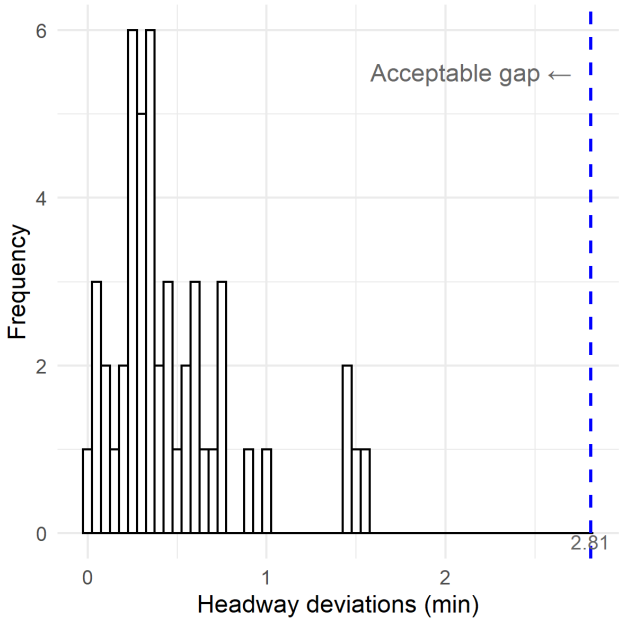
At an example station, northbound, 7:30–8:00 a.m., the scheduled headways on this platform-interval range between 2 and 4 minutes. To guarantee the reliable service of early peak hours in the morning, we determine that the acceptable headway deviations should be lower than 75% of the scheduled headway. This threshold is set arbitrarily, based on the intuition that if no train is delayed by more than another scheduled headway, including a 25% safety gap, then it is very unlikely that significant disruptions happened within the 30-minute interval. Figure 4.6(a) to 4.6(d) display the histogram of the observed headway deviations under different scheduled headways. The dashed lines represent the 75% boundaries defined above. In plots (a) and (d), there are observed headway deviations above the acceptable level, which means that the input is type II and we cannot exclude the presence of disruptions. Therefore, we proceed to the next step of our analysis.



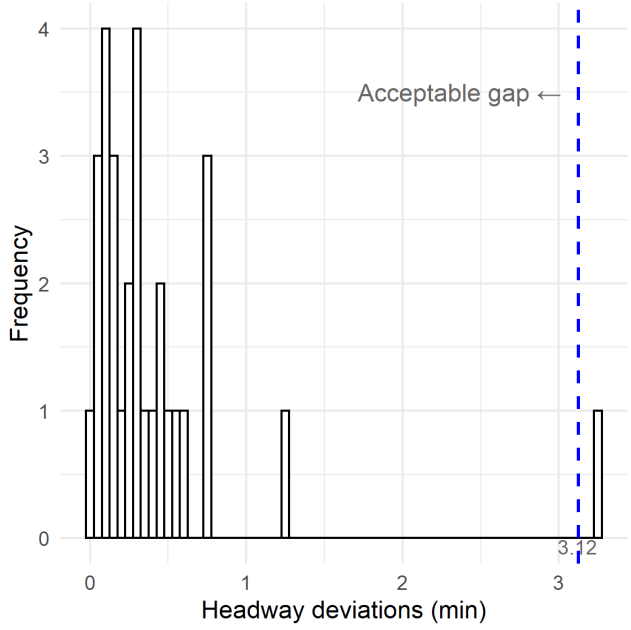
(a) Scheduled headway = 2 min



(b) Scheduled headway = 3.25 min



(c) Scheduled headway = 3.75 min



(d) Scheduled headway = 4.17 min

Figure 4.6: The histogram of observed headway deviations for different scheduled headways from the given platform-interval of the example station

4.4.2 Optimal number of GMM clusters and probability threshold

Before applying detection models, we run semi-synthetic simulations to obtain the optimal values of two critical parameters: the number of clusters in our GMM approach and the probability threshold above which observations in the right-most cluster indicate a disruption (see Section 4.2.1).

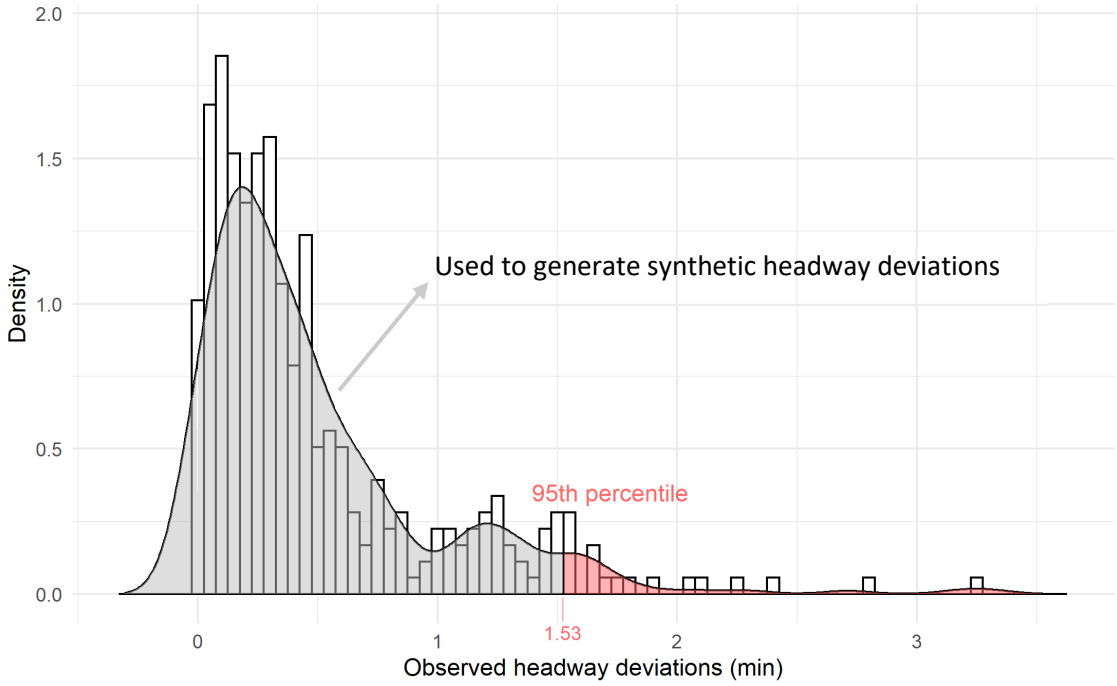


Figure 4.7: Histogram of overall headway deviations observed from the given platform-interval. Observations below the 95th percentile are used to generate the simulation data

Figure 4.7 shows the overall distribution of headway deviations regardless of schedules. The graph shows that 95% of the headway deviations are less than 1.5 minutes. We first generate a synthetic dataset of *undisrupted* headway deviations, which is drawn from the empirical distribution of the observed deviations truncated at the 95th percentile. We ensure that the sample size of the synthetic data matches with that of the empirical data. Subsequently, a certain proportion of disruptions are generated based on a log-normal distribution. The mean of the lognormal distribution μ_{syn} is set according to the scheduled headway, and the standard deviation σ_{syn} is set to achieve the varied lengths of disruptions. These disruptions are then randomly allocated and added to the original synthetic deviations, thus forming the *disrupted* headway deviations.

In the present example, 5% of the observed headway deviations are above the acceptable level. We choose this proportion to generate disruptions. For the distribution of disruption durations, the μ_{syn} is set as 1.2 times the scheduled headway, and the standard deviation σ_{syn} is set to be 0.3. Finally, the disrupted synthetic deviations range between 1.5 and 6 minutes. Figure 4.8 displays the empirical distribution of an example synthetic dataset. The grey bars represent undisrupted headway deviations while the orange bars indicate disrupted data items. In Appendix A, we perform robustness checks regarding the chosen percentile for the sampling of undisrupted observations. We also perform sensitivity analysis relative to the proportion of disruptions in the data generating process and different station-time-interval pairs.

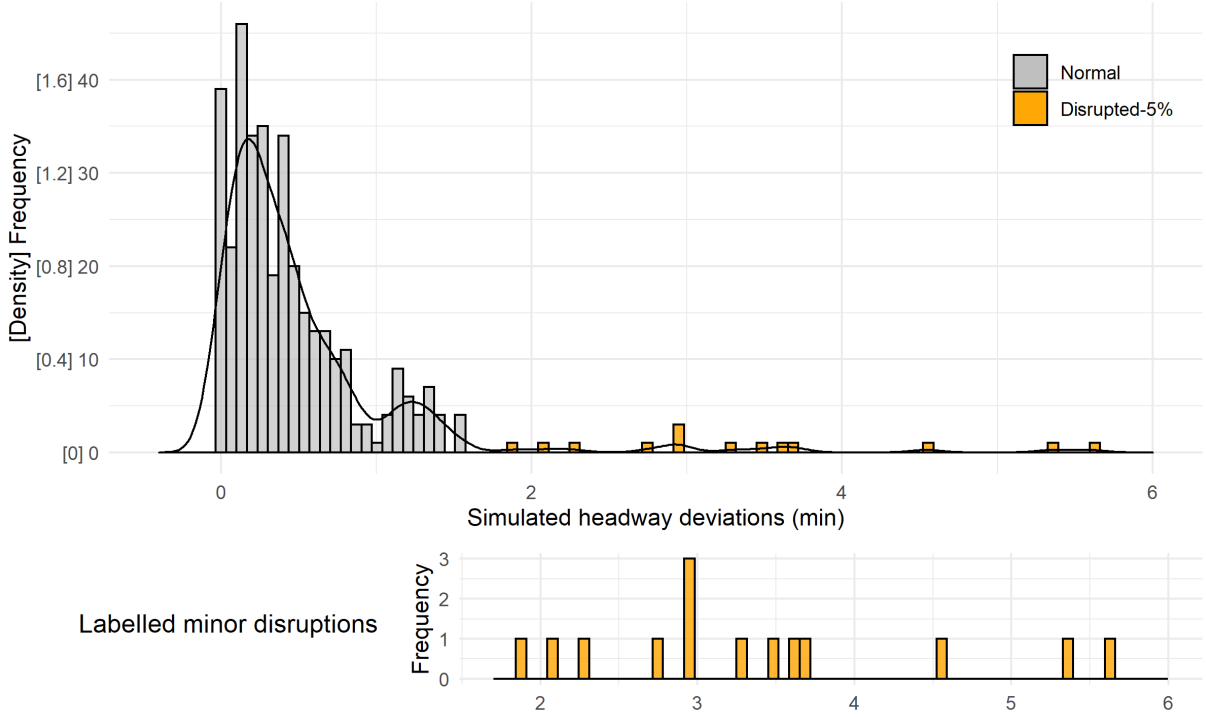


Figure 4.8: Histogram of a sample synthetic headway deviations for the given platform-interval. The proportion of the disrupted synthetic deviations is 5%

Table 4.1: The two-way table of simulation performance (averages of 1000 runs) and optimal GMM parameters

Number of clusters (M)	Precision	Recall	Accuracy	Optimal threshold
2	0.152	0.998	0.725	1.000
3	0.176	0.996	0.769	1.000
4	1.000	0.813	0.991	0.933
5	1.000	0.871	0.994	0.958

6	1.000	0.896	0.995	0.960
7	1.000	0.907	0.995	0.965
8	1.000	0.923	0.996	0.974
9	1.000	0.930	0.997	0.982
10	1.000	0.931	0.997	0.980
11	1.000	0.914	0.996	0.975
12	1.000	0.937	0.997	0.984
13	1.000	0.946	0.997	0.994
14	1.000	0.943	0.997	0.991
15	**1.000**	**0.947**	**0.997**	**0.994**
16	1.000	0.912	0.996	0.979
17	1.000	0.903	0.995	0.968
18	1.000	0.890	0.995	0.959

With pre-defined labels of service status, in simulation we have transformed the disruption detection task into a supervised learning problem. For a wide range of possible combinations of the GMM cluster number and the probability threshold, we calculate the precision, recall, and accuracy of detection. The simulation is then repeated 1,000 times to obtain the average performance metrics for every combination of parameters. Table 4.1 summarises these metrics (each row represents the average results of 1000 simulations) for the dataset visualised above. Due to limited space, three performance measures (precision, recall, accuracy) and the optimal probability thresholds for the given cluster number are compared in the table, with cluster numbers ranging from 2 to 18.

We find that, when the number of clusters is set to 15 and the probability threshold is 0.994, both the detection precision rate and overall accuracy reach their maximum values (above 0.997). The balance between the precision and recall rate also reaches the best. Figure 4.9 shows how the right-most cluster changes under different choice of cluster numbers. As the GMM clusters increase from 2 to 30, the right-most cluster gradually shifts to the right of x-axis with higher mean and lower standard deviations. Meanwhile, the probability of all disrupted headway deviations belonging to the right-most cluster keeps increasing until the number of clusters reaches 15. When the cluster number continues to grow, such probability starts to drop as the less spread right-most cluster tends to cover fewer disruptions. Thus, in the formal GMM-based detections, the optimal 15 clusters and the 0.994 probability threshold are applied for the platform-interval we consider in this example.

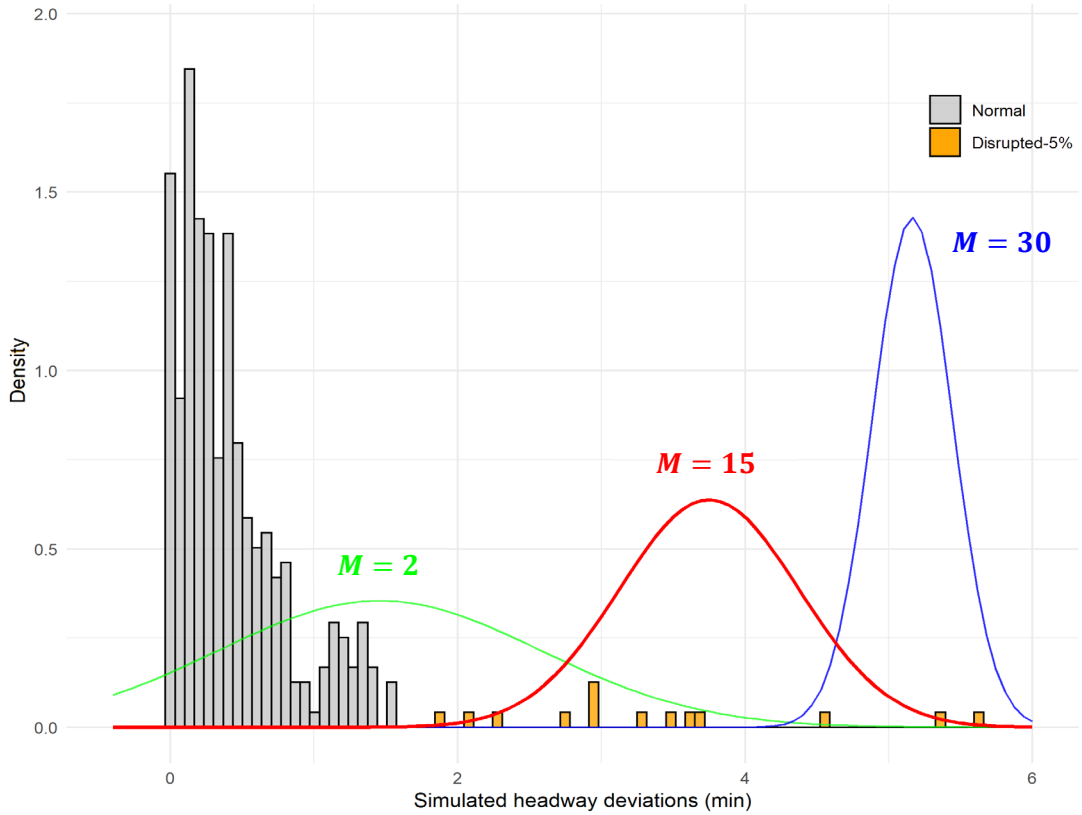


Figure 4.9: Changes in the right-most clusters of the estimated GMM, under different number of clusters (M). Each solid line represents the distribution of the right-most cluster for a given cluster number

4.4.3 GMM detection results

Figure 4.10 presents the GMM detection results of the above example in the form of a three-dimensional plot. The y-axis of the 3D plot represents observed headway deviations in the given platform-interval, the x-axis represents the corresponding scheduled headways, and the z-axis represents the probability of belonging to the right-most cluster (refer to Section 4.2.1). The colour of scatter points indicates detection decisions. The grey points refer to normal headway deviations that are within the corresponding acceptable levels. Their disruption probabilities are less than 22%. The yellow points tend to include all possible outliers, with the disruption probability ranging from 10% to 99%. To achieve the highest detection accuracy, we rely on the optimised probability threshold. In this case, only two observations (highlighted in purple) are above 0.994, and they are finally identified as disruptions.

In terms of the entire line, we compare our detection results with manual incident logs from the Hong Kong MTR. For medium to severe interruptions that are between 5 minutes and

several hours long, all reported disruptions have been detected by the proposed GMM method. For minor service interruptions that range from 2 minutes to 5 minutes, 96% of them have been successfully identified. The remaining undetected minor incidents are generally of very short duration (just over 2 minutes), especially when compared to their scheduled headway. Our data-driven detection also provides more supplementary results that may have been omitted in human inspections.

As for the validation via simulated detections, in all stations of the selected line under both minor and mixed disruption scenarios,[9] the average detection accuracy is above 0.99. Specifically, the average precision is nearly uncompromised, and the average recall rate is over 0.9. The corresponding sensitivity analysis is demonstrated in Appendix A.

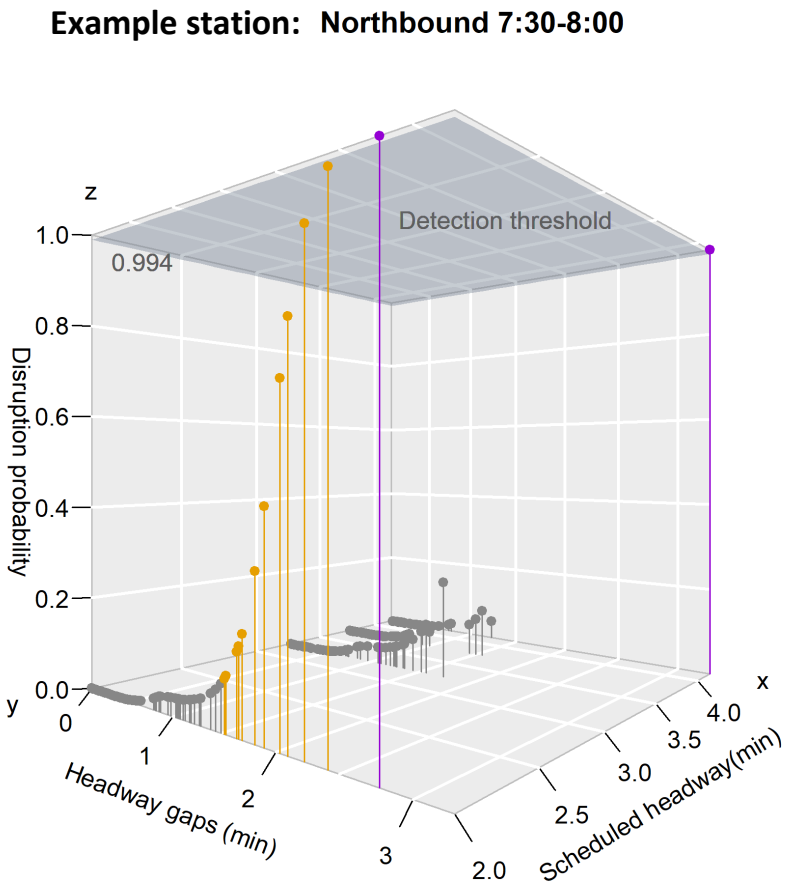


Figure 4.10: Final detection results of the given platform-interval: probabilities of belonging to the disrupted cluster and the optimal threshold.[10] The purple dots represent the identified disruptions

[9] The simulated minor disruptions range from 1.5 to 8 minutes. The mixed disruptions are referred to as a mixture of minor interruptions and severe interruptions (over a few hours).
[10] The same detections were applied to all platform-intervals in the selected metro line.

4.4.4 Secondary disruptions and recovery interventions

In this subsection, we demonstrate how to apply the algorithm presented in Section 4.2.2 to identify secondary disruptions. Since the MTR system closes after midnight and reopens the next morning, the identification is implemented on a daily basis. After the first step of pooling disruptions into the level of the entire line and partitioning based on date, Table 4.2 shows a sample of the detected disruptions during two off-peak periods (10:30 – 11:30 and 20:00 – 21:00). Considering the average scheduled headway of the line is around 3.5 minutes in these periods, we focus on detected disruptions over 4 minutes.

Table 4.2: A sample of selected disruption records with the corresponding category identification results

Disruption ID	Start time	Station ID	Train ID	Duration (min)	Category
18	10:48:44	2	42	00:09:48	Primary
19	10:52:03	5	53	00:04:33	Intervention
20	10:52:05	3	42	00:10:09	Secondary
21	10:55:57	4	42	00:08:22	Secondary
22	10:59:41	5	42	00:06:58	Secondary
23	11:02:06	6	42	00:06:35	Secondary
24	11:03:55	7	42	00:06:34	Secondary
25	11:05:05	11	53	00:04:59	Intervention
26	11:05:50	8	42	00:06:38	Secondary
27	11:07:30	9	42	00:06:46	Secondary
28	11:07:53	12	53	00:05:09	Intervention
29	11:09:14	10	42	00:07:02	Secondary
30	11:09:51	13	53	00:05:09	Intervention
31	11:11:32	14	53	00:05:08	Intervention
32	11:13:09	11	42	00:04:57	Secondary
33	11:13:33	15	53	00:04:58	Intervention
34	11:14:44	16	53	00:04:58	Intervention
35	11:16:07	12	42	00:04:46	Secondary
36	11:18:05	13	42	00:04:47	Secondary
37	11:19:45	14	42	00:04:42	Secondary
38	11:21:36	15	42	00:04:47	Secondary
39	11:22:46	16	42	00:04:47	Secondary
40	13:32:14	2	40	00:04:04	Primary
…	…	…	…	…	…
118	20:24:10	2	70	00:06:49	Primary
119	20:26:17	4	44	00:04:43	Intervention
121	20:27:38	3	70	00:07:12	Secondary
122	20:28:10	5	44	00:04:58	Secondary

In the second step we sort these detection records by start time, as shown in Table 4.2. Figure 4.11 visualises the disruptions and their categories with the corresponding train trajectory data in a space-time diagram. The first record (Disruption 18) is marked as the initial primary disruption. When moving to the next record, Disruption 19 starts slightly later than the primary one (within the regular time of a full journey) and the platform location is downstream, but their train IDs are not the same. Thus, this record is marked as a secondary disruption with an intentional dispatching intervention from the operator. Then, moving to the third record, compared with the primary one, Disruption 20 satisfies all three conditions of a secondary disruption; it starts later, at a downstream station, with the same train ID. We repeat the above procedure until we encounter a new record that breaks the temporal and spatial proximity. For instance, after Disruption 39, the location of Disruption 40 is once again at Station 2 and it occurs nearly two hours later. In this case, Disruption 40 will be marked as a primary disruption again. We repeat the screening steps until all the records are processed.

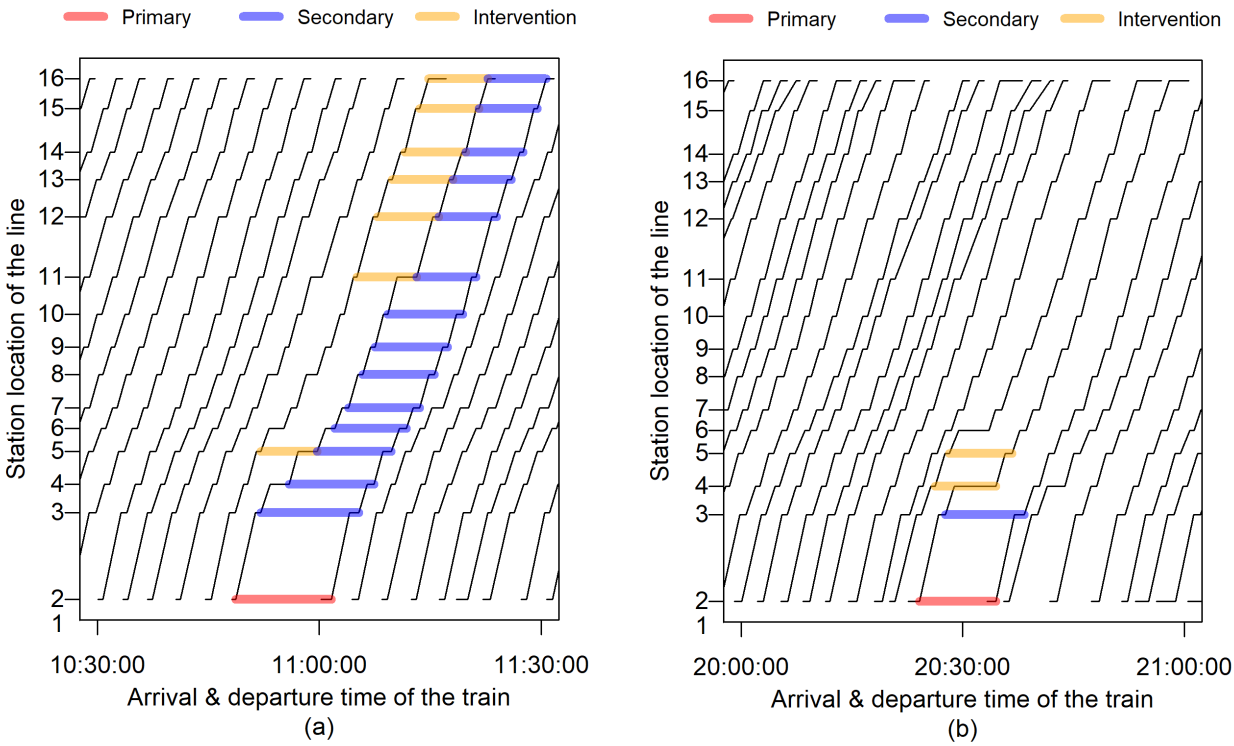


Figure 4.11: Spatial-temporal train movement diagram with detected disruptions and their categories. The propagation process of two primary disruptions

The identification results within the two sample periods are showed separately in Figure 4.11. The horizontal axis represents the arrival and departure time of trains at each platform,

while the vertical axis shows the location of and distance between the stations along the line. The black solid lines are the trajectories of train movement, and the bold lines are detected disruptions. We confirm that both detections and secondary identification results match well with train trajectories, which is in line with the law of interruption propagation. For example, in Figure 4.11(a), a primary disruption occurred at 10:48:44 at Station 2. On the one hand, this disruption spreads downstream along the line until the terminal station. On the other hand, metro operators act promptly at Stations 4, 5, and 11 to increase the dwell time of the last train preceding the disrupted one, thus avoiding further bunching effects. Due to these interventions, after Station 5, the disrupted train does not accumulate further delays. Similarly, in Figure 4.11(b), a primary disruption occurred at 20:24:10 at Station 2. Then, immediate interventions take place at Stations 4 and 5 to slow down the previous train and relieve the delayed one from excessive passenger load. No further delays are identified after the primary disruption spreading to Station 5, and the train services return to normal.

This visual analysis indicates that the proposed detection framework is valid and highly effective for identifying disruptions and their categories. Furthermore, Figure 4.11 also illustrates how identifying secondary disruptions can contribute to practical metro operations. In the space-time diagram of train movements, by labelling the secondary disruptions due to interventions, operators can easily locate the interventions used for mitigating delays, such as adjusting the dwell time of upstream trains. More importantly, with automated disruption classification, the operator can disentangle the frequency and severity of primary disruptions from subsequent time loss due to delay propagation and dispatching measures. This information is essential for both preparing recovery plans and analysing the resilience of metro systems.

4.5 Conclusions and future work

Service disruptions cause various challenges in urban metro systems, including delays, crowding, and declining passenger satisfaction. Operators need to monitor disruption occurrences closely in order to reduce their detrimental effects. With accurate information on the location, time, duration, and propagation process of disruptions, they can comprehensively assess the reliability and resilience of metro systems. Thus, the detection of service disruptions is a prerequisite of any further research on disruption management.

This research proposes a novel, probabilistic, unsupervised clustering framework to quantify the probability of an observed train headway being identified as abnormal. In contrast to traditional manual inspections and other detection methods based on social media data or smart card data, which suffer from human errors, limited monitoring coverage, and potential bias, our approach uses information on train trajectories derived from automated vehicle location (train movement) data. The proposed GMM approach assumes that the observed headway distributions are composed of a disrupted and multiple undisrupted subcomponents, where disruptions belong to the right-most subcomponent with the highest mean headway deviation. Our approach estimates the probability that a headway deviation observation belongs to the right-most cluster. We develop a simulation algorithm to infer the threshold probability, above which the headway observations are classified as disruptions. Finally, we extend the detection framework from the platform level to entire metro lines. We distinguish three categories of service delays: primary disruptions, secondary delays of the disrupted train at downstream stations, and delays of other trains due to dispatching interventions. To the best of our knowledge, this is the first study in the literature which identifies secondary disruptions and the operator's recovery interventions using automated data and algorithms.

The proposed method is applied in the Hong Kong MTR with the case study of a densely used line. This illustrative application indicates that the detection accuracy of our method is very high. In all simulated scenarios for the entire selected line, the average precision is nearly uncompromised and the average detection accuracy is above 0.991. For minor service delays in the range of 1.5 to 8 minutes, the average recall rate is over 0.90. Even though the proposed simulation framework is based on simple assumptions and idealised conditions, these results highlight promising prospects for practical adaption.

Let us conclude this chapter by acknowledging some of the present limitations of the proposed method. In daily metro operations, service disruptions can be caused by unexpected infrastructure malfunctions (e.g., signal failures and track blockages), rolling stock breakdowns and accidents, planned maintenance works, or temporal dispatching adjustments. The fact that we cannot obtain the cause of service disruptions from automated train movement data is clearly a limitation of our data-driven detection method, as compared to manual data collection. Indeed, the main reason for this limitation is that we perform the detection only based on one data source. In line with this limitation, our future research will focus on merging more data

sources to infer the cause of disruptions, such as manual incident logs, smart card data, news, and data from social media.

Chapter 5

A causal inference approach to measure the vulnerability of urban metro systems

Transit operators need vulnerability measures to understand the level of service degradation under disruptions. This chapter contributes to the literature with a novel causal inference approach for estimating station-level vulnerability in metro systems. The empirical analysis is based on large-scale data on historical incidents and population-level passenger demand. This analysis thus obviates the need for assumptions made by previous studies on human behaviour and disruption scenarios. We develop four empirical vulnerability metrics based on the causal impact of disruptions on travel demand, average travel speed and passenger flow distribution. Specifically, the proposed metrics based on the irregularity in passenger flow distribution extends the scope of vulnerability measurement to the entire trip distribution, instead of just analysing the disruption impact on the entry or exit demand (that is, moments of the trip distribution). The unbiased estimates of disruption impact are obtained by adopting a propensity score matching method, which adjusts for the confounding biases caused by non-random occurrence of disruptions. In applying the proposed framework to the Hong Kong Mass Transit Railway (MTR), we learn that the vulnerability of a metro station depends on the location, topology, and other characteristics. The core methodology elaborated in this chapter has been published as part of

> Zhang, N., Graham, D. J., Hörcher, D., & Bansal, P. (2021). A causal inference approach to measure the vulnerability of urban metro systems. *Transportation*, 1-32.

5.1 Introduction

Disruptions occur frequently in urban metro systems, causing delays and overcrowding, which can lead to safety hazards and losses in social productivity. Operators may consider investing in new technologies to improve metro facilities and to mitigate the effect of incidents. However, it is often not known how those investments compare in achieving improvements. To facilitate project selection, metros are increasingly relying on disaggregate performance metrics that

reveal the most vulnerable parts of the network. Performance can be measured in various ways. Popular examples are risk, resilience, reliability and vulnerability related metrics. These concepts are often conflated by researchers as well as by practitioners. Interested readers can refer to Faturechi and Miller-Hooks (2015) and Reggiani et al. (2015) to understand the most agreed relationship among these concepts. In this research, we focus on the vulnerability of urban metro systems, for which the performance measures of interest are passenger demand, average travel speed and passenger flow distribution.

Since the 1990s, the concept of vulnerability has been widely used to characterise the performance of transport systems (Mattsson and Jenelius, 2015; Reggiani et al., 2015); vulnerability is often defined as a measure of susceptibility of the transport system to incidents (Berdica, 2002; Jenelius et al., 2006; O'Kelly, 2015). In this study, the vulnerability of metro systems refers to the extent of degradation in the level of service due to service disruptions. Service disruptions are defined as events that interrupt normal train operations for a specific period of time.[11] Disruptions should be distinguished from the broader term 'incidents', as incidents might not always affect services. Examples of such incidents include elevator failure or corridor congestion in metro stations. Vulnerability metrics can measure the consequences of service interruptions, in the form of performance outputs such as train kilometres, passenger volumes or the quality of travelling. For operators, such metrics have important implications in identifying weak stations or links in metro systems and efficiently allocating resources to the most affected areas. Given the rising interest in utilising vulnerability metrics in disruption prevention and management, obtaining an accurate measure of such metrics is crucial.

Traditionally, vulnerability in urban metros is investigated based on complex network theory and graph theory. Complex network theory converts metro networks into graphs, which enables the quantitative measurement of vulnerability in metro systems (Derrible and Kennedy, 2010; Yang et al., 2015; Chopra et al, 2016). The adoption of graph theory has facilitated the evolution of vulnerability indicators from simply capturing the characteristics of network topology to also considering travel demand patterns and their land use dependencies (Jiang et al., 2018). However, most of these studies rely on simulation-based approaches to quantify vulnerability in hypothetical disruption scenarios. These simulation experiments are based on assumptions, both in terms of passenger behaviour and the type and scale of disruptions (Cats

[11] Five minutes to ten minutes are commonly used thresholds to define disruptions. Different metro systems around the world adopt several thresholds, primarily based on the regular frequency of operations.

and Jenelius, 2014; Sun et al., 2015; Sun and Guan, 2016; Cats and Jenelius, 2018; Lu, 2018; Sun et al., 2018). With an empirical approach, such assumptions can be avoided, and thus more reliable metrics of vulnerability can be achieved using historical evidence.

The empirical approach is rare but not unique in the literature. The exception we are aware of is Sun et al. (2016), who first detect incidents based on abnormal ridership and use the real incidents data to assess the influence of disruptions on the Beijing Subway. However, their method has some limitations. First, they assume the occurrence of incidents to be random, which is a strict and unrealistic assumption, as we demonstrate in this study. Also, the abnormal ridership may not be a reliable indicator of incidents if the fluctuations in ridership are merely manifestations of changes in travel demand due to external factors.

This research proposes a novel alternative methodology to quantify vulnerability, by empirically estimating the causal impact of service disruptions on travel demand, average travel speed and passenger flow distribution at the station level. The application of a propensity score matching method (PSM) accounts for the non-randomness of disruptions and ensures the unbiasedness of causal estimates. What makes this method attractive is that it gives a clear criterion by which to select the control group. The PSM balances the distribution of confounding factors between the disrupted and normal units by matching them based on propensity scores. Such design ensures the estimated disruption impacts to be unbiased. We make this approach scalable for the entire network, including stations where disruptions are not observed, by predicting the level of vulnerability at these stations with an advanced machine learning algorithm. In this way, we eliminate the need for ad hoc assumptions regarding passenger behaviour and the nature of disruptions.

In this chapter, we use the Hong Kong MTR as a case study and apply the methodology with large-scale automated fare collection and incident data. The station-level vulnerability is heterogeneous across the network, depending on the considered performance metrics. In terms of the demand loss and gross speed loss (overall delay), the most affected stations are more likely to be found in Hong Kong's urban areas. When considering average speed loss (individual delay) and irregularity in relative passenger flows, the most affected stations are scattered around suburban and extended urban areas due to lack of alternative routes. These results can potentially aid the investment decisions of metro operators.

The rest of chapter is organised as follows. Section 5.2 outlines the empirical framework used to compute vulnerability metrics. More specifically, this section discusses the proposed causal inference approach to estimate the unbiased disruption impact, which is the key input in building vulnerability metrics. Section 5.3 then describes the case study and data sources. Results are discussed in Section 5.4. Finally, Section 5.5 concludes and highlights the potential avenues for future research.

5.2 Methodology

From a methodological point of view, our empirical approach has three stages. First, we apply a causal inference method to estimate the impact of disruptions on station-level travel demand and travel speed (see Section 5.2.1). Then, in Section 5.2.2, we construct vulnerability metrics based on the disruption impact estimated in the first stage. Finally, the third stage imputes[12] missing vulnerability metrics for non-disrupted stations using machine learning algorithms. Figure 5.1 illustrates all steps of the proposed empirical framework.

[12] In Statistics, "imputation" is the process of replacing missing data with substituted values. Here we retrieve these missing values based on a relationship between vulnerability metrics and covariates of the disrupted stations.

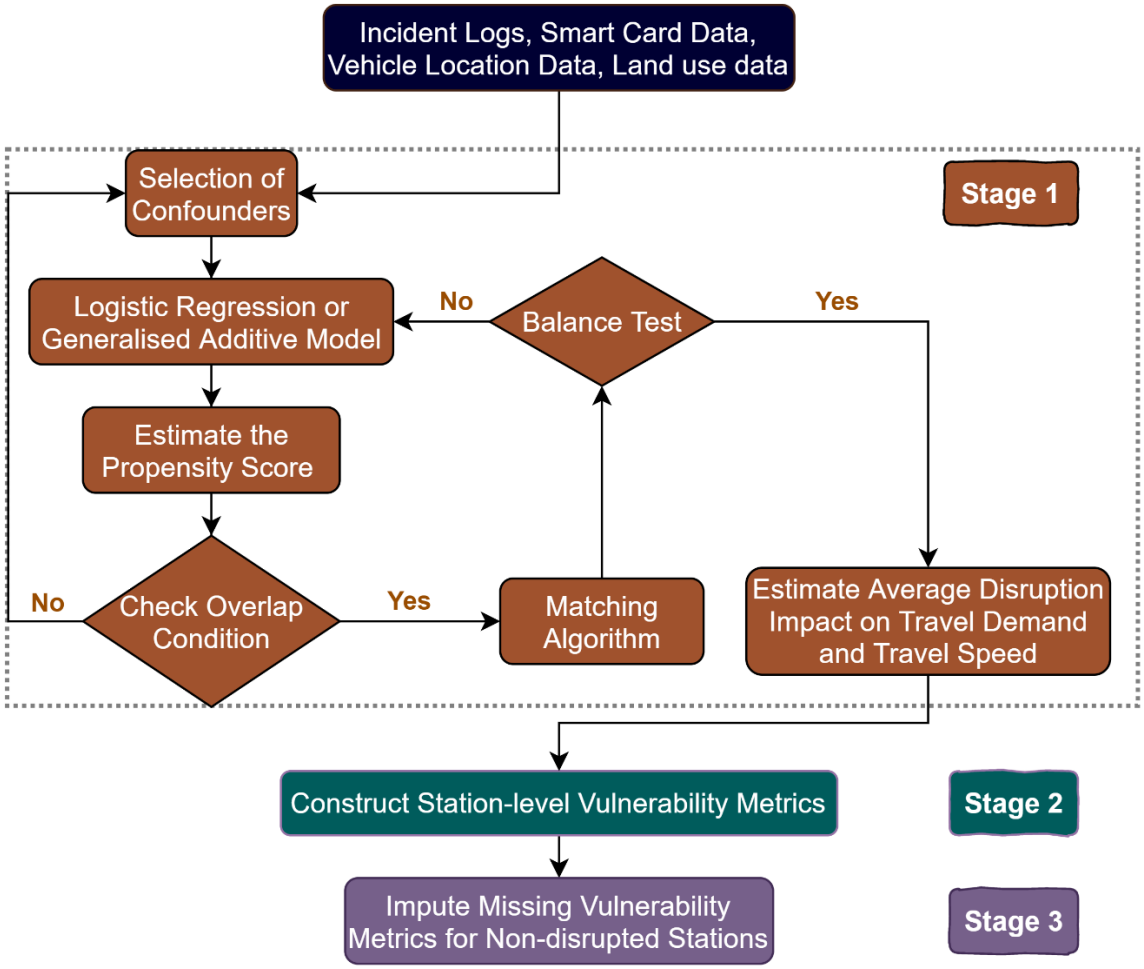


Figure 5.1: Flowchart of the chapter's methodological framework

5.2.1 Causal inference method to estimate disruption impacts

To evaluate the impact of a disruption on a metro system, we use Rubin's potential outcome framework to establish causality (Rubin, 1974). As introduced in Section 3.2.1, we define metro disruptions as 'treatments and the objective of our analysis is to quantify the causal effect of treatments on 'outcomes' related to system performance.[13] Specifically, we are interested in estimating the station-level causal effects of disruptions on (i) travel demand, (ii) travel speed of passengers, and (iii) passenger flow distributions from/to a station. From the literature, we know that factors such as passenger demand, weather conditions, network topology and

[13] In causal inference, 'treatment' means the intervention or exposure assigned to (or encountered by) study units, and 'outcomes' means the observed results or effects of the intervention on a response variable of interest. In the context of this study, service disruptions that occurred at metro stations are the 'treatment', and 'outcomes' are the performance of metro services as measured by indicators such as travel demand, journey speed, and passenger flow distribution.

engineering design influence the likelihood of disruption occurrence (Brazil et al., 2017; Melo et al., 2011; Wan et al., 2015). Therefore, the assignment of the treatment is not random. This is important to our study because the factors associated with the assignment of the treatment are also likely to affect the outcomes of interest, and are thus potential confounders in the estimation of impacts. Since previous studies on disruption impact have ignored the non-randomness of treatments, their estimated impact may be biased.

We adopt propensity score matching (PSM) methods to address this issue, which potentially eliminates such confounding biases. The propensity score is defined as the conditional probability that a unit receives treatment given its baseline confounding characteristics. If the observed characteristics sufficiently capture the sources of confounding, then the propensity score can be used to consistently estimate impacts given conditional independence between treatment assignment and outcomes (e.g., conditional on the propensity score) (Imbens and Rubin, 2015). This index is obtained by estimating a relationship between treatment assignment and baseline confounding characteristics using a regression model. The estimated propensity score is then used to form various semi-parametric estimators of the treatment effect such as weighting, regression, and matching. In this section, we first provide a contextual formulation of PSM and then describe how we apply PSM to quantify the causal impact of disruptions on the performance of metro systems.

Propensity score matching (PSM)

The system-level impact, which averages the impact of all disruptions that occurred in the metro system, is too generic to represent network vulnerability. Thus, we focus instead on estimating station-level disruption impacts. We define study unit i as the observation of a metro station within a 15-minute interval. The treatment variable, denoted by $W_{it} \in \{0, 1\}$, records whether study unit i at time t is observed in a disrupted ($W_{it} = 1$) or undisrupted state ($W_{it} = 0$). To quantify disruption impacts, we define outcomes of interest as the changed travel demand, flow distribution and average speed of trips that start from the given study unit, denoted by Y_{it}.

$$Y_{it}(W_{it}) = Y_{it}(0) \times (1 - W_{it}) + Y_{it}(1) \times W_{it} \tag{5.1}$$

$$Y_{it} = \begin{cases} Y_{it}(0) & if\ W_{it} = 0 \\ Y_{it}(1) & if\ W_{it} = 1 \end{cases}$$

$$i = 1, \dots, n \quad t = 1, \dots, T,$$

where n is the total number of stations within the metro system, and T is the total number of time intervals during the study period (for example, T=4 if the study period is 1 hour). $Y_{it}(0)$ and $Y_{it}(1)$ are counterfactual potential outcomes, only one of which is observed. The propensity score, denoted by $\mathrm{e}(X_{it})$, is obtained by regressing W_{it} on confounding factors, denoted by X_{it}. We discuss potential confounding factors in the empirical case study in Section 5.3.

To derive valid causal inference using PSM, our model needs to satisfy three key assumptions. The first is the conditional independence assumption (CIA),

$$W_{it} \perp (Y_{it}(0), Y_{it}(1)) \mid X_{it}, \tag{5.2}$$

which states that conditional on the observed confounding factors X_{it}, the treatment assignment should be independent of the potential outcomes. The advantages of the propensity score stems from the fact that this conditional independence can be achieved by just conditioning on a scalar rather than high-dimensional baseline covariates (Rosenbaum and Rubin, 1983). Thus, the CIA based on the propensity score can be written as

$$W_{it} \perp (Y_{it}(0), Y_{it}(1)) \mid \mathrm{e}(X_{it}). \tag{5.3}$$

The second assumption requires common support in the covariate distributions by treatment status:

$$0 < pr(W_{it} = 1 | X_{it} = x) < 1 \qquad \text{for all } x, \tag{5.4}$$

which states that the conditional distribution of X_{it} given $W_{it} = 1$ should overlap with that of the conditional distribution of X_{it} given $W_{it} = 0$. This assumption can be tested by comparing the distributions of propensity scores between treatment and control groups.

The third assumption, also known as the stable unit treatment value assumption (SUTVA), requires that the outcome for each unit be independent of the treatment status of other units (Graham et al., 2014).

If all three assumptions hold and the outcome variable is entry demand or travel speed, the average treatment effect of disruptions (ATET) on a station i can be derived using the following equations (Imbens and Wooldridge, 2009)

$$\tau^{i}{}_{ATET} = \hat{\tau}^{i}{}_{match} = \frac{1}{T_d} \sum_{t=1}^{T_d} \left(\hat{Y}^{i}{}_{t}(1) - \hat{Y}^{i}{}_{t}(0) \right), \tag{5.5}$$

$$\hat{Y}^i{}_t(1) = Y_{it},$$

$$\hat{Y}^i{}_t(0) = \frac{1}{M} \sum_{t_c \in J_M(it)} Y_{it_c},$$

$$i = 1, \dots, n \qquad t = 1, \dots, T_d,$$

where $t \in \{1, \dots, T_d\}$ denotes all the disrupted time intervals of station i during the study period and Y_{it_c} is the outcome of the control unit t_c corresponding to station i disrupted or treated at time t. $J_M(it)$ is a set of indices of the closest M control units (in terms of propensity scores) for station i disrupted at time t, during the same 15-minute interval but on a different day.[14] Thus, $\hat{\tau}^i{}_{match}$ represents the average of the difference between the outcomes of treated and matched control units.

When the outcome variable is trip distribution, ATET can be expressed as

$$\tau^i{}_{ATET} = \hat{\tau}^i{}_{match} = \frac{1}{T_d} \sum_{t=1}^{T_d} \left[dif\left(\hat{Y}^i{}_t(1),\ \hat{Y}^i{}_t(0)\right)\right], \tag{5.6}$$

$$\hat{Y}^i{}_t(1) = Y_{it} = \left(r^1_{1it},\ r^2_{1it}, \dots,\ r^k_{1it}\right),$$

$$\hat{Y}^i{}_t(0) = \frac{1}{M} \sum_{t_c \in J_M(it)} Y_{it_c} = \left[\frac{1}{M} \sum_{t_c \in J_M(it)} \left(r^1_{0it_c}\right), \dots, \frac{1}{M} \sum_{t_c \in J_M(it)} \left(r^k_{0it_c}\right) \right],$$

$$i = 1, \dots, n \qquad k = 1, \dots, n \qquad t = 1, \dots, T_d,$$

where for a treated or disrupted unit, Y_{it} denotes the distribution of trips made from (outward) and to (inward) station i at time t, r^k_{1it} denotes the ridership from the disrupted station i to station k in case of outward flow (or from station k to station i in case of inward flow) at time t. Correspondingly, Y_{it_c} denotes a composite distribution which averages the ridership distribution of all closest M control units during the same 15-minute duration, but on a different day. $r^k_{0it_c}$ denotes the ridership between station i and station k for a non-disrupted period t_c in the control group. $dif(a, b)$ is a function to calculate the distance between discrete distributions a and b. In the context of this study, we consider three distance functions:

$$dif_1\left(\hat{Y}^i{}_t(1),\ \hat{Y}^i{}_t(0)\right) = \sqrt{\sum_{k=1}^{n} \left(r^k_{1it} - \frac{1}{M} \sum_{t_c \in J_M(it)} \left(r^k_{0it_c}\right)\right)^2}, \tag{5.7}$$

[14] Please note that the study period of this study is 54 days. Therefore, we observe the same station across multiple days (see Section 5.3 for details).

$$dif_2\left(P(\hat{Y}^i{}_t(1)),\ P\left(\hat{Y}^i{}_t(0)\right)\right) = \frac{1}{\sqrt{2}} \times \sqrt{\sum_{k=1}^{n}\left(\sqrt{P_{it}^k(1)} - \sqrt{P_{it}^k(0)}\right)^2}, \tag{5.8}$$

$$dif_3\left(P(\hat{Y}^i{}_t(1)) \,||\, P\left(\hat{Y}^i{}_t(0)\right)\right) = \sum_{k=1}^{n}\left[P_{it}^k(1) \times \log\left(\frac{P_{it}^k(1)}{P_{it}^k(0)}\right)\right], \tag{5.9}$$

$$P\left(\hat{Y}^i{}_t(1)\right) = \left(p_{it}^1(1), \dots, p_{it}^k(1)\right),$$

$$P\left(\hat{Y}^i{}_t(0)\right) = \left(p_{it}^1(0), \dots, p_{it}^k(0)\right),$$

$$p_{it}^k(1) = \frac{r_{1it}^k}{\sum_{k=1}^{n}\left(r_{1it}^k\right)},$$

$$p_{it}^k(0) = \frac{\frac{1}{M}\sum_{t_c \in J_M(it)}\left(r_{0it_c}^k\right)}{\sum_{k=1}^{n}\left(\frac{1}{M}\sum_{t_c \in J_M(it)}\left(r_{0it_c}^k\right)\right)},$$

where $dif_1(.)$ represents the Euclidean distance, which directly aggregates the difference between each element of the input distributions without normalising. The latter two functions compare the probability mass functions $P(\hat{Y}^i{}_t(1))$ and $P(\hat{Y}^i{}_t(0))$. $dif_2(.)$ represents the Hellinger distance and $dif_3(.)$ represents Kullback–Leibler divergence (also known as relative entropy). Each distance function has its strength and weakness, which we highlight in Section 5.4 while discussing results of the empirical study.

In the next subsection, we explain how the causal inference framework introduced in Equations (5.1), (5.5) and (5.6) can be implemented in the present application. Following the framework summarised in Figure 5.1, we first provide details of the propensity score model, followed by a description of our matching algorithms and the estimation of disruption impacts.

Application of PSM

To predict the propensity score, i.e., the probability of encountering disruptions at a metro station within a 15-minute interval conditional on the baseline confounding characteristics, we use the logistic regression model with a linear link function:

$$\mathrm{e}(X_{it}) = pr\left(W_{it} = 1 \middle| X_{it} = x^{\{c\}}\right) = p(it) \tag{5.10}$$

$$log\left[\frac{p(it)}{1 - p(it)}\right] = \alpha + \beta x^{\{c\}} \qquad i = 1, \dots, n \quad t = 1, \dots, T,$$

where α is the intercept and β is the vector of regression coefficients related to the vector of confounding factors $x^{\{c\}}$. In our empirical study, a station with a higher number of incidents in the past is more likely to encounter a new disruption in the future, just like the black spot on highways. To account for this temporal correlation among disruption occurrence, we ensure that confounding factors contain the history of past disruptions that occurred in the study period.

Additionally, we consider a more advanced generalised additive model (GAM), in which the logarithm of the odds ratio is modelled via semi-parametric smoothing splines. A GAM has the potential to uncover flexible relationships between the likelihood of disruption occurrence and confounding factors. The GAM with temporal correlation is presented in Equation (5.11):

$$\mathrm{e}(X_{it}) = pr\left(W_{it} = 1 \middle| X_{it} = x^{\{c\}}\right) = p(it), \tag{5.11}$$

$$log\left[\frac{p(it)}{1-p(it)}\right] = \alpha + f\left(x^{\{c\}};\beta\right) \quad i = 1,\dots,n \quad t = 1,\dots,T,$$

where $f(x^{\{c\}};\beta)$ is a flexible spline function of baseline characteristics. After estimating propensity scores, we check the common support (overlap) assumption to ensure the effective matching and reliability of the propensity score estimates (Lechner, 2001).

The next step is matching. Every treated unit i at time t is paired with M similar control units based on the value of their propensity scores and time-of-day characteristics. Since there is no theoretical consensus on the superiority of matching algorithms, we adopt two commonly used approaches: subclassification matching and nearest neighbour matching. We then compare them with different replacement conditions and pairing ratios, finally select the one that balances the greatest disparity among the mean of confounding factors. It is also necessary to check the conditional independence assumption after matching. We conduct balancing tests to check whether the disrupted units and the matched units are statistically similar across the domain of confounders. If significant differences are found, we try another specification of the propensity score model and repeat the above-discussed procedure.

In the last step, we estimate station-level disruption impact using Equations (5.5) and (5.6). Given the matched pairs, the treatment effect for a station at a specific period is estimated as the difference between outcomes of the treated unit and its matched control units. Then the average station-level disruption impact is obtained by averaging these differences across all

disrupted periods. We separately estimate the average treatment effects for three measures of metro performance:

i). *Entry ridership:* the number of passengers who enter the study unit.

ii). *Average travel speed:* the average of the speed of all trips that start from the study unit. For each trip, speed is computed as travel distance divided by observed journey time. Whereas journey time is directly obtained using the automated fare collection (AFC) data, travel distance (track length) of the most probable route is derived using the shortest path algorithm.[15] Passengers who left the system and used other transport modes to reach their final destination are not included in the computation of this metrics. If the origin station is entirely closed and no passenger can continue trips by metro, then the average speed will be zero. If the origin station is partially closed, this metrics reflects the average speed of passengers who remain in the system.

iii). *Distribution of passenger flow:* the distribution of completed trips that start from (outward flow) and arrive to (inward flow) the study units.

5.2.2 Constructing vulnerability metrics

We propose four station-level vulnerability metrics that are constructed from the empirical estimates of disruption impacts on the above-discussed performance measures.

i). The *loss of travel demand* is expressed as

$$d_i = -\tau^i{}_{ATE}(entry), \tag{5.12}$$

where $\tau^i{}_{ATE}(entry)$ (calculated using Equation 5.5) denotes the station-level change in the number of entry passengers due to service disruptions. d_i is the loss of demand from external passengers who have not entered the metro system during a 15-minute interval due to disruption.

ii). The *loss of average travel speed* quantifies the decline in the level of service experienced by each passenger at a metro station (individual delay), which is expressed as

[15] For future research, conditional on the availability of vehicle location data, the shortest path algorithm can be replaced by the passenger-train assignment algorithm (Hörcher et al., 2017; Zhu and Goverde, 2019) to infer the most likely path chosen by passengers.

$$s_{avg}{}^{i} = \tau^{i}{}_{ATE}(speed), \tag{5.13}$$

where $\tau^{i}{}_{ATE}(speed)$ (calculated using Equation 5.5) denotes the decrease in the average travel speed of trips starting from station i during a 15-minute disruption period. By definition, $s_{avg}{}^{i}$ accounts for the changes in both travel distance and journey time of passengers.

iii). The *loss of gross travel speed* reflects the loss of passenger kilometres per unit time, which is expressed as:

$$s_{gross}{}^{i} = \tau^{i}{}_{ATE}(speed) \times r_i, \tag{5.14}$$

where r_i denotes the average entry ridership of all disrupted 15-minute intervals at the corresponding station. Thus, $s_{gross}{}^{i}$ denotes the total decrease in average travel speed for all passengers who start their journeys from station i during a 15-minute service disruption.

iv). The *irregularity in passenger flow* reflects the degree of deviation in the distribution of trips from/to the disrupted station as compared to regular conditions, which is expressed as:

$$f_i = \tau^{i}{}_{ATE}(flow) \tag{5.15}$$

where $\tau^{i}{}_{ATE}(flow)$ (calculated using Equation 5.6) denotes the average irregularity in flows that start from or arrive at station i during a 15-minute disruption period. This metrics extends the scope of vulnerability measurement in terms of the entire distribution of entry/exit ridership, instead of just analysing the disruption impact on the entry or exit demand (that is, moments of the trip distribution).

5.2.3 Imputing missing vulnerability metrics

Some stations may not encounter any disruptions within the study period. Thus, the empirical disruption impact and the vulnerability metrics cannot be estimated directly for these stations. To predict the missing metrics of non-disrupted stations, we propose to apply the extreme gradient boosting (XGBoost) algorithm (Chen and Guestrin, 2016).

For a given dataset $E = \left\{(x_i^{\{s\}}, y_i)\right\}$ ($i = 1, \dots, n;\ x_i^{\{s\}} \in \mathbb{R}^m, y_i \in \mathbb{R}$) with n observations, m features, a vector of features $\boldsymbol{x}^{\{s\}}$ and a corresponding variable y. Let $\hat{y}_i$ denote the prediction output given by an ensemble model using B addictive functions

$$\hat{y}_i = \phi\left(x_i^{\{s\}}\right) = \sum_{b=1}^{B} f_b\left(x_i^{\{s\}}\right), \tag{5.16}$$

where f_b is a regression tree, $f_b\left(x_i^{\{s\}}\right)$ denotes the score given by the b^{th} tree to the i^{th} observation. To learn the set of functions, we minimise the following regularised objective:

$$\mathcal{L}(\phi) = \sum_i L(y_i\,,\ \hat{y}_i) + \sum_B \Omega(f_b), \tag{5.17}$$

$$\Omega(f_b) = \gamma T + \frac{1}{2}\lambda\|w\|^2,$$

where $L(\cdot)$ is the loss function that measures how well the model fits on training data. $\Omega(\cdot)$ denotes the regularisation term that measures the complexity of the model and prevents overfitting problems. T denotes the number of leaves in the tree, and w denotes the leaf weights. γ and λ are parameters controlling the penalty for T and w, respectively.

An iterative method is used to minimise the objective function in Equation (5.17). Let $\hat{y}_i^{\{t\}}$ be the prediction of the i^{th} instance at the z^{th} iteration. We greedily add f_z to minimise the following objective:

$$\mathcal{L}^{(z)} = \sum_{i=1}^{n} L\left(y_i\,,\ \hat{y}_i{}^{(z-1)} + f_z(x_i^{\{s\}})\right) + \Omega(f_z). \tag{5.18}$$

This function can be simplified by using the Taylor expansion,

$$\mathcal{L}^{(z)} \simeq \sum_{i=1}^{n}\left[L\left(y_i\,,\ \hat{y}_i{}^{(z-1)}\right) + g_i\, f_z\left(x_i^{\{s\}}\right) + \frac{1}{2}h_i\, f_z{}^2\left(x_i^{\{s\}}\right)\right] + \Omega(f_z) \tag{5.19}$$

where $g_i = \partial_{\hat{y}_i{}^{(z-1)}} L(y_i\,,\ \hat{y}^{(z-1)})$ denotes the first order gradient statistics, and $h_i = \partial^2_{\hat{y}_i{}^{(z-1)}} L(y_i\,,\ \hat{y}^{(z-1)})$ denotes the second order. Letting $I = I_L \cup I_R$, I_L and I_R be the instance sets of the left and right nodes after the tree split from the given node, the loss reduction can be derived as

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma. \tag{5.20}$$

The XGBoost is a scalable system for learning tree ensembles, which applies the regularised objective for better models compared to general gradient boosting algorithms. Interested readers are referred to Chen and Guestrin (2016) for details of this algorithm, who explain the reasons behind its superior prediction accuracy as compared to other competing machine learning methods. For this vulnerability imputation problem, the objective is predicting the missing impacts for non-disrupted stations based on their

characteristics/surrounding facilities and those of disrupted stations. A key property of such problem is the small sample size of predictors, given the size of metro systems which typically range between tens to few hundred of stations (up to 472) and the naturally low disruption rate. The selected XGBoost method is well suited to small samples, flexible, easy to use and can provide more accurate results than most prediction methods. However, the performance of machine learning models always depends on the characteristics of the input data. Also considering that the field of machine learning is evolving rapidly, we encourage readers to explore state-of-the-art alternatives to XGBoost and to test different prediction algorithms to find the most suitable algorithm for their data.

5.3 Data and case study

As mentioned in Section 4.3, the case study in this chapter is also based on the MTR system. Considering the data availability, we focus on four urban lines, including the Island Line, Tsuen Wan Line, Kwun Tong Line and Tseung Kwan O Line, with a total of 49 stations (as shown in Figure 5.2). The following data are collected to estimate the station-level vulnerability metrics. We conducted data processing and analysis using open-source R software (version 4.1.1).

Pseudonymised AFC data

The MTR provided automated fare collection data from 01/01/2019 to 31/03/2019. Excluding holidays and days of incomplete data, we consider 54 weekdays as our study period. The AFC data contain information on transaction date and time, entry and exit locations, encrypted card ID, and ticket type. The resolution of time stamps exacts to one second. By using AFC data, at each station we compute entry/exit ridership, passengers' average journey time, average travel speed and the distribution of inward/outward flows.

Incident logs and disruption detection results

The MTR provided incident information data during this study period. By combing incident logs and the detection results from Chapter 4, we construct an accurate database of service disruptions, which includes their occurrence time, location and duration.

MTR network topology information

We collect data on station coordinates, topology structure and the length of tracks between adjacent stations from DATA.GOV.HK and ArcGIS open databases.[16]

Weather data

We collect temperature (°C), wind speed (km/h) and rain status (cm) from the web portals: Time and Date / Weather Underground of Hong Kong.[17] Based on hourly historical observations, we estimate weather conditions for all selected stations at 15-minute intervals during the study period.

MTR station characteristics

These station-level features include daily ridership, station age, rolling stock age, sub-surface/deep-tube stations and terminal/origin stations. We also obtain supplementary factors, which capture the characteristics of the affected areas around metro stations. To compute these factors, we define the *affected area* as a circular area with a radius of 500 metres around the station (see Figure 5.3). The demographics and transport facility information are collected from the open database Esri China (HK), and land use information is collected from the Planning Scheme Area of Outline Zoning Plans in Hong Kong.[18] To calculate these supplementary factors, for land use we select all Statutory Plan Zones whose boundaries are within the 500-metre radius of the affected area. Then, the related statistics of the selected Statutory Plan Zones are averaged with the weights according to their areas in the circle. For transport facilities, we directly count the number of facilities covered in the affected area. Figures 5.3 and 5.4 illustrate the calculation processes.

To construct the causal inference framework for the MTR, our study units are the observations of metro stations during each consecutive 15-minute interval throughout any service day. We define *metro disruption* as the state when scheduled train services are interrupted for at least 5 minutes at a station. Over the study period, the four urban lines of the MTR encountered 106 disruptions lasting from 5 minutes to over 24 hours. The aim of causal

[16] Source: https://data.gov.hk/en-data/dataset/mtr-data-routes-fares-barrier-free-facilities.
Source: https://www.arcgis.com/home/item.html?id=eae269a6d8d045bea911a13d7f134a74.
[17] Source: https://www.timeanddate.com/weather/hong-kong/hong-kong/historic?month=1&year=2019.
Source: https://www.weather.org.hk/english/wxreport.html.
[18] Source: https://opendata.esrichina.hk/.
Source: https://opendata.esrichina.hk/datasets/esrihk::hong-kong-outline-zoning-plans-land-use-zonings-1/explore?location=22.358476%2C114.143402%2C10.98.

inference is to estimate the unbiased impact of these observed disruptions (i.e., treatment) on system-performance measures (outcome). The treatment status W_{it} is assigned according to the aforementioned disruption database. To match the disruption duration with the timeframe of study units, we define the following rule to assign the treatment status: if a disruption occurs within a 15-minute interval t of a given station i, we regard this study unit as disrupted (i.e., $W_{it} = 1$), regardless of whether disruptions start or end in the middle or last for the entire 15-minute interval. Conversely, if the station is under normal service during the entire 15-minute interval, we regard this study unit as non-disrupted (i.e., $W_{it} = 0$). The treatment outcomes Y_{it} are presented as three station-level performance indicators: entry ridership, average travel speed and flow distribution.

As discussed earlier, metro disruptions may not occur randomly. We list all potential confounding factors for the MTR in Table 5.1, which may be used in estimating the propensity score model (Section 5.2.1). These confounders are selected according to the literature and expertise, including travel demand, weather conditions, engineering design, time of day and past disruptions (Brazil et al., 2017; Melo et al., 2011; Wan et al., 2015). Table 5.1 also shows available covariates for the imputation of missing vulnerability metrics in Stage 3 (Section 5.2.3), which not only include some of confounders, but also include supplementary factors of the MTR station characteristics.

Table 5.1: Available covariates for PSM (stage 1) and vulnerability imputation (stage 3)

Variable	Description	Stage 1	Stage 3
Real-time travel demand			
15-minute entry ridership	The number of passengers that enter a station within 15 minutes before the study unit.	✓	
15-minute exit ridership	The number of passengers that exit a station within 15 minutes before the study unit.	✓	
Average travel demand and speed			
Daily entry ridership	The daily average number of passengers that enter a station during the study period.		✓
Daily exit ridership	The daily average number of passengers that exit a station during the study period.		✓
Daily travel speed	The daily average speed of passengers that start their trips from the study unit.		✓

Weather conditions			
Temperature	Atmospheric temperature around study units. Observations range from 15℃ to 27℃.	✓	
Wind speed	The wind speed around study units, ranging from 4 to 44 km/h.	✓	
Rain status	Rain precipitation around study units, ranging from 0 to 4 mm/h.	✓	
Engineering design characteristics			
Rail connectivity	Dummy variable, representing whether the station is connected to other rail systems.	✓	✓
Overground	Dummy variable, representing whether the station is on surface or closed deep in tube.	✓	✓
Terminal	Dummy variable, representing whether the station is an origin or terminal station.	✓	✓
Number of lines	The number of lines within the given station.	✓	✓
Average adjacent distance	The average distance between the given station and its adjacent stations (km).	✓	✓
Station age	Age of the oldest metro line served by the station.	✓	✓
Rolling stock age	Average age of all rolling stocks operated in the given station.	✓	✓
Time of day	Time of day divided into five intervals; AM peak: 7:30 to 10:30, PM peak: 16:30 to 19:30	✓	
Past disruptions			
Number of past disruptions occurred in the study period	Representation of the temporal correlation of disruption occurrence.	✓	
Station supplementary factors			
Socio-economic characteristics			
Population density*	The density of population within the district to which the given station belongs.		✓
Land use characteristics			
Residential area*	Area of residential buildings (10^3 m^2)		✓
Commercial area*	Area of commercial buildings (10^3 m^2)		✓
Industrial area*	Area of industrial buildings (10^3 m^2)		✓
Open-space area*	Area of open spaces (10^3 m^2)		✓

Road area (m^2) *	Area of major roads (10^3 m^2)	✓
Pedestrian area (m^2) *	Area of pedestrians (10^3 m^2)	✓
Transport accessibility measures		
Bus*	Number of bus stops around the station	✓
Bicycle*	Number of bicycle parking spaces	✓
Car*	Number of metered car parking spaces	✓

**Computed for the affected area around each station*

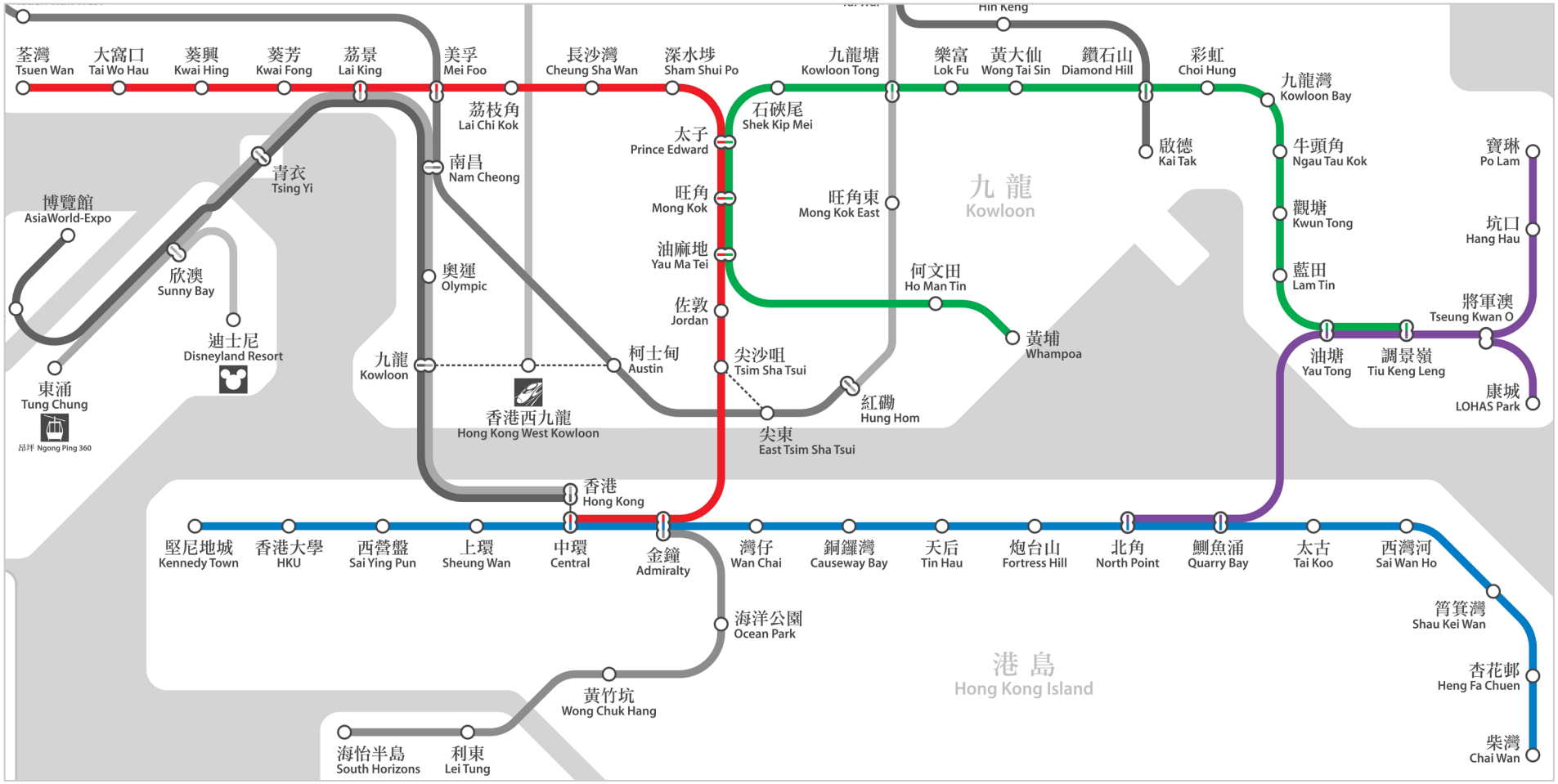
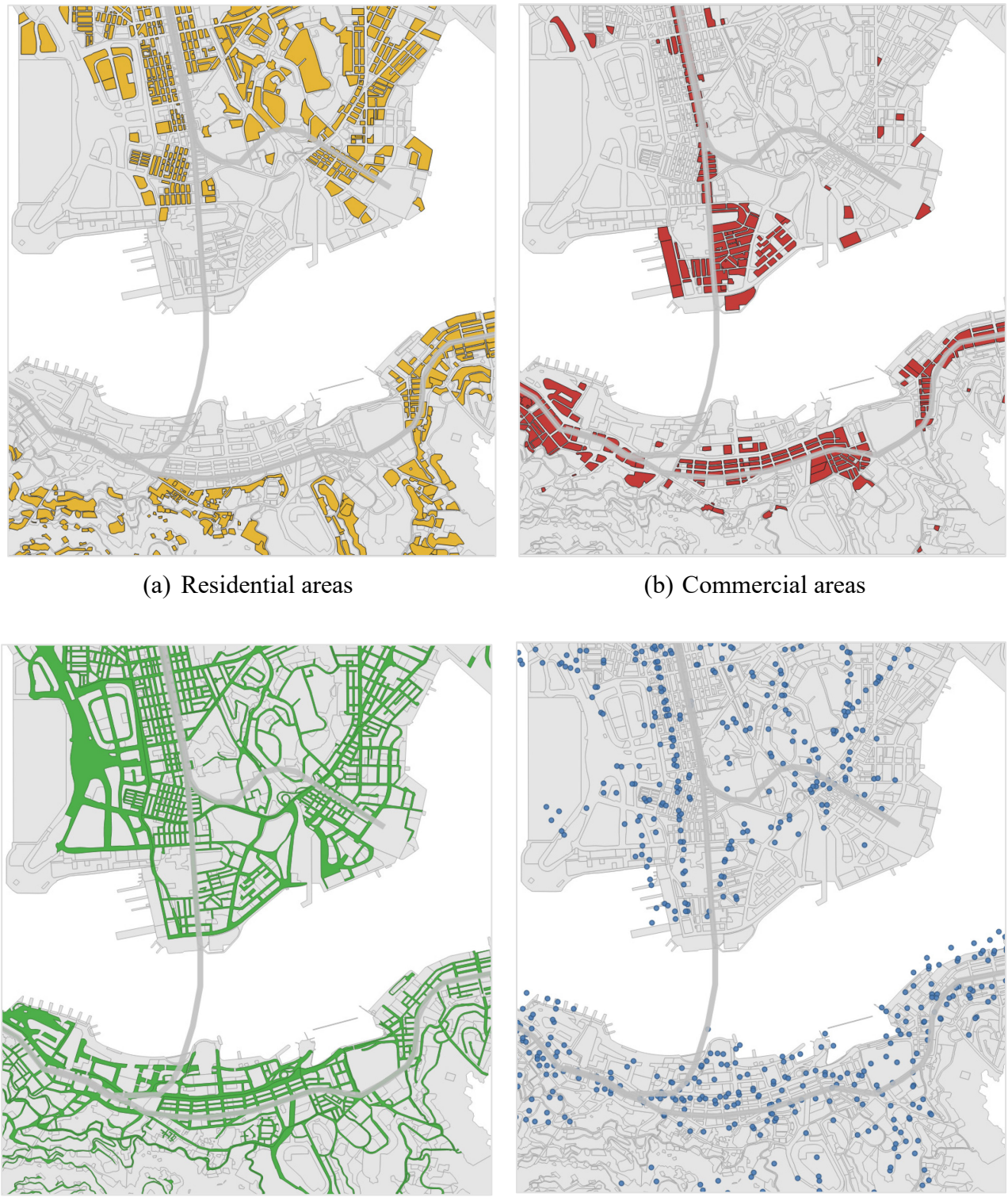


Figure 5.2: The four urban lines that we study in the MTR network (highlighted in colour)

Figure 5.3: Station affected areas that are used to calculate the supplementary factors surrounding metro stations. The radius of each circle is 500 meters



(a) Residential areas

(b) Commercial areas

(c) Major roads

(d) Bus stops

Figure 5.4: Examples of the distribution of land use and transport facilities in Hong Kong

5.4 Results and discussions

From 01/01/2019 to 31/03/2019, excluding holidays and days of incomplete data, our analysis covers 54 weekdays. The MTR system is open for 18 hours per day, from 6:00 a.m. until midnight, which is divided into 72 intervals of 15 minutes. Based on the assumption of exchangeability of weekdays (Silva et al., 2015), we generate a panel dataset for the 49 stations with a total of 49×54×72=190,512 study units. Although the PSM method is a *data-hungry method*, the untreated pool (control group) is large enough to ensure adequate matches for treated units. Specifically, the ratio of the number of control and treatment units is around 190:1.

5.4.1 Propensity score models

We initially include three key baseline covariates – past disruptions, time of day and real-time travel demand – in the logistic regression. We then iteratively add one of the remaining covariates at a time from potential confounders listed in Table 5.1, and conduct the likelihood ratio test to determine whether the additional covariate should be included in the final specification. The Generalised additive models (GAM) have also been tested, but we do not observe any improvement in the model fit. A high proportion of dummy variables (5 out of 10) may limit the gains from a flexible spline specification of the link function. The estimation results of the logistic regression model are summarised in Table 5.2.

Table 5.2: Estimation results of the propensity score model (logistic regression)

Confounders	Coef.	S.E.
Intercept	-3.672***	0.901
Past disruptions	0.560***	0.057
Time 0 (6:00–7:30) (1)	0.010	0.495
Time 1 (7:30–10:30) (1)	0.977***	0.259
Time 2 (10:30–16:30) (1)	-0.223	0.393
Time 3 (16:30–19:30) (1)	1.431***	0.279
Temperature (°C)	-0.282***	0.045
Rain (mm/h)	0.784***	0.209
Overground (1)	1.985***	0.488
Pre 15-minute entry ridership	3.088e-04***	1.099e-04

Overground*wind speed (km/h)	0.114***	0.036
McFadden's pseudo R-squared		0.193

Note: (1) represents dummy variables

The base dummy for the time of day is Time 4 (19:30-24:00).

*p < 0.1; **p < 0.05; ***p < 0.01.

The role of propensity score models is to establish a comprehensive index to represent all confounding factors, rather than predicting treatment assignment.[19] While noting that the logistic regression model does not reveal the causal effect of covariates on the likelihood of disruption occurrence, we succinctly discuss the multivariate correlations uncovered by this model. The coefficients of time dummies indicate that incidents are more likely to occur in peak hours. Positive signs on coefficients of the remaining confounders (except for 'Temperature' and 'Time 2') confirm that all these factors increase the probability of encountering a disruption. Specifically, surface stations are more susceptible to the surrounding environment than those in tubes. We find statistically significant interaction effects between wind speed and overground dummy. The accumulated number of past disruptions increases the probability of encountering another disruption.[20] Conclusively, the propensity score model reveals that the occurrence of metro disruptions is non-random, which, in turn, also justifies the application of causal inference methods in estimating disruption impacts.

Alternatively, the estimated propensity score model can be viewed as a binary classifier that conditionally predicts whether or not metro disruptions occur. To illustrate its diagnostic ability, we compute the area under the receiver operating characteristic curve: AUC= 0.830, which again indicates that the occurrence of metro disruptions is non-random.

[19] The propensity score is the conditional probability of receiving treatment, given a set of confounding covariates.

[20] Please note that disruption frequency is only used in stage 1 to calculate propensity scores which support the matching process. The aim of such design is to remove the bias caused by different frequency of past disruptions in impact quantification. While, the disruption severity in this research, as an important output, is measured as the average impacts of all observed disruptions.

5.4.2 Matching results

Before the estimated propensity scores are utilised for matching, we inspect the *common support* condition (Assumption 2 of the PSM method). Figure 5.5 presents the propensity score distributions for both disrupted and normal observations. The histograms display apparent overlap between the treatment and control groups, even for large propensity scores. There is no treated unit outside the range of common support, which means we do not need to discard any observations. We thus conclude that the overlap assumption is tenable in this empirical study.

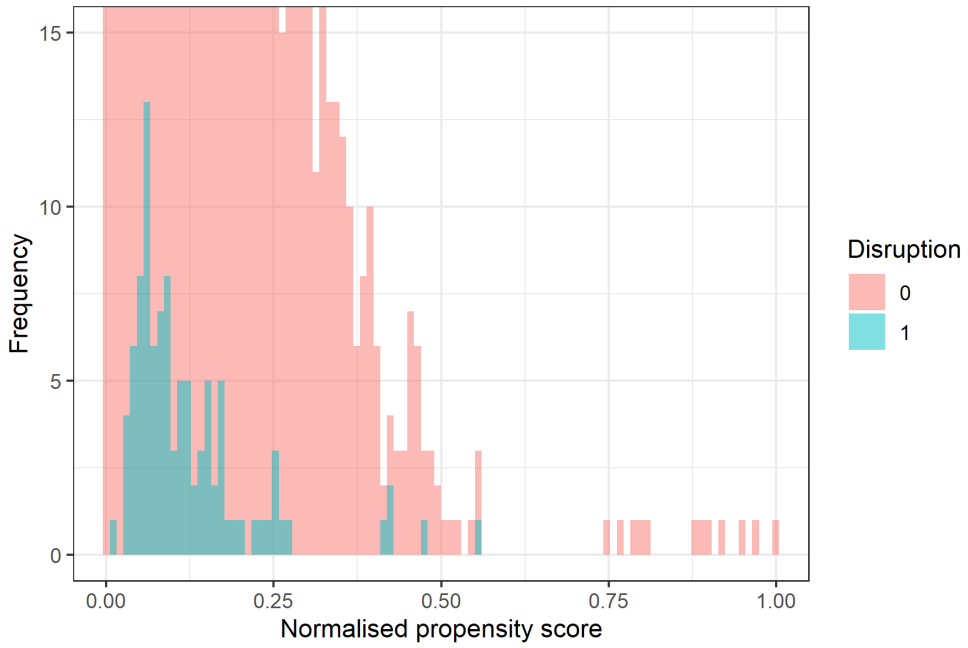


Figure 5.5: Histogram of the normalised propensity scores (common support check).[21] Red and green colour represent the control group and disruptions respectively

The PSM method aims to balance the distribution of confounders between the treatment and control groups after the matching stage. To assess the quality of matching, we perform balance tests for three algorithms: subclassification matching, nearest neighbour matching with replacement ($M = 1$) and nearest neighbour matching with replacement ($M = 2$), where M is the number of matched control units for each treatment unit. It is worth noting that the proposed matching scheme not only conditions on the estimated propensity scores, but also condition on the location and time of day of the treatment (disruption). We

[21] Due to a higher share of the control group, the frequency in Figure 5.5 ranges up to 95,000 for lower propensity scores. However, we truncate frequency at 15 to clearly show the validity of the overlap condition across the entire domain of the propensity score.

find that nearest neighbour matching with replacement ($M = 1$) performs the best, improving the overall balance of all confounding factors by 78.5%. This improvement indicates that within matched pairs, the difference of propensity scores and date-related characteristics between treatment and control units has been reduced by 78.5%, compared with the original data before matching.

5.4.3 Imputation of missing vulnerability metrics

During the study period, 12 out of 49 stations did not encounter any service disruptions. We apply the extreme gradient boosting (XGBoost) algorithm to predict the missing vulnerability metrics of these stations. The input features of the model are indicated in the 'Stage 3' column of Table 5.1, consisting of station-level supplementary factors and a subset of confounding factors. For each vulnerability metrics, the XGBoost algorithm is implemented using the 'xgboost' package of R (Chen et al., 2021). In terms of model settings, we consider the maximum depth of each tree to be 18. We summarise the prediction performance of the XGBoost algorithm in Table 5.3 and benchmark it against three competing methods: linear regression, random forests and support vector machines.

Table 5.3: Comparison of prediction accuracy of different imputation methods

Vulnerability metrics	Performance measures	Imputation methods			
		Linear regression	Random forests	Support vector machines	XGBoost
Demand loss	MAE	19.093	11.114	13.313	1.415
	RMSE	22.549	14.004	24.114	1.942
	RAE	0.781	0.454	0.544	0.058
	RSE	0.471	0.182	0.538	0.003
Avg. travel speed loss	MAE	0.564	0.276	0.326	0.021
	RMSE	0.748	0.407	0.735	0.028
	RAE	0.656	0.321	0.379	0.024
	RSE	0.379	0.112	0.366	0.001
Gross travel speed loss	MAE	1091.812	677.771	727.564	41.272
	RMSE	1485.499	925.377	1760.853	55.420
	RAE	0.715	0.443	0.476	0.027
	RSE	0.367	0.142	0.515	0.001
Irregularity in flow (Euclidean-entry)	MAE	22.582	12.057	12.679	0.902
	RMSE	28.530	15.856	35.221	1.138

	RAE	0.683	0.364	0.383	0.027
	RSE	0.241	0.074	0.367	0.001
Irregularity in flow (Hellinger-entry)	MAE	0.007	0.016	0.009	0.001
	RMSE	0.009	0.020	0.013	0.001
	RAE	0.250	0.597	0.346	0.027
	RSE	0.071	0.368	0.164	0.001
Irregularity in flow (KL-entry)	MAE	0.130	0.195	0.141	0.012
	RMSE	0.177	0.264	0.290	0.016
	RAE	0.481	0.719	0.522	0.046
	RSE	0.224	0.497	0.598	0.002

Four measures are considered to benchmark the performance of XGBoost against other methods – mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE), and relative squared error (RSE). Whereas MAE measures the average magnitude of the errors in predictions, RMSE represents the standard deviation of the unexplained variance (Willmott and Matsuura, 2005). A better prediction model produces lower values for each of these performance measures. The results in Table 5.3 indicate that the XGBoost algorithm outperforms other competing methods with the lowest MAE, RMSE, RAE and RSE for all vulnerability metrics.

5.4.4 The MTR insights

The estimated vulnerability metrics vary across stations in the MTR system (of the four urban lines). We first discuss results for loss of entry demand, loss of average travel speed, and loss of gross travel speed metrics. For 49 stations in the first quarter of 2019, during a 15-minute period of service disruption, the loss of station entry demand ranges from 0 to 119.7 passengers, the loss of average travel speed ranges from 0 to 6.47 kilometres/hour, and the loss of gross travel speed ranges from 0 to 12319.1 passenger-kilometres/hour. The spatial distributions of these vulnerability metrics are visualised in Figures 5.6(a) to 5.6(c). For the demand loss and gross speed loss, a large proportion of vulnerable stations are in Hong Kong central urban areas, while a small number of vulnerable stations are also located in suburban areas. Conversely, for the loss of average travel speed, the most vulnerable stations are scattered around Hong Kong extended urban areas. These stations usually have only one metro line (internal alternatives) and have very limited access to other transport modes (external alternatives) compared to central urban areas. When passengers encounter

disruptions, to continue their trips they need to wait longer in the system until train services are recovered. In other words, due to a lack of alternative routes,[22] passengers at these stations tend to experience more individual delays.

We firstly sort all 49 stations based on demand and speed loss metrics; the top 5 stations are presented in Table 5.4. In terms of demand loss and gross speed loss, stations such as Kowloon Bay and Kwun Tong are among the leading vulnerable stations. However, based on the loss of average travel speed metrics, the most vulnerable stations are Heng Fa Chuen, LOHAS Park and Chai Wan in Hong Kong east coast areas, where passengers suffer the longest delays due to a lack of alternative routes. The above rankings based on different vulnerability metrics can assist metro operators in preparing effective plans for ridership evacuation and service recovery.

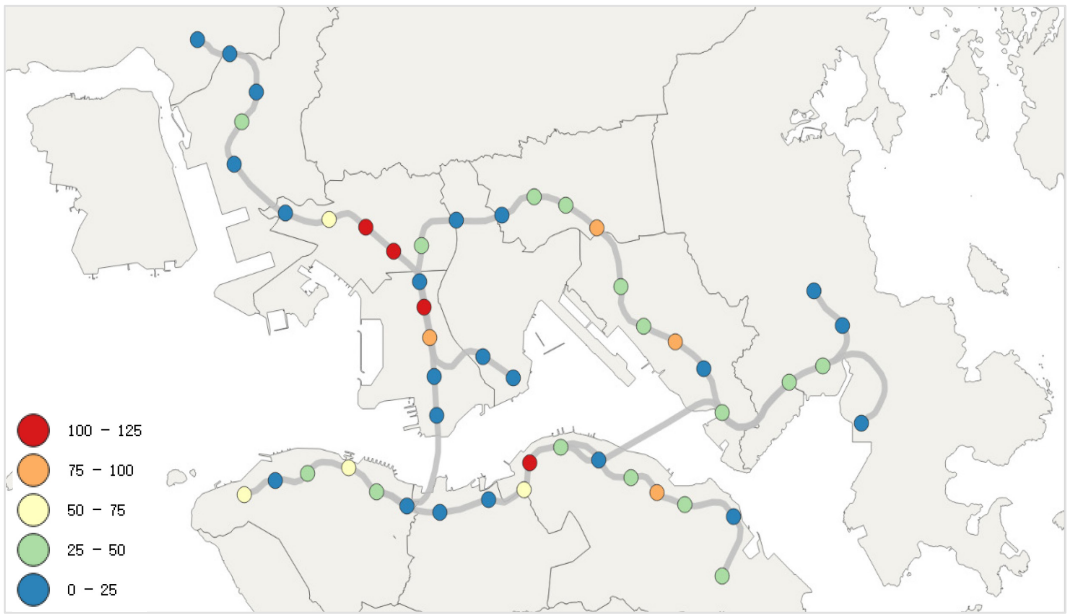
Table 5.4 also presents normalised vulnerability metrics for these top 5 stations, which is the relative percentage change as compared to the undisrupted performance measure (baseline). Note that all baseline situations for these three metrics are calculated by using the average across undisrupted observations. We find that the rankings based on relative vulnerability metrics can be different to those based on absolute metrics, especially for the loss of travel demand and gross speed. In the western part of the Island Line, stations such as Kennedy Town, HKU and Sheung Wan can lose up to 32.0% of their normal demand and 74.85% of the overall travel speeds, due to service interruption. In terms of relative metrics of average travel speed, the same top three vulnerable stations – Heng Fa Chuen, LOHAS Park and Chai Wan – experienced decreases in average travel speed by 32.1%, 18.8% and 15.0%, respectively.

We propose three distance measures for the irregularity in flow metrics: Euclidean distance (ED), Hellinger distance (HD) and Kullback–Leibler (KL) divergence for both outward (from) and inward (to) flows. ED directly compares the difference between each element of the trip distribution, where the *element* represents the ridership between a specific station and the disrupted station. Thus, ED reflects changes in the magnitude as well as in the proportion of the flow of each element because it is not normalised. HD and KL divergence, meanwhile, are normalised measures as they compare the difference between probability mass function of trip distributions, which only capture changes in the

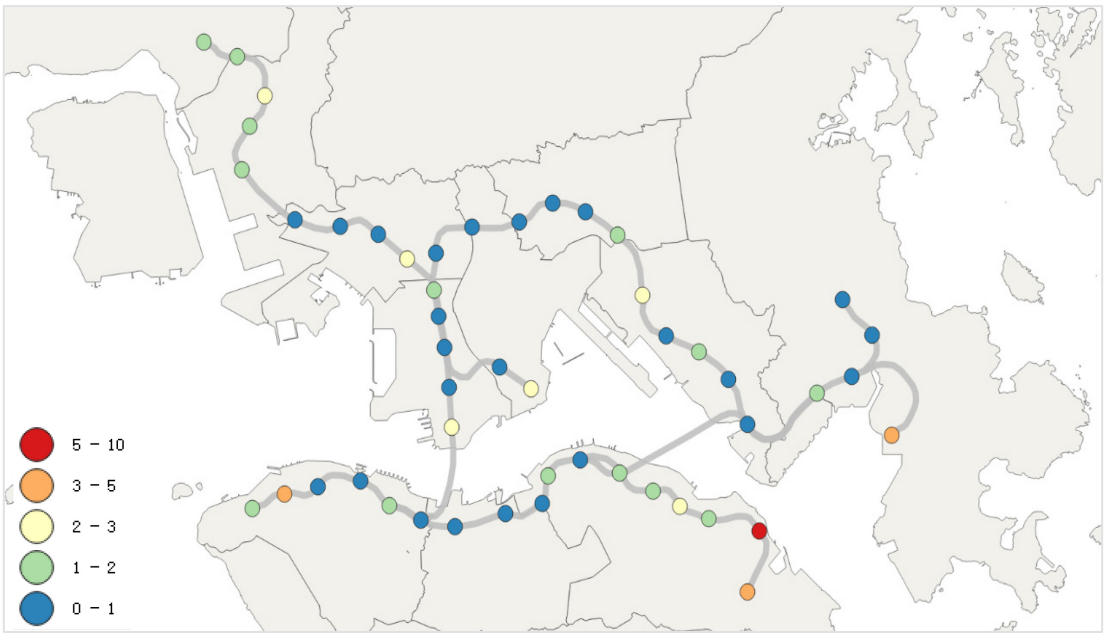
[22] There can be two types of alternative routes under disruptions – within the metro system (interchange to use other operated lines) and outside of it (in the form of other modes).

proportion of trips completed between the disrupted and non-disrupted stations. Unlike ED, HD and KL divergence would not be useful measures if disruption leads to a decrease in ridership across all stations by the same proportion. HD and KL divergence are similar in principle, but the latter can be interpreted as the change in relative entropy, which is meaningful in the context of disruptions in metro systems. As an analogy with the concept of entropy in thermodynamics, we may interpret the extra entropy in metro systems as an additional *generalised cost* (in terms of time and congestion costs) that passengers have to pay under disruptions.

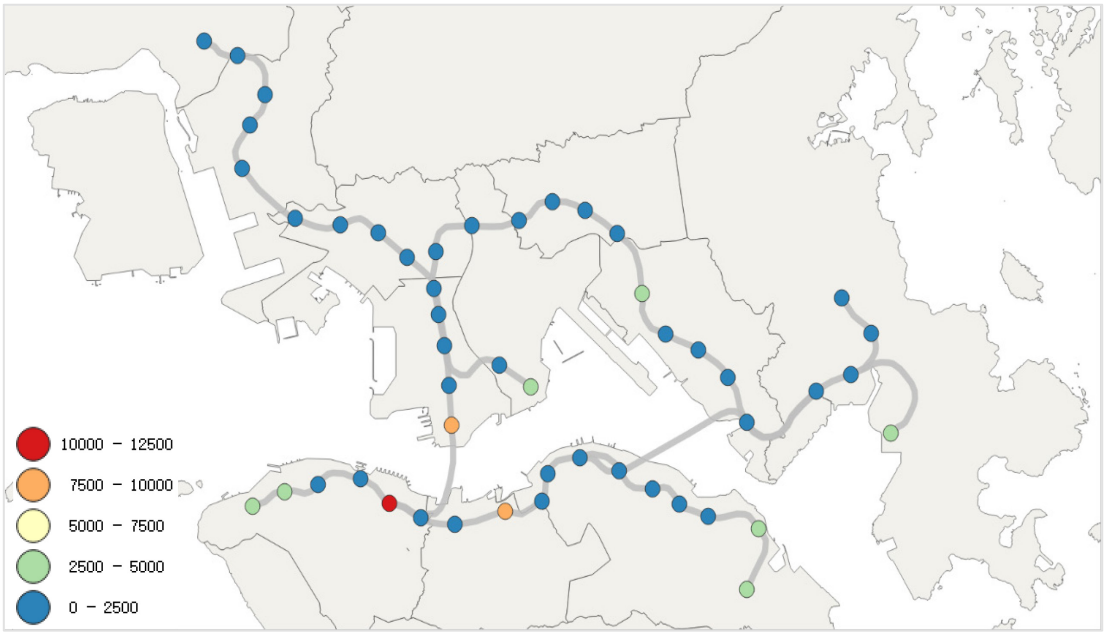
We plot the spatial distribution of all these distance measures in Figures 5.6(d) to 5.6(f). We also sort all 49 stations based on ED, HD and KL divergence; the top 5 vulnerable stations are presented in Table 5.5. We find that the station rankings for outward flow (i.e., the entry ridership distribution) based on ED are similar to those obtained based on gross speed loss metrics. They also share a similar spatial distribution of vulnerable stations. As for the distribution of inward flow (i.e., the exit ridership distribution), the most affected stations are mostly busy stations in central business districts (CBD). However, station rankings based on HD and KL divergence show contrasting results. For both inward (exit) and outward (entry) flow distributions, suburban stations are more severely affected than urban stations on a normalised scale. The top two stations based on HD and KL divergence are LOHAS Park and Po Lam, both being located in the east end of the Tseung Kwan O Line.



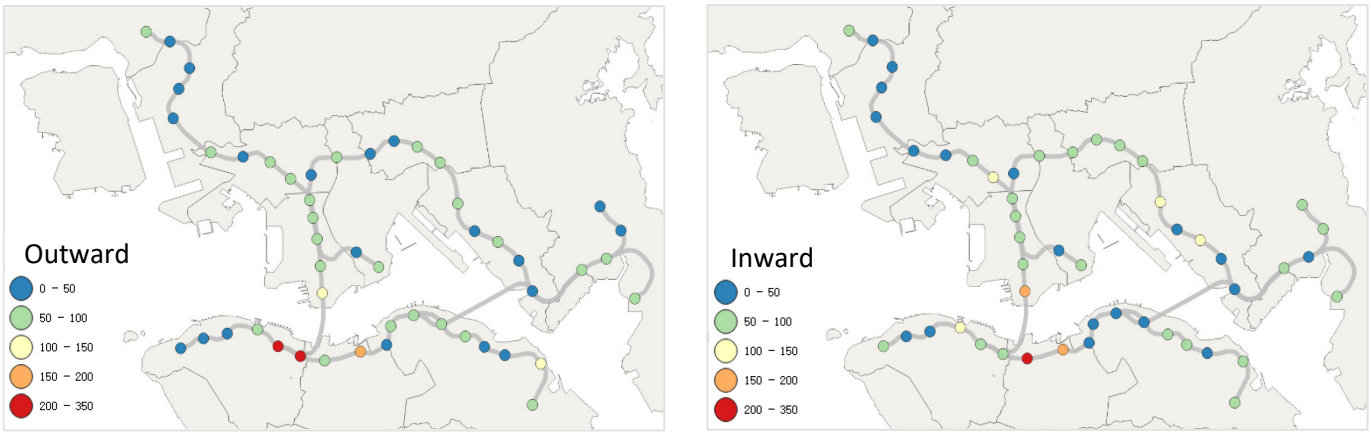
(a) The loss of travel demand (passengers)



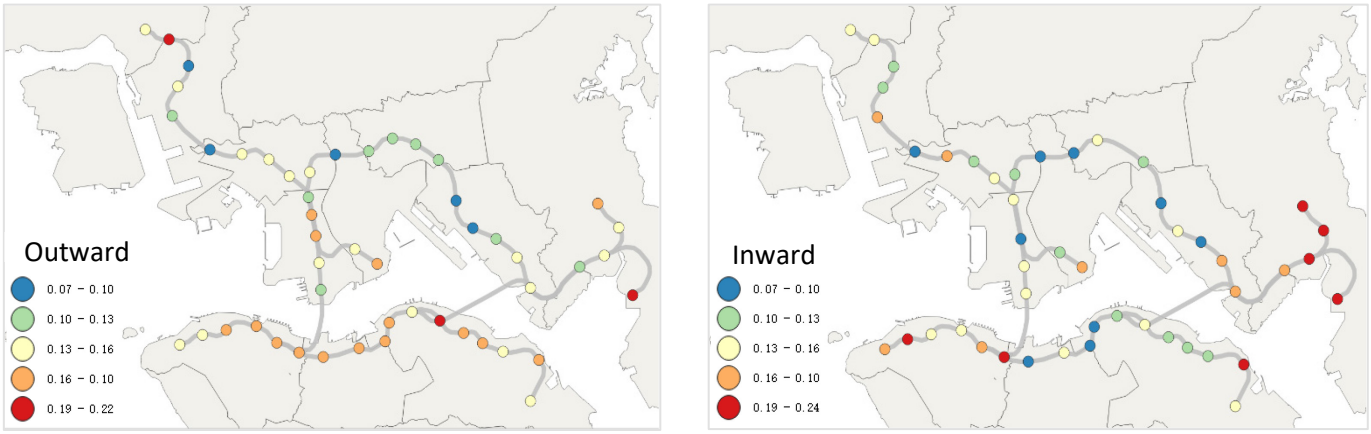
(b) The loss of average travel speed (km/h)



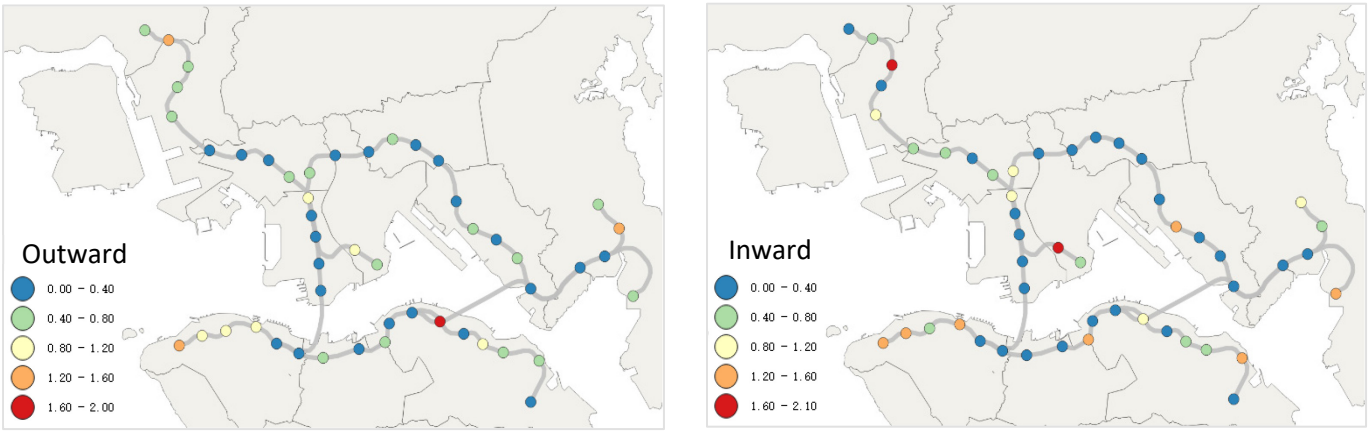
(c) The loss of gross travel speed (passengers*km/h)



(d) The irregularity in flow distribution from/to the disrupted station (ED)



(e) The irregularity in flow distribution from/to the disrupted station (HD)



(f) The irregularity in flow distribution from/to the disrupted station (KL divergence)

Figure 5.6: Spatial distribution of station-level vulnerability metrics in the MTR (four urban lines). Each dot represents a metro station

Table 5.4: Top 5 vulnerable stations based on demand loss and speed loss vulnerability metrics

Station	Demand loss in passenger/15-minute (% of baseline)	Station	Avg. travel speed loss in km/h (% of baseline)	Station	Gross travel speed loss in passenger-km/h (% of baseline)
Cheung Sha Wan	119.67 (14.34%)	Heng Fa Chuen	6.47 (32.05%)	Central	12319.1 (29.53%)
Mong Kok	116.50 (5.98%)	LOHAS Park	4.33 (18.84%)	Tsim Sha Tsui	8973.9 (22.47%)
Sham Shui Po	107.36 (8.25%)	Chai Wan	3.37 (14.97%)	Causeway Bay	7984.8 (23.02%)
Fortress Hill	103.33 (20.74%)	HKU	3.26 (14.93%)	Kowloon Bay	4276.4 (14.00%)
Yau Ma Tei	98.00 (9.00%)	Kwai Hing	2.40 (10.49%)	Chai Wan	3964.4 (25.85%)

Table 5.5: Top 5 vulnerable stations based on irregularity in flow vulnerability metrics

Station	ED (outward)	Station	ED (inward)	Station	HD (outward)	Station	HD (inward)	Station	KL (outward)	Station	KL (inward)
Central	313.06	Wan Chai	213.29	LOHAS Park	0.204	Po Lam	0.238	Quarry Bay	1.82	Ho Man Tin	2.01
Admiralty	265.37	Causeway Bay	175.40	Tai Wo Hau	0.199	LOHAS Park	0.221	Tai Wo Hau	1.41	Kwai Hing	2.00
Causeway Bay	193.58	Tsim Sha Tsui	174.63	Quarry Bay	0.197	Heng Fa Chuen	0.220	Kennedy Town	1.22	Ngau Tau Kok	1.59
Tsim Sha Tsui	148.54	Kwun Tong	117.94	Fortress Hill	0.189	Admiralty	0.215	Hang Hau	1.20	Tin Hau	1.49
Heng Fa Chuen	101.16	Kowloon Bay	116.91	Wan Chai	0.183	HKU	0.206	Sheung Wan	1.18	HKU	1.41

Note: ED: Euclidean distance, HD: Hellinger distance, KL: Kullback–Leibler divergence.

5.5 Conclusions and future work

Disruptions occur frequently in urban metro systems, causing delays, crowding and substantial loss of social welfare. Operators need accurate estimates of vulnerability measures to identify bottlenecks in metro networks. We propose a novel causal inference framework to estimate station-level vulnerability metrics in urban metro systems and empirically validate it for the Hong Kong MTR. In contrast to previous simulation-based studies, which largely assume virtual disruption scenarios and necessitate the adoption of unrealistic assumptions regarding passenger behaviour, our approach relies on real disruption data and avoids making behavioural assumptions by leveraging automated fare collection data. We also illustrate that disruptions can occur non-randomly, which further justifies the importance of the proposed causal inference framework in obtaining the unbiased estimate of disruption impacts.

The proposed empirical framework consists of three stages. First, we conduct propensity score matching methods and estimate unbiased disruption impacts (average of all observed disruptions) at the station level. The estimated impacts are subsequently used to establish vulnerability metrics. In the last stage, for non-disrupted stations, we impute their vulnerability metrics by using the extreme gradient boosting (XGBoost) algorithm. We propose three empirical vulnerability metrics at station level, which are loss of travel demand, loss of average travel speed and loss of gross travel speed. The demand loss metrics reflects the amount of passenger who (i) switched to other transport modes, (ii) switched their departure time, trip origin or destination, (iii) or ended their trip, before entering the disrupted metro system. In other words, it implies the demand for alternative transport services during disruptions, which can guide metro operators to prepare effective service replacement plans. The two speed-related metrics reflect the degradation in the level of service for passengers who still use the metro system under disruptions. These metrics provide essential information for service recovery to mitigate the adverse influence on passengers and the overall performance of stations. The proposed irregularity in flow metrics extends the scope of vulnerability measurement to the changes in trip distribution. This irregularity metrics can be used to reflect the level of disorder within metro systems.

Results of the case study of the MTR in 2019 indicate that the effect of service disruption is heterogeneous across metro stations, and it depends on the location of a station in the network and other station-level characteristics. In terms of the travel demand loss and gross speed loss (overall delay), the most affected stations are more likely to be found in Hong Kong central

urban areas. On the other hand, considering average speed loss (individual delay), the most affected stations are scattered around the suburban or extended urban areas (e.g., LOHAS Park and Kwai Hing) due to a lack of alternative routes.

Disruption impact estimates are probabilistic relative to the sample data, that is, causal estimates and vulnerability metrics estimates have sampling distribution. Since our analysis is based on the data of the MTR from January 1 to March 31, 2019, the results of our case study reflect the vulnerability status of MTR for this specific period. If we use data from other periods, the estimates of vulnerability metrics might change due to inherent temporal variations in travel demand and incidents. Therefore, to improve the generalisability of vulnerability metrics estimates, the study period needs to be sufficiently long for the sample to be representative of the population. That is, a sample should capture supply-side interruptions as much as possible, including service disruptions due to maintenance. In addition, the sample should also reflect the possible fluctuations of travel demand.

The proposed methodology to obtain the unbiased estimates of disruption impact thus provides crucial information to metro operators for disruption management. More specifically, it helps with identifying the bottlenecks in the network and with preparing targeted plans to evacuate ridership as well as to recover services in case of disruptions. The direct integration of the estimated vulnerability metrics in preparing these target plans remains an avenue for future research. It is worth noting that the proposed framework can be applied to other metro systems conditional on the availability of the required data on incident logs, confounding characteristics and performance outcomes. Future empirical studies can also incorporate other context-specific and relevant confounders or outcome indicators in their analysis. For example, they can explore the disruption impacts on interchange passengers if the required datasets are available. We do not include this part of ridership in our MTR case study because it cannot be directly derived from the AFC data. A more advanced assignment algorithm is required to identify passengers' routes by matching AFC data with vehicle location data and reproduce the spatiotemporal flow distribution in the metro network.

In line with the limitations of this study, there are three potential directions for future research. First, stations surrounding the disrupted stations may also be affected due to indirect propagation, but this study does not account for such spillover effects. Modelling spatiotemporal propagation disruption impacts requires significant methodological developments, which would be an important improvement over the current method. For

instance, recent developments in Bayesian nonparametric sparse vector autoregressive models (Billio et al., 2019) can be adapted to model the spatiotemporal effect of service disruptions in transit networks. Second, the proposed vulnerability metrics can reveal static disruption impacts at different stations, but passengers need real-time service information to reschedule their trips. Thus, the current framework can be extended to update the vulnerability metrics dynamically. Considering the interaction between information provision and how it influences passengers' decisions under disruptions, this advancement would improve the dissemination of the incident alerts to passengers in real time. Finally, by merging data from other travel modes (e.g., bus, urban rails, shared bike or taxi) with metro datasets, we can estimate multi-modal vulnerability metrics in the same causal inference framework and understand the characteristics of the mode shift due to disruptions. In a potential extension of our method to multi-modal transport systems, the lost demand would not include passengers who shift to other public transport modes due to metro disruptions. Compared to metro-only vulnerability metrics, multi-modal demand loss metrics would focus on passengers who cancel their trips entirely or switch to private transport modes. Therefore, for metro stations linked to multi-modal hubs, the multi-model demand loss metrics might be lower than the metro-only metrics. The magnitude of this gap would depend on the attractiveness of alternative public transport services compared to private modes.

Chapter 6
Quantifying the direct and spillover effects of disruptions in urban metro networks

As mentioned in the previous chapter, urban metro systems are often affected by disruptions, causing delays, crowding and decline in passenger satisfaction. To mitigate such adverse impacts, it is important for metro operators to comprehensively understand disruption impacts. Therefore, this chapter proposes to utilise modified synthetic control methods to quantify the direct and spillover causal effects of disruptions. We relax the non-interference assumption and allow disruption impacts to propagate in metro networks. The proposed method is unique in the sense that the weighted average of unaffected observations is used to simulate the counterfactual outcomes of disruptions. Such a framework enables the estimation of the indirect impacts on other non-disrupted stations. This research also delivers an innovative analysis of the propagation progress of disruption impacts along metro lines.

6.1 Introduction

In urban metro systems, to manage disruptions and mitigate their adverse influence, it is important for operators to measure and understand disruption impacts. Effective recovery strategies need detailed information on the affected ridership, delayed time and crowding level in stations or trains. For passengers, updates of real-time disruption impacts can also help them reschedule travel plans. Therefore, in this chapter, we focus on accurately quantifying disruption impacts and analysing their spatiotemporal propagation across the metro network.

In the literature, the empirical-based studies have analysed the disruption impacts via survey data and smart card data. However, most of the studies tacitly assume that metro disruptions occur randomly. Ignoring the existence of factors that can influence both disruption occurrence and the corresponding outcomes, their estimates of disruption impacts are not causal and can be biased. Zhang et al. (2021a) relaxed this assumption and avoided such confounding bias by applying the propensity score matching methods. Although this recent

research is conducted under the casual inference framework, there are still some limitations. First, they focused on the direct disruption impacts, which are restricted to the stations where disruptions occurred, assuming that other parts of the network would not be affected. In reality, however, with the presence of interference among connected stations, disruption impacts can spread along metro lines and influence the entire network. To the best of our knowledge, the propagation of indirect impacts (spillover effects) in metro systems has not yet been explored empirically[23]. Second, their station-level outputs are the average impacts of all disruptions that occurred in the study period, from which the influence of individual disruptions is hard to be distinguished.

We propose a new approach to quantify the direct and indirect causal effects of metro disruptions on system performances, where the measures of interest are passenger demand, average travel speed and journey time. The modified synthetic control method is unique in the way that it allows interference among metro stations, and the weighted average of unaffected observations (*synthetic control*) is used to simulate the counterfactual outcomes of disruptions. Under the setting that disruptions are regarded as a treatment or intervention, we determine the optimal weights by best approximating the outcomes and predictors (attributes) of the treated unit using pre-treatment data. The proposed framework can not only be applied to the disrupted station, but also to other affected stations in the network, thus enabling the analysis of spillover effects of disruptions and their spatiotemporal propagation.

A case study of four urban lines in the Hong Kong Mass Transit Railway (MTR) has been conducted, focusing on the selected disruption on the Island Line. This application indicates that the synthetic controls outperform the before-after comparison and the simple average of control units. In terms of the direct disruption impacts, the exit ridership of the interrupted station drops by 50%, with an increase of over 11 minutes in the average journey time per trip and a maximum of 9 km/h decrease in the average travel speed. For other stations in the MTR network, the impacts of this disruption spread throughout the entire Island Line and reach further stations on another two connected lines. With the increase of propagation distance, the magnitude of disruption impacts gradually decreases. Two hours after the disruption, the service of the entire network returns to normal.

[23] The spillover effects of public transport disruptions have been analysed based on various simulation-based methods. Section 2.5.1 reviews the related literature in detail.

The rest of the chapter is organised as follows. Section 6.2 presents the synthetic control framework and the method of weight choice. In Section 6.3, we detail the case study on the Hong Kong MTR with an example of disruption impacts propagation. Results are then discussed in Section 6.4. Finally, Section 6.5 concludes and discusses limitations and future work for this research.

6.2 Methodology

To measure the impact of disruptions on a metro system, we use Rubin's potential outcome framework to establish causality (Rubin, 1974). As introduced in Section 3.2.2, We define metro disruptions as 'treatments (or interventions)' and the objective of our analysis is to quantify the direct and indirect causal effect of treatments on 'outcomes' related to service performance. Specifically, we are interested in estimating station-level impacts on (i) travel demand, (ii) journey time and (iii) the travel speed of passengers. From the literature, most of the empirical research concentrates on direct disruption impacts. The performance of other stations is tacitly assumed to be independent of the disrupted stations. However, metro stations are connected by tracks and consecutive train services, meaning that disruption impacts can propagate along lines to the entire network. Moreover, due to the non-random occurrence of metro disruptions, confoundedness needs to be considered in real estimations (Zhang et al. 2021a). To address these issues, we adopt a modified synthetic control method, which relaxes the non-interference assumptions and also eliminates potential confounding biases.

6.2.1 The modified synthetic control method

In this research, we define the study unit as the status of a metro station $a = 1, \dots, A$ on a given day $d = 1, \dots, D$ during interval $t = 1, \dots, T$. We consider 15-minute-long intervals. The station is considered to be *treated* in case it encounters service disruptions of above five minutes in the 15-minute interval. The treatment assignment, denoted by $W_{adt} \in \{0,1\}$, records whether the station a has been exposed to disruptions during interval t on day d. Under the Consistency Assumption [24] (Imbens and Rubin 2015), we use $Y_{adt}(W_{adt})$ to denote the

[24] The assumption that there are no hidden versions of treatment.

potential outcomes, which are defined as the passenger demand, average journey time and average travel speed.

$$Y_{adt}(W_{adt}) = Y_{adt}(0) \times (1 - W_{adt}) + Y_{adt}(1) \times W_{adt}, \tag{6.1}$$

$$Y_{adt} = \begin{cases} Y_{adt}(0) & if\ W_{adt} = 0 \\ Y_{adt}(1) & if\ W_{adt} = 1, \end{cases}$$

where $Y_{adt}(0)$ and $Y_{adt}(1)$ are counterfactual potential outcomes, only one of which is observed. Causal inference studies commonly make the stable unit treatment value assumption (SUTVA), which requires that the outcome for each unit should be independent of the treatment status of other units (Graham et al., 2014). However, due to interference between stations in the metro network, SUTVA is unlikely to hold. We illustrate how this modified synthetic control method could estimate causal effect in the absence of SUTVA.

To create the synthetic control, we utilise the observations on days when no disruption in the entire metro network as the *donor* pool. $\boldsymbol{d_N}$ is a set of such undisrupted days with cardinality J. This design of the donor pool benefits from the fact that metro automated data contain observations from multiple days. To quantify the impact of a disruption that starts at station $a_I = 1, \dots, A$, on day $d_I = 1, \dots, D$, at time $T_{IS} = 1, \dots, T$ and ends at time $T_{IE} = T_{IS}, \dots, T$, we construct a vector of outcomes $\boldsymbol{p}$. We assume that this disruption has no effect on outcomes before the treatment period T_{IS}. Conversely, after T_{IS}, all stations in the network can be affected by this disruption.

$$\boldsymbol{p} = \{\boldsymbol{p_1}, \boldsymbol{p_2}, \dots, \boldsymbol{p_A}\} \tag{6.2}$$

where $\boldsymbol{p_a}$ is the vector of outcomes for station a during time intervals $t = T_{IS}, \dots, T$ on the disrupted day d_I and J undisrupted days (i.e., $J + 1$ days). Since we stack the data of the treated day followed by undisrupted days, $p_{ajt} = Y_{ad_It}(W_{ad_It})$ for $j = 1$ and $p_{ajt} = Y_{ad_jt}(W_{ad_jt})$ for $j = 2, \dots, J + 1$, $d_j \in \boldsymbol{d_N}$. Note that $W_{ad_It} = 1$ if $t \geq T_{IS}$ and $W_{ad_It} = 0$ otherwise.

The main idea behind the proposed method is to use the untreated and unaffected units (from the donor pool) to construct a 'synthetic' control unit, of which the characteristics approximate that of the treated unit. Therefore, the counterfactual outcome of the treated unit can be estimated by the outcomes of the synthetic control unit. For a station-interval pair of the treated/affected station a, we create a synthetic control by weighted combination of the same station-interval pair on J undisrupted days. The counterfactual outcome is defined as a

weighted average of the outcomes of units in the donor pool, where $\boldsymbol{C^a} = (c^a{}_2, \dots, c^a{}_{J+1})'$ is a $J \times 1$ vector of non-negative weights that sum to one (see the next subsection for the weight computation details). The modified synthetic control estimators of the counterfactual outcomes ($\hat{Y}^N_{ad_It}$) and causal effect estimate ($\hat{\tau}_{ad_It}$) are, respectively,

$$\hat{Y}^N_{ad_It} = \sum_{j=2}^{J+1} c^a{}_j \cdot Y_{ad_jt}(0) \qquad t = T_{IS}, \dots, T, \tag{6.3}$$

$$\hat{\tau}_{ad_It} = Y_{ad_It} - \hat{Y}^N_{ad_It} \qquad t = T_{IS}, \dots, T. \tag{6.4}$$

With the above settings, during the given disruption, the direct causal effects on the treated station a_I is derived as

$$\tau_{a_Id_It} = Y_{a_Id_It}(1) - \sum_{j=2}^{J+1} c^{a_I}{}_j \cdot Y_{a_Id_jt}(0) \qquad t = T_{IS}, \dots, T_{IE}. \tag{6.5}$$

After the disruption, the remaining impacts (indirect) on station a_I is derived as

$$\tau_{a_Id_It} = Y_{a_Id_It}(0) - \sum_{j=2}^{J+1} c^{a_I}{}_j \cdot Y_{a_Id_jt}(0) \qquad t = T_{IE} + 1, \dots, T, \tag{6.6}$$

where $Y_{a_Id_It}(1\ or\ 0)$ denotes the observed outcome of the treated unit on the disrupted day in interval t. $c^{a_I}{}_j$ denotes the weight of the j^{th} day in the corresponding donor pool for station a_I, and $Y_{a_Id_jt}(0)$ denotes the observed outcomes for the same station-interval pair on the j^{th} day.

Similarly, the indirect causal effects (spillover effects) of the disruption at station a_I on the performance of other station a_O ($a_O \in 1, \dots, A \setminus a_I$) is derived as

$$\tau_{a_Od_It} = Y_{a_Od_It}(0) - \sum_{j=2}^{J+1} c^{a_O}{}_j \cdot Y_{a_Od_jt}(0) \qquad t = T_{IS}, \dots, T, \tag{6.7}$$

where $Y_{a_Od_It}(0)$ denotes the observed outcomes for the affected units of other stations (non-treated), during and after the given disruption. $c^{a_O}{}_j$ and $Y_{a_Od_jt}(0)$ denote the weight and outcomes of the j^{th} day in the corresponding donor pool for station a_O. Figure 6.1 illustrates the design of the modified synthetic control framework for metro disruptions.

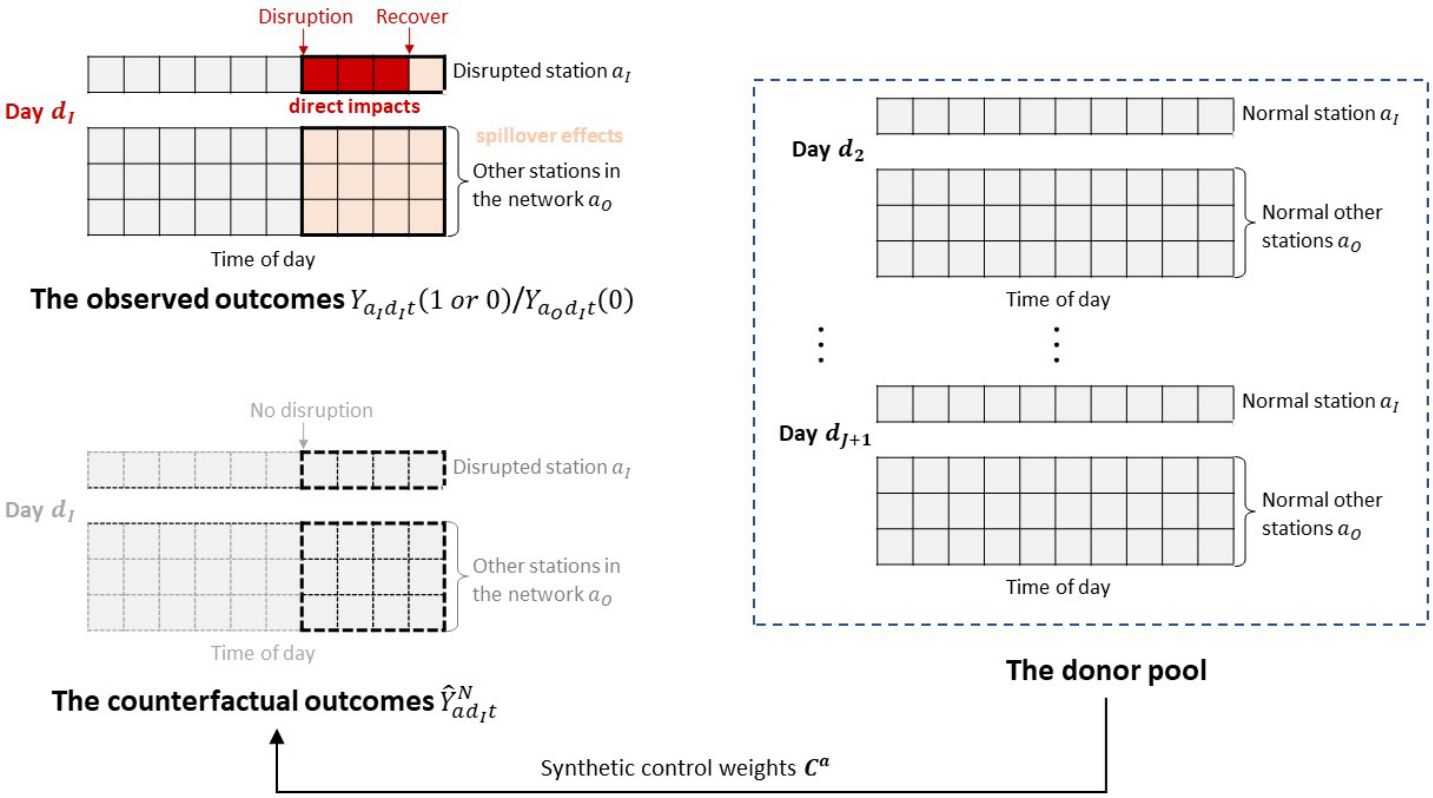


Figure 6.1: Schematic overview of the modified synthetic control method for metro disruptions. The donor pool consists of observations from non-disrupted days. a_o represent any other station in the network, it can be upstream, downstream or a surrounding station to the disruption.

6.2.2 The choice of weights

A naïve example is that of equally assigning weights $c^a{}_j = 1/J$ to each of the units in a comparison control group (donor pool). The estimator for $\tau_{ad_I t}$ is

$$\hat{\tau}_{ad_I t} = Y_{ad_I t} - \frac{1}{J}\sum_{j=2}^{J+1} Y_{ad_j t} \qquad t = T_{IS}, \dots, T, \tag{6.8}$$

where the synthetic control is the simple average of all the units in the donor pool.

In this research, we apply the method proposed by Abadie and Gardeazabal (2003) and Abadie et al. (2010) to determine the choice of weights $\boldsymbol{C^a}$. For each unit j in the donor pool corresponding to station a at time t, we first observe a set of k predictors of the outcomes, denoted by $X^{at}{}_{j1}, \dots, X^{at}{}_{jk}$. The $k \times 1$ vectors $\boldsymbol{X^{at}}_{\boldsymbol{1}}, \dots, \boldsymbol{X^{at}}_{\boldsymbol{J+1}}$ collect the values of such predictors for units $j = 1, \dots, J+1$, individually. The $k \times J$ matrix $\boldsymbol{X^{at}}_{\boldsymbol{0}} = \left[\boldsymbol{X^{at}}_{\boldsymbol{2}}, \dots, \boldsymbol{X^{at}}_{\boldsymbol{J+1}}\right]$ represents the predictors for the J unaffected units within this donor pool (Abadie, 2021).

Predictors $\boldsymbol{X}$ are defined not to be influenced by the treatments (interruptions) and with a certain amount of influence on the outcomes. Thus, to account for the non-randomness of disruption occurrence, we include partial confounding factors of metro disruptions when selecting predictors, such as weather conditions and pre-intervention values of outcome variables.

Then, weights $\boldsymbol{C^a}$ are determined according to whether the resulting synthetic control can best resemble the values for the predictors of the treated unit ($j = 1$) before disruption. That is, given a set of non-negative constants $\boldsymbol{V^a} = ({v^a}_1, \dots, {v^a}_k)$, the optimal synthetic control $\boldsymbol{C^{a*}} = ({c^a}_2^*, \dots, {c^a}_{J+1}^*)'$ is obtained from the following minimisation problem:

$$\min_{\boldsymbol{C^a}} \|\boldsymbol{X^{at}}_{\boldsymbol{1}} - \boldsymbol{X^{at}}_{\boldsymbol{0}} \cdot \boldsymbol{C^a}\| = \sqrt{\sum_{h=1}^{k} {v^{at}}_h \left({X^{at}}_{1h} - {c^a}_2 \cdot {X^{at}}_{2h} - \cdots - {c^a}_{J+1} \cdot {X^{at}}_{(J+1)h}\right)^2},$$

$$\text{such that } \sum_{j=2}^{J+1} {c^a}_j = 1, \quad {c^a}_j > 0 \tag{6.9}$$

where the positive constants ${v^a}_1, \dots, {v^a}_k$ reflect the relative importance of the k predictors. To determine the value of weights $\boldsymbol{V^a}$ and $\boldsymbol{C^a}$, we follow the steps below (Abadie, 2021).

i). Divide the pre-intervention periods (before disruption occurs) into an initial training period ($t = 1, \dots, t_0$) and a subsequent validation period ($t = t_0 + 1, \dots, T_{IS} - 1$). The lengths of the training and validation periods can be application specific.

ii). With training period data on the predictors, compute the synthetic control weights $\widetilde{\boldsymbol{C^a}}(\boldsymbol{V^a})$ according to the optimisation as Equation (6.9).

iii). In the validation period, the mean squared prediction error (MSPE) of this synthetic control with respect to $Y_{ad_1 t}^N$ is:

$$\min_{\boldsymbol{V^a}} \sum_{t=t_0+1}^{T_{IS}-1} \left(Y_{ad_1 t} - \widetilde{c^a}_2(\boldsymbol{V^a}) \cdot Y_{ad_2 t} - \cdots - \widetilde{c^a}_{J+1}(\boldsymbol{V^a}) \cdot Y_{ad_{J+1} t}\right)^2. \tag{6.10}$$

Minimise the MSPE in Equation (6.10) with respect to $\boldsymbol{V^a}$.

iv). With the validation period data on the predictors, use the resulting $\boldsymbol{V^{a*}}$ to calculate weights $\boldsymbol{C^{a*}} = \boldsymbol{C^a}(\boldsymbol{V^{a*}})$, according to Equation (6.9).

6.3 Data and case study

In this chapter, the case study is also based on the Hong Kong MTR, focusing on the four urban lines (the Island Line, Tsuen Wan Line, Kwun Tong Line and Tseung Kwan O Line) with 49 stations (refer to Section 5.3). Figure 5.3 displays the partial network that we study. The following data are used to estimate the direct and spillover causal effects of disruptions. We conducted data processing and analysis using open-source R software (version 4.1.1).

Pseudonymised smart card data

The Hong Kong MTR provided smart card data from 01/01/2019 to 31/03/2019. The smart card data contain information on the time and location of tap-in and tap-out transactions throughout the system, recording individual trips. Based on the data, we compute entry and exit ridership, passenger's average journey time and average travel speed for each target station. The resolution of time stamps exacts to one second.

Incidents logs and disruption detection results

The MTR provided incident information data during this study period. By combing incident logs and the detection results from Chapter 4, we construct an accurate database of service disruptions, which includes their occurrence time, locations and durations.

Weather data

We collect temperature (°C), wind speed (km/h) and rain status (cm) from the web portals: Time and Date / Weather Underground of Hong Kong. Based on the hourly historical observations, we estimate weather conditions for all selected stations at 15-minute intervals (refer to Section 5.3).

Mega events in Hong Kong

From 01/2019 to 03/2019, we collect the information – including the location and event time - of three types of mega-events held in Hong Kong: concerts, sports matches and exhibitions. Data sources include official news and government records.[25]

In this case study, historical disruption data are obtained from the aforementioned disruption database. Minor disruptions that lasted less than five minutes are excluded from the impact estimation. During the study period, 106 disruptions (of over 5 minutes) were observed

[25] https://www.mevents.org.hk/en/index.php.
https://www.lcsd.gov.hk/tc/programmes/programmeslist/mqme_prog.html.

on the four urban lines. Considering a primary disruption can spread along metro lines and lead to service interruption at other stations (secondary disruptions), the impacts of these two types of disruptions will be superimposed on each other and hence will be virtually indistinguishable. Thus, the causal effects estimated via the synthetic control framework are the integrated impacts from both the primary disruption and its corresponding secondary disruptions. For each individual primary disruption, we implement synthetic control methods to quantify its station-level effects (direct and spillover) on four performance measures: entry/exit ridership, average journey time and average travel speed. Refer to Section 5.2.1 for a detailed definition of all measures.

The daily service time of the four lines starts at 6:00 and ends at 24:00, a time period which is then divided into 72 intervals of 15 minutes each. During the day of a disruption, the synthetic control is implemented for each 15-minute interval at the 49 stations, respectively. The estimation results of disruption impacts are discussed in the next subsection.

6.4 Results and discussion

During the study period from 1/1/2019 to 31/3/2019, our analysis covers 54 weekdays, excluding holidays and days of incomplete data. The results of the case study are presented through a randomly selected disruption, which occurred during evening peak hours at Chai Wan station (westbound), Island Line, and lasted for 27 minutes. First, we illustrate the design of synthetic control methods and how to choose the predictors and weights in the context of metro disruption. The second subsection presents the results of synthetic control estimation of four outcome measures for the disrupted station. Convincing evidence is found to support the validity of synthetic control. Finally, we visualise the spillover effects of the selected disruption spatially and temporally.

6.4.1 Modified synthetic control design

As mentioned, the time of a service day is divided into 72 intervals of 15 minutes each, and the metro station in each 15-minute interval (station-interval) is our study unit. On Monday 11/3/2019, the selected disruption occurred at 17:41 and ended at 18:08. Thus, Chai Wan

station is interrupted (treated) during this period (time interval: 47 to 48),[26] while the other 48 stations on the four urban lines were still functioning on this day. Please note that within the entire network, no other disruption occurred on the same day.[27] Under the proposed framework, a treated station-interval is compared with the synthetic control unit, which consists of the unaffected units from the same station at the same time but on different dates (donor pool). In this case study, 13 weekdays with no disruption are used to construct donor pools (refer to Section 3.2). Therefore, when we choose the predictors of metro performances, there is no need to consider station characteristics and time of day. The predictors are selected mainly based on factors that change with date. Possible predictors for all four outcome measures are summarised in Table 6.1.

As for the optimisation of weights, we divide the pre-intervention periods (time intervals: 1 to 46) into a training period (first 23 intervals) and a subsequent validation period (last 23 intervals). With the data in the above two periods, weights of predictors and donor pools are determined so that the outcomes and predictors in the treated station-interval are best replicated by the synthetic counterpart before treatment. For the disrupted station (Chai Wan), the estimation results of weights and causal impacts are discussed in the next subsection.

[26] After matching the disruption duration with time intervals, the selected disruption occurred at the end of interval 46 for approximately 3 minutes. Such an impact can be neglected, so we set this disruption to start from interval 47.

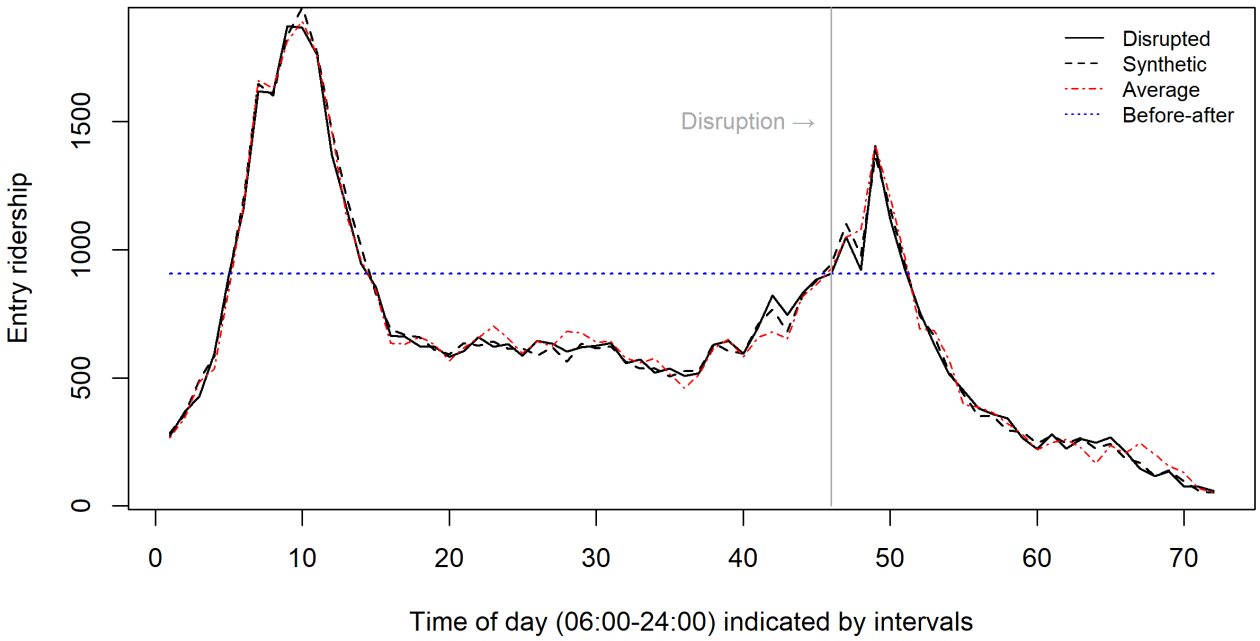
[27] If multiple disruptions occur on the same day, the synthetic control methods can only quantify their joint impacts rather than the individual impact.

Table 6.1: Potential predictors of metro performance

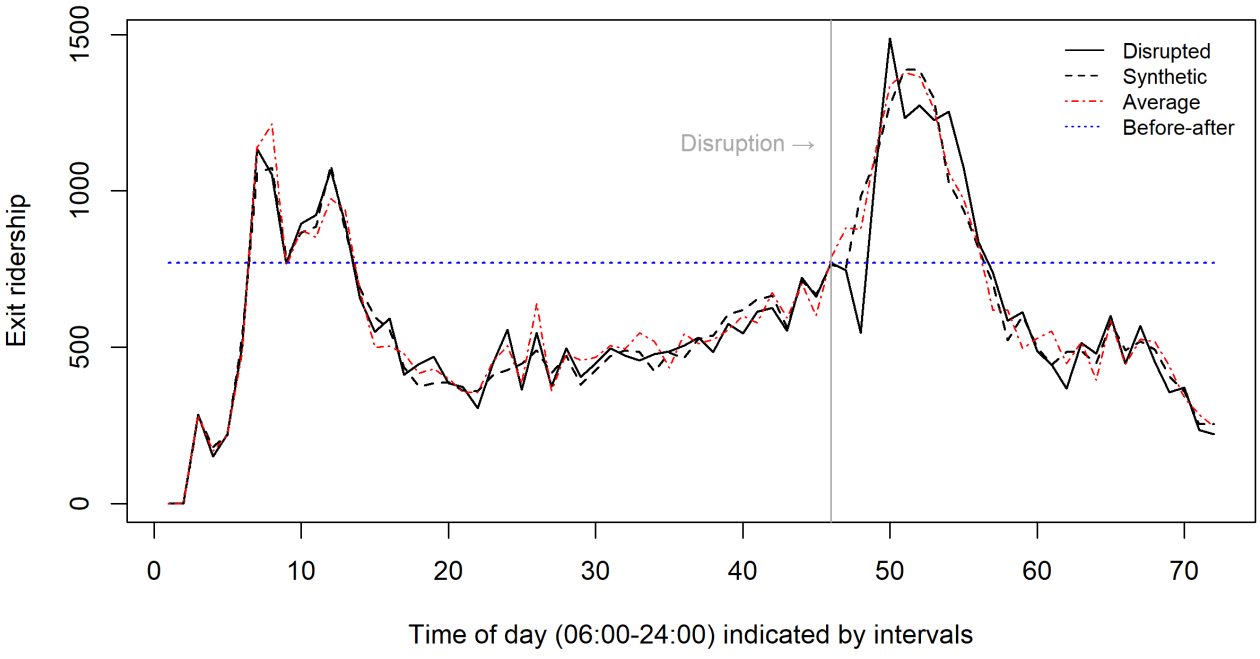
Category	Predictors	Description
Pre-intervention outcomes (15-minutes)	Entry ridership	The number of passengers that enter the study unit before the disruption starts.
	Exit ridership	The number of passengers that exit the study unit before the disruption starts.
	Average journey time	The average journey time of passengers that enter the study unit before the disruption starts.
	Average travel speed	The average travel speed of passengers that enter the study unit before the disruption starts.
Weekday	Day of week	Dummy variable, representing whether it is on the same day of the week as the disrupted date.
Weather conditions	Temperature	Atmospheric temperature around study units, ranging from 15℃ to 27℃.
	Wind speed	The wind speed around study units, ranging from 4 to 44 km/h.
	Rain status	Rain precipitation around study units, ranging from 0 to 4 mm/h.
External events	Concert	Dummy variable, representing whether there is a concert held in Hong Kong.
	Sports	Dummy variable, representing whether there is a sports match held in Hong Kong.
	Exhibition	Dummy variable, representing whether there is a large exhibition held in Hong Kong.
	Overall-mega	Dummy variable, representing whether there are external mega-events held in Hong Kong.

6.4.2 Synthetic control results of the disrupted station

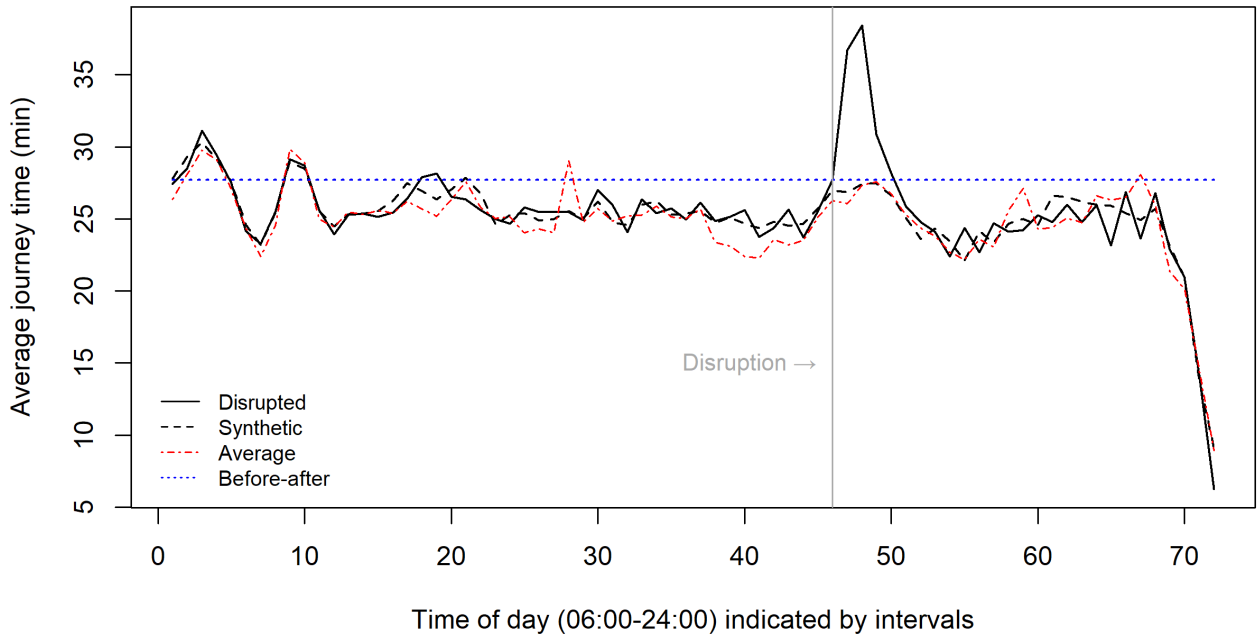
For Chai Wan station, the synthetic control estimations are displayed in Figure 6.2, which compares the trajectory of four outcome measures before and after the disruption at Chai Wan station and (i) a modified synthetic control, (ii) a simple average of the station-intervals in the donor pool and (iii) a constant value before disruption. This figure shows that a synthetic control (weight average) can closely approximate the trajectory of four outcome measures for Chai Wan station before the disruption occurrence, while the simple average sometimes fails. The naive before-and-after comparison cannot capture the changes in the trajectory at all.


Disrupted
Synthetic
Average
Before-after
Disruption →
Entry ridership
0
500
1000
1500
0
10
20
30
40
50
60
70
Time of day (06:00-24:00) indicated by intervals

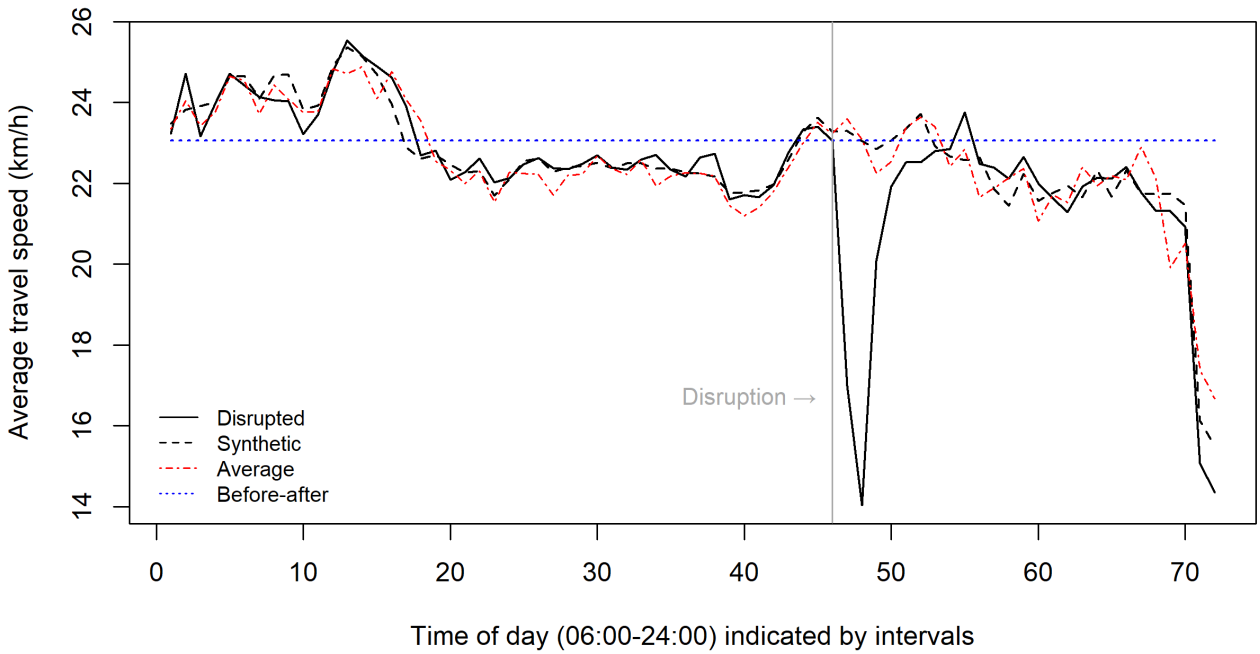
(a)


Disrupted
Synthetic
Average
Before-after
Disruption →
Exit ridership
0
500
1000
1500
0
10
20
30
40
50
60
70
Time of day (06:00-24:00) indicated by intervals

(b)



(c)



(d)

Figure 6.2: Results of synthetic control estimation and causal effects on the disrupted station – with comparison of other impact quantification methods

Table 6.2 reports the mean value of average speed predictors before the disruption, columns (1) to (4) represent $\boldsymbol{X^{a_I}}_{\mathbf{1}}$ observed on 11/3/2019, $\boldsymbol{X^{a_I}}_{\mathbf{0}}\boldsymbol{C^{a_I}}^{*}$ for the synthetic control,

$\frac{1}{J}\boldsymbol{X^{a_I}_0}$ for the simple average of donor pool and $\boldsymbol{X^{a_I}_4}$ for a single unit in the donor pool (observed on 11/2/2019). The results in Table 6.2 illustrate that the synthetic control provides a rather accurate approximation of the value of the predictors for the disrupted date. In contrast, the simple average and the single control unit both lose accuracy in the reproduction of predictors prior to the disruption. Meanwhile, we also validate the approximation of the pre-intervention outcomes as shown in Table 6.3. For all metro performance measures, the synthetic control outperforms the other two methods, which is in line with the above analysis.

Table 6.2: Mean values of predictors for average speed measures before the disruption occurred

Predictors	Disrupted station	Synthetic control	Average control	Single control
Entry ridership	796.956	795.012	794.309	809.311
Exit ridership	532.089	532.551	527.815	537.822
Ave journey time (min)	25.945	25.954	25.960	25.475
Ave speed (km/h)	23.040	23.035	23.034	22.947
Day of week (dummy)	1	0.152	0.154	0
Temperature (℃)	19.272	20.995	22.051	18.235
Wind (km/h)	7.244	10.463	13.460	13.444
Rain (mm)	0.133	0.126	0.087	0
Overall mega	0	0.421	0.612	0.822

Table 6.3: Mean square prediction errors of four outcome measures before the disruption

Outcome measures	Mean square prediction error (S.E.)		
	Synthetic control	Average control	Single control
Entry ridership	844.869 (204.994)	1874.413 (530.373)	2436.478 (579.482)
Exit ridership	1844.911 (416.405)	2356.370 (644.580)	5171.478 (1185.877)
Ave journey time	0.496 (0.093)	1.632 (0.410)	2.524 (0.467)
Ave speed	0.038 (0.018)	0.140 (0.029)	0.270 (0.062)

Table 6.4 displays the final synthetic control weights of the disrupted station $\boldsymbol{C^{a_I}}^*$. In the table, weights in a column reflect the contribution of each normal day in the donor pool $\boldsymbol{d_N}$ to the synthetic control of the disruptive day. For different outcome measures, the distribution of weights is different. Generally, dates 16/1/2019, 05/3/2019 and 25/3/2019 tend to carry greater weights, while the remaining dates are also more or less helpful for the synthetic control.

Table 6.4: Synthetic control weights of the disrupted station (for four outcome measures)

d_N	Entry ridership	Exit ridership	Ave journey time	Ave speed
09/1/2019	-	0.005	0.064	0.006
16/1/2019	0.323	0.124	0.231	0.006
11/2/2019	-	0.005	0.116	0.186
13/2/2019	0.001	0.004	0.022	0.297
21/2/2019	-	0.229	0.024	0.098
28/2/2019	-	-	0.025	0.006
05/3/2019	0.177	0.211	0.040	0.193
13/3/2019	-	-	0.025	0.006
14/3/2019	-	-	0.346	0.008
20/3/2019	-	-	0.019	0.006
25/3/2019	0.499	0.422	0.036	0.179
26/3/2019	-	-	0.030	0.006
28/3/2019	-	-	0.022	0.006

In terms of the direct disruption impacts on Chai Wan station, for passenger demand, the selected disruption significantly reduced the exit ridership (50%) during the service interruption. There is only a small impact on the entry ridership, with a decrease of just 5% during the disruption. With regards to the average journey time, this disruption dramatically increases the passenger delay by over 11 minutes per trip. Therefore, the average travel speed also experiences a sharp drop by up to 9 km/h. However, when the disruption is over, with the resumption of train services, impacts on exit ridership, average journey time and average speed reach their turning point and are gradually reduced to zero.

6.4.3 Spillover disruption effects and propagation

In this subsection, we use average travel speed as an example to illustrate how the impacts of this disruption spread to the other 48 stations spatially and temporally. In the same manner as implementing the proposed framework for the disrupted station, we obtained the weights and synthetic control estimations for other non-disrupted stations in the metro network (see Appendix C for more details). At different time intervals after the disruption occurred, Figure 6.3 visualises the spatial distribution of the impacts on average travel speed. The points in the plot represent metro stations, and their colour indicates the magnitude of speed decrease.

Severe decreases of more than 5 km/h are marked in red. No effect or minor decreases below 1 km/h are marked in blue, with three levels of orange, yellow and green in between.

The disruption occurred at the origin station of the Island Line, westbound. In the first 15 minutes of the disruption, Figure 6.3(a) shows that the second station in the downstream direction has been severely affected, and the impacts spread from the third to the seventh station. Then, during the next 15 minutes, as shown in Figure 6.3(b), the disruption impacts continue to propagate along the Island Line until the tenth station, with the first four stations all in severe delay. Eventually, the connected Tseung Kwan O Line starts to be affected. After this point, the disruption is over and train services are restored. In Figure 6.3(c), we find the speed decrease in the first four stations is declining to a moderate level. On the other hand, following the downstream direction the disruption impacts run through the entire Island Line and even spread to partial stations on the Tsuen Wan line. In Figure 6.3(d), another 30 minutes later, the spillover effects on the top two-thirds of the Island Line continue to drop until there is virtually no impact on the originally disrupted station. For the stations left, the spillover effects reach the highest level. Finally in Figure 6.3(e), one hour after the disruption, the average travel speed at most stations has returned to normal.

Based on the progress of impact propagation as described above, we conclude some insights. A disruption first affects its adjacent stations, then following the direction of train movement, such influence spreads along the metro line where it occurred. The propagation of spillover effects takes time, with impacts on downstream stations lagging behind that on upstream stations. Meanwhile, the magnitude of disruption impacts gradually decreases, as it spreads. Lines directly connected to the disrupted line are more likely to be affected. For example, in this case study, passengers on the Kwun Tong Line experience less drop in travel speed. Moreover, an interesting finding is that interchange stations with more than two metro lines are more resistant to the disturbance from disruptions, especially for journey time and travel speed. A possible explanation is that passengers in interchange stations have access to alternative routes to continue their trips, thus reducing the probability of waiting and overall waiting time.

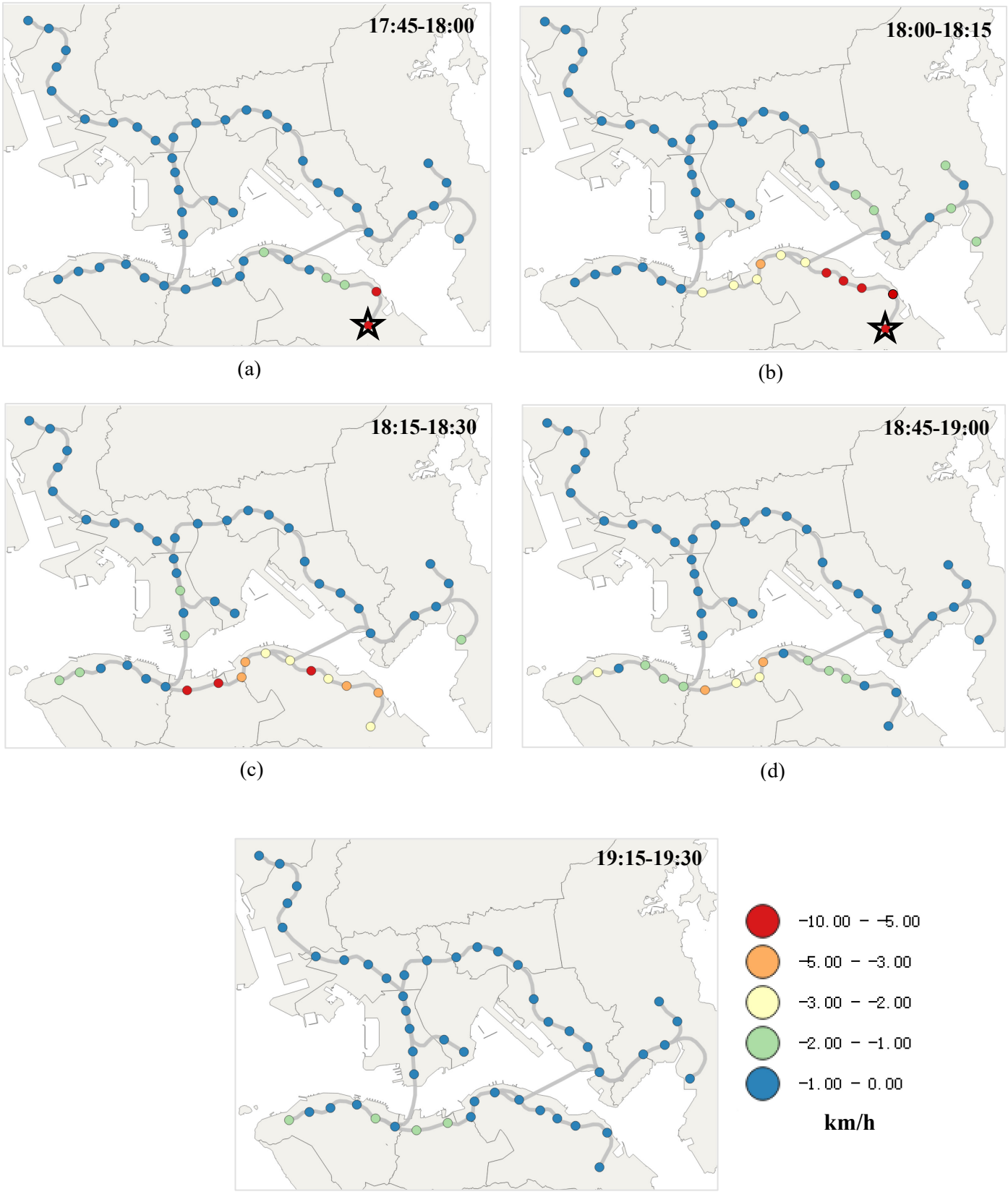


Figure 6.3: Spillover effects on average travel speed at different time periods. The star symbol indicates the location of the example disruption.

6.5 Conclusions and future work

Service disruptions pose various challenges for urban metro systems, including delays, crowding, and declining passenger satisfaction. To mitigate these adverse impacts and build effective recovery strategies, detailed information on disruptions is required. For passengers, updates on real-time disruption impacts can also help them reschedule travel plans. Thus, it is important for operators and passengers to comprehensively understand disruption impacts.

This study proposes a novel causal inference framework to quantify the direct and spillover effects of disruptions on passenger demand, average journey time and average travel speed. In contrast to previous empirical-based studies, which assume disruptions occur randomly and only focus on direct impacts, our approach not only accounts for the existence of confounding factors but also explores the wider propagation of disruption impacts. The proposed synthetic control framework relaxes the non-interference assumption; thus, disruption impacts that spread along lines throughout the entire network can be captured. Our approach compares the treated station-interval with a synthetic control unit which consists of the unaffected units within the donor pool (from the same station at the same time but on different dates). We obtain optimal weights by minimising the difference between the treated unit and the synthetic control regarding the pre-intervention predictors and outcomes of interest. Finally, we extend the estimation framework from direct impacts to spillover effects, by implementing the modified synthetic control framework on both disrupted and non-disrupted stations. To the best of our knowledge, this is the first empirical study that estimates indirect disruption impacts and analyses the propagation of such effects.

The proposed method is applied in a case study of four urban lines in the Hong Kong MTR, with a selected disruption on the Island Line. This illustrative application indicates that the modified synthetic controls provide a rather accurate approximation of the characteristics of the treated units. For the disrupted station itself, during disruptions the exit ridership drops by 50%, the average journey time increases by over 11 minutes per trip, and the passenger speed reduces by up to 9 km/h. For other stations in the metro network, the impacts of this disruption spread through the Island Line and reach part of the Tseung Kwan O Line and Tsuen Wan Line.

Let us conclude the chapter by acknowledging some of the limitations of the proposed method. The effectiveness of the modified synthetic control method heavily relies on the input data for a number of reasons. First, there must be sufficient days without disruptions, otherwise,

the donor pool will fail due to a lack of relevant data. Second, there need to have sufficient pre-treatment data. With a small number of pre-treatment periods, close or even perfect fit of the predictor values for the treated unit may be spuriously attained (Abadie, 2021), which leads to difficulties in the impact estimation of early-morning disruptions. Finally, if multiple primary disruptions occur on the same day, the synthetic control framework will estimate the joint impacts of all disruptions rather than their individual impacts.

As for future works, it could be a great idea to extend the research scope to individual-level impacts and incorporate sophisticated passenger-to-train assignment algorithms into the impact estimation.

Chapter 7
Conclusions

7.1 Main findings and contributions

This thesis presents new insights into understanding disruptions in urban metro systems. To achieve this aim, the research objectives are listed in Section 1.3. This section revisits these objectives and summarises the relevant findings and contributions.

i) Detect metro service disruptions via a data-driven probabilistic unsupervised learning approach, with identifying disruption propagation and operator's intervention.

Chapter 4 focused on the first objective. The main contribution of this research is a probabilistic GMM-based detection approach with large-scale AVL data. We used the AVL data to observe the headway of train services at platform level. The deviations between the observed headway and scheduled headway are the input of detection. Abnormal (overlong) headway deviations are regarded as a sign of service interruption. Compared with other indicators from social media data and smart card data, detection based on abnormal headway deviations is straightforward, comprehensive and more accurate. The GMM-based detection method is unique in the sense that it is a probabilistic unsupervised clustering approach, revealing the probability of each observed headway being disrupted. The parameters used in the detection are dynamically learned from the semi-simulations based on observed headway deviations, rather than being subjectively determined. This design improves the applicability of our detection model to different stations and time of day (various distributions of headway data). The empirical case study on a densely used line of the MTR showed the validity of this data-driven detection method. Benchmarked with both manual inspection logs and simulation scenarios, our detection model is highly effective for any type of service disruption.

Another notable output of this research is the analysis of secondary disruptions and recovery interventions. By merging the detection results with the train movement trajectories that are also obtained from the AVL data, we identified the relationship among the platform-

level abnormal headway deviations. These spatial and temporal links reveal the propagation progress of the interruption of services along metro lines. Furthermore, such links also enable the metro operators to perform effective interventions for service recovery.

ii) Measure the vulnerability of urban metros based on empirical causal effects of disruptions.

The second objective is to construct reliable vulnerability metrics based on empirical disruption impacts. This objective has been addressed in Chapter 5. This chapter's contribution is the introduction of the causal inference framework into metro disruption analysis. Compared to previous studies with tacit random disruption assumptions, the proposed propensity score matching methods successfully adjust for the source of confounding bias in the estimation of disruption impacts, such as weather conditions, time of day, real-time demand and frequency of historical disruptions. The proposed causal inference framework is combined with large-scale smart card data and the detected disruptions from the AVL data. We carried out a system-based analysis of metro vulnerability based on metro AVL data.

The empirical evidence in this chapter justifies the relaxation of the assumption on random disruption occurrence, i.e. the statistically significant influence of the above confounding factors on disruptions as a treatment. This finding in return illustrates the rationale of applying causal inference methods. From the estimated causal impacts, new and more reliable empirical insights into the vulnerability measurement are provided. The results of the case study suggest that vulnerability metrics are heterogeneous across metro stations, depending on their location and other station related characteristics. For the MTR in 2019, in terms of the travel demand loss and gross speed loss (overall delay), the most affected stations are more likely to be found in Hong Kong urban areas. Conversely, considering average speed loss (individual delay), the most affected stations are scattered around the suburban and extended urban areas due to a lack of alternative routes.

iii) Quantify the direct and spillover causal effects of disruptions, and analyse the impact propagation in urban metros.

The third objective has been addressed in Chapter 6. A key contribution of this chapter is that we modified the classic synthetic control methods to relax the non-interference assumption (SUTVA). Under the potential outcomes framework and considering the existence of confounding factors, the proposed modification lies in the novel design of the donor pool. We utilised the feature that metro systems reboot every day, and the panel structure of the large-scale datasets (multi-day observations). The modified donor pool consists of observations on days when no disruption occurred in the entire metro network, which eliminates the bias in synthetic controls caused by interference.

In practice, this chapter contributes to the literature by quantifying the *indirect* disruption impacts based on empirical evidence. In the presence of interference, the modified synthetic control framework is suitable for both disrupted stations and other indirectly affected stations. For the first time, we empirically explored the direct and spillover *causal* effects of disruptions in a metro network, and carried out a novel analysis of spatiotemporal propagation of disruption impacts. The case study used an example disruption of 27 minutes on the Island Line to illustrate the proposed method. The empirical results reveal that the disruption impacts spread throughout the entire Island Line and reached some further stations on other connected lines. With the increase of propagation distance, the magnitude of disruption impacts gradually decreased.

7.2 Potential applications

There are a number of potential applications of the research presented in the thesis, which can be broadly categorised into applications related to transport operations and practical policy-making. The transferability of the proposed methods to other public transport modes has also been discussed.

i) Operations related applications

- **Service disruption database** – Reliable detection results from Chapter 4 could be used to construct a high-quality database of service disruptions, including the information of occurrence location and time, duration, and disruption propagation status (primary or secondary). Disruptions in the range of two minutes to several days could be recorded, especially a large number of minor delays in train operations, with precision exact to

second. Such databases have the potential to be the cornerstone of any disruption related performance monitoring or policy assessment.

- **Metro performance evaluation** - Results from the Chapter 4 to Chapter 6 could be used in development of useful key performance indicators based on large-scale automated data.

  Reliability

  Based on the AVL data and the detection results from Chapter 4, the service reliability could be measured by new indicators such as the frequency of abnormal headway, the average train interruption time or the duration between two severe disruptions.

  Vulnerability

  As discussed in Chapter 5, metro vulnerability could be measured by empirical disruption impacts (Zhang et al., 2021a). Besides the proposed metrics related to direct impacts, new metrics could be developed to account for spillover effects or propagation characteristics.

  Resilience

  Based on the results of recovery intervention identification from Chapter 4, indicators such as response time after disruption, the number of interventions and effective rate imply the ability of metro operators to mitigate adverse disruption impacts, so-called recoverability. From the system perspective, such recoverability may be measured by a new indicator: recovery time from the end of disruption to services fully restored.

  Metro benchmarking

  The integrated research design of this thesis is applicable and scalable to the majority of metro systems in the globe, with automatic fare collection system and automatic vehicle location system. The proposed performance indicators enable the smart benchmarking based on big data.

- **Disruption management** – With the comprehensive analytics of metro disruptions, from their occurrence to corresponding impacts, the findings in this thesis could support the prevention of future disruptions and prediction of disruption impacts.

  Prediction of disruption occurrence and duration

  Mainly based on the detection results from Chapter 4 (offline – historical disruption data), and the supplementary information such as passenger demand, weather

conditions, external events, topological and station characteristics, advanced machine learning techniques could be applied to predict the probability and duration of future disruptions. Such knowledge could be used as a reliable input for timetable recovery or train rescheduling after disruptions.

Disruption impact prediction - offline

With the above-mentioned datasets, the estimated historical disruption impacts could be used to predict passenger delays, demand loss or crowding caused by future disruptions.

- **Passenger information provision (online service)** – The proposed detection method in Chapter 4 and the impact quantification method in Chapter 6 could be applied to provide real-time information of disruptions to passengers. The frequency of information updates will be determined by the minimum interval of GMM detection and impact estimation, for example per 15 minutes. Such short-term information of train delays and passenger delays could be incorporated into the online route planners (Cats and Jenelius, 2014; Drabicki et al., 2021).

  To provide alerts on disruption occurrence and severity, the required database includes static data such as metro topology, facility status, service schedules and station characteristics, as well as dynamic data such as passenger demand, weather conditions and external events. For monitoring a single platform, the running time on a laptop featuring 2.80 GHz CPU and 16 GB RAM can be less than 5 minutes. When the proposed application is equipped with high performance computing facilities and parallel design to handle multiple platforms, it would be suitable for monitoring large metro networks under the expected information refresh rate.

ii) Applications in practical policy-making

- **Appraisal of metro investments** – The vulnerability metrics derived in Chapter 5 are important inputs in the economic appraisal of metro projects, particularly for the investments in system maintenance and facility upgrades. Such disaggregate performance metrics could reveal the most vulnerable parts of the network, and guide the allocation of resources to achieve the maximum improvements. In important

intermediate step in achieving this goal is to transform the vulnerability metrics into monetary valuations of loss of social welfare.

iii) Transferability to other public transport modes

- **Disruption detection** – The feasibility depends on whether the distribution of headway deviations of other modes has similar characteristics to that of the metro. Since metro systems use fully enclosed lines, dedicated tracks and signals, the independent operation ensures less variation in train headway under normal conditions, unless service disruption occurs. For similar modes, such as light rail and commuter rail, the GMM-based detection method is suitable. However, for other public transport modes operated on urban roads, such as bus, tram and taxi, their service headway may experience large variations due to different road conditions. In this case, the probabilistic detection method may not be applicable.
- **Impact estimation/vulnerability measurement** – the proposed causal inference methods can be easily transferred to other transport modes, provided that there are enough relevant data of these modes, such as disruption occurrence, performance measures, weather conditions and system characteristics.

7.3 Future research

A number of potential avenues for future research related to the work in this thesis are summarised below.

For disruption detection, the indicator of service interruptions is overlong headway, which is extracted from the AVL data. Since the AVL data only contain information on train locations and the inherent characteristics of trains, there is no way to infer the cause of disruption from it. This is a limitation of our data-driven detection method, as compared to manual inspections. A feasible solution is to include more sources of data in the detection process, such as the manual inspection logs, social media data, news and smart card data. Future research may focus on merging the AVL-based detection results with these new data to infer the reason for disruptions.

For disruption impact estimation, this thesis concentrates on station-level impacts rather than impacts on the individual level. The journey time and travel speed measures are the averages of all passengers that enter a station in a 15-minute interval. Inspired by Yap et al. (2017, 2021), it is worth focusing on individual-level impacts. By merging the AVL data and smart card data, we can infer the most possible route (with transfer choice) for each passenger within the metro network via passenger to train assignment algorithms (Hörcher et al., 2017; Zhu et al., 2017b). With such information, the disruption impacts could be estimated for individual passengers, especially under the synthetic control framework. It would be interesting to integrate sophisticated assignments into the proposed causal inference framework.

For both impact estimation and vulnerability measurement, this research only considers the disruption impacts within metro systems. This is due to limited access to the SCD and AVL data of other public transport modes. If possible, it would be interesting to extend these studies to other transport modes. For example, by merging data from other public transport modes (e.g., bus, urban rails, shared bike or taxi) with metro datasets, we could estimate multi-modal vulnerability metrics in the same causal inference framework and understand the characteristics of the mode shift due to disruptions. It would also be interesting to analyse the propagation of disruption impacts between different modes (Yap et al., 2021).

For the application of a modified synthetic control framework, it is worth mentioning that the data requirements of this method are relatively high. First, the construction of the donor pool requires a certain numbers of unaffected study units. In the context of this research, this would mean that there have to be a sufficient number of days where no disruption occurs in the entire metro system. However, in some cases, this condition may not be met. The insufficient donor pool will lead to biased estimates. It would be interesting to analyse how the composition of the donor pool influences the estimation of disruption impacts. Second, the credibility of synthetic control estimators partially depends on sufficient pre-treatment information (Abadie, 2021). Thus, for disruptions that occurred within the first fifteen-minute interval, there is no such pre-treatment data. It would be necessary to prepare alternative solutions for this particular case.

References

Abadie, A. (2021). Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects. *Journal of Economic Literature*, *59*(2), 391–425.

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, *105*(490), 493–505.

Abadie, A., & Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, *93*(1), 113–132.

Adjetey-Bahun, K., Birregah, B., Châtelet, E., & Planchet, J.-L. (2016). A model to quantify the resilience of mass railway transportation systems. *Reliability Engineering & System Safety*, *153*, 1–14.

Ahmed, M., Naser Mahmood, A., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, *60*, 19–31.

An, W. (2018). Causal Inference with Networked Treatment Diffusion. *Sociological Methodology*, *48*(1), 152–181.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434), 444-455.

Aronow, P. M. (2012). A General Method for Detecting Interference Between Units in Randomized Experiments. *Sociological Methods & Research*, *41*(1), 3–16.

Aronow, P. M., & Samii, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, *11*(4).

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. *The Review of Economics and Statistics*, *60*(1), 47.

Ashenfelter, O., & Card, D. (1985). Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs. *The Review of Economics and Statistics*, *67*(4), 648.

Augusteijn, M. F., & Folkert, B. A. (2002). Neural network classification and novelty detection. *International Journal of Remote Sensing*, *23*(14), 2891–2902.

Bagchi, M., & White, P. R. (2005). The potential of public transport smart card data. *Transport Policy*, *12*(5), 464–474.

Bansal, P., Daziano, R. A., & Guerra, E. (2018). Minorization-Maximization (MM) algorithms for semiparametric logit models: Bottlenecks, extensions, and comparisons. *Transportation Research Part B: Methodological*, *115*, 17–40.

Barabino, B., Di Francesco, M., & Mozzoni, S. (2015). Rethinking bus punctuality by integrating Automatic Vehicle Location data and passenger patterns. *Transportation Research Part A: Policy and Practice*, *75*, 84–95.

Basse, G., & Feller, A. (2018). Analyzing Two-Stage Experiments in the Presence of Interference. *Journal of the American Statistical Association*, *113*(521), 41–55.

Berdica, K. (2002). An introduction to road vulnerability: what has been done, is done and should be done. *Transport Policy*, *9*(2), 117–127.

Bereziński, P., Jasiul, B., & Szpyrka, M. (2015). An Entropy-Based Network Anomaly Detection Method. *Entropy*, *17*(4), 2367–2408.

Bertini, R. L., & El-Geneidy, A. (2003). Generating Transit Performance Measures with Archived Data. *Transportation Research Record: Journal of the Transportation Research Board*, *1841*(1), 109–119.

Billio, M., Casarin, R., & Rossini, L. (2019). Bayesian nonparametric sparse VAR models. *Journal of Econometrics*, *212*(1), 97–115.

Bishop, C. M. (1994). Novelty detection and neural network validation. *IEE Proceedings - Vision, Image, and Signal Processing*, *141*(4), 217.

Bojinov, I., Chen, A., & Liu, M. (2020). The Importance of Being Causal. *Harvard Data Science Review*.

Bolla, R., & Davoli, F. (2000). Road traffic estimation from location tracking data in the mobile cellular network. *2000 IEEE Wireless Communications and Networking Conference. Conference Record (Cat. No.00TH8540)*.

Brazil, W., White, A., Nogal, M., Caulfield, B., O'Connor, A., & Morton, C. (2017). Weather and rail delays: Analysis of metropolitan rail in Dublin. *Journal of Transport Geography*, *59*, 69–76.

Briand, A.-S., Côme, E., Khouadjia, M., & Oukhellou, L. (2019). Detection of Atypical Events on a Public Transport Network Using Smart Card Data. *European Transport Conference 2019*.

Cano, J. A. L., Kovaceva, J., Lindman, M., & Brännström, M. (2009). Automatic incident detection and classification at intersections. *Volvo Car Corporation*, 09–0234.

Cao, J., & Dowd, C. (2019). *Estimation and inference for synthetic control methods with spillover effects*. *arXiv preprint arXiv:1902.07343*.

Cats, O. (2013). Multi-agent transit operations and assignment model. *Procedia Computer Science*, 19, 809-814.

Cats, O., & Jenelius, E. (2014). Dynamic Vulnerability Analysis of Public Transport Networks: Mitigation Effects of Real-Time Information. *Networks and Spatial Economics*, *14*(3-4), 435–463.

Cats, O., & Jenelius, E. (2018). Beyond a complete failure: the impact of partial capacity degradation on public transport network vulnerability. *Transportmetrica B: Transport Dynamics*, *6*(2), 77–96.

Chan, J. (2007). *Rail transit OD matrix estimation and journey time reliability metrics using automated fare data* [Doctoral dissertation]. Massachusetts Institute of Technology

Chen, J., Roberts, C., & Weston, P. (2008). Fault detection and diagnosis for railway track circuits using neuro-fuzzy systems. *Control Engineering Practice*, *16*(5), 585–596.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Chen, T., He, T., Benesty, M., & Khotilovich, V. (2021). Package 'xgboost'.

Chopra, S. S., Dillon, T., Bilec, M. M., & Khanna, V. (2016). A network-based framework for assessing infrastructure resilience: a case study of the London metro system. *Journal of the Royal Society Interface*, *13*(118), 20160113.

Cook, T. D. (2008). "Waiting for Life to Arrive": A history of the regression-discontinuity design in Psychology, Statistics and Economics. *Journal of Econometrics*, *142*(2), 636–654.

D'Andrea, E., & Marcelloni, F. (2017). Detection of traffic congestion and incidents from GPS trace analysis. *Expert Systems with Applications*, *73*, 43–56.

Davy, M., & Godsill, S. (2002). Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation. *IEEE International Conference on Acoustics Speech and Signal Processing*, 1313–1316.

De Bruin, T., Verbert, K., & Babuška, R. (2017). Railway Track Circuit Fault Diagnosis Using Recurrent Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(3), 523–533.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Derrible, S., & Kennedy, C. (2010). The complexity and robustness of metro networks. *Physica A: Statistical Mechanics and Its Applications*, *389*(17), 3678–3691.

Drabicki, A., Kucharski, R., Cats, O., & Szarata, A. (2020). Modelling the effects of real-time crowding information in urban public transport systems. *Transportmetrica A: Transport Science*, 1–39.

El Attar, A., Khatoun, R., & Lemercier, M. (2014). Clustering-based anomaly detection for smartphone applications. *2014 IEEE Network Operations and Management Symposium (NOMS)*, 1–4.

Eskin, E. (2000). Anomaly Detection over Noisy Data Using Learned Probability Distributions. *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000)*, 255–262.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., & Stolfo, S. (2002). A Geometric Framework for Unsupervised Anomaly Detection. *Advances in Information Security*, 77–101.

Faturechi, R., & Miller-Hooks, E. (2015). Measuring the Performance of Transportation Infrastructure Systems in Disasters: A Comprehensive Review. *Journal of Infrastructure Systems*, *21*(1), 04014025.

Forastiere, L., Airoldi, E. M., & Mealli, F. (2021). Identification and Estimation of Treatment and Interference Effects in Observational Studies on Networks. *Journal of the American Statistical Association*, *116*(534), 901–918.

Forastiere, L., Mealli, F., & VanderWeele, T. J. (2016). Identification and Estimation of Causal Mechanisms in Clustered Encouragement Designs: Disentangling Bed Nets Using Bayesian Principal Stratification. *Journal of the American Statistical Association*, *111*(514), 510–525.

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: Journal of the Transportation Research Board*, *2343*(1), 17–24.

Graham, D. J., McCoy, E. J., & Stephens, D. A. (2014). Quantifying Causal Effects of Road Network Capacity Expansions on Traffic Volume and Density via a Mixed Model Propensity Score Estimator. *Journal of the American Statistical Association*, *109*(508), 1440–1449.

Grossi, G., Lattarulo, P., Mariani, M., Mattei, A., & Öner, Ö. (2020). *Synthetic Control Group Methods in the Presence of Interference: The Direct and Spillover Effects of Light Rail on Neighborhood Retail Activity*. *arXiv preprint arXiv:2004.05027*.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *Journal of Computational and Graphical Statistics*, *2*(4), 405–420.

Gu, Y., Qian, Z. (Sean), & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies*, *67*(3), 321–342.

Heckman, J. J., Ichimura, H., & Todd, P. (1998). Matching As An Econometric Evaluation Estimator. *Review of Economic Studies*, *65*(2), 261–294.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Horbury, A. X. (1999a). Guidelines for specifying automatic vehicle location and real-time passenger information systems using current best practice. *Transport Reviews*, *19*(4), 331–351.

Horbury, A. X. (1999b). Using non-real-time Automatic Vehicle Location data to improve bus services. *Transportation Research Part B: Methodological*, *33*(8), 559–579.

Hörcher, D. (2017). *The economics of crowding in urban rail transport* (Doctoral dissertation, Imperial College London).

Hörcher, D., Graham, D. J., & Anderson, R. J. (2017). Crowding cost estimation with large scale smart card and vehicle location data. *Transportation Research Part B: Methodological*, *95*, 105–125.

Hudgens, M. G., & Halloran, M. E. (2008). Toward Causal Inference With Interference. *Journal of the American Statistical Association*, *103*(482), 832–842.

Imbens, G. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710.

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences : an introduction*. Cambridge University Press.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, *47*(1), 5–86.

James, A., Jie, W., Xulei, Y., Chenghao, Y., Ngan, N. B., Yuxin, L., Yi, S., Chandrasekhar, V., & Zeng, Z. (2018). TrackNet - A Deep Learning Based Fault Detection for Railway Track Inspection. *2018 International Conference on Intelligent Rail Transportation (ICIRT)*, *2018*, 1–5.

Jang, W. (2010). Travel Time and Transfer Analysis Using Transit Smart Card Data. *Transportation Research Record: Journal of the Transportation Research Board*, *2144*(1), 142–149.

Jasperse, F. (2020). Automated offline detection of disruptions using smart card data: A case study of the metro network of Washington DC. *TU Delft*.

Jenelius, E., Petersen, T., & Mattsson, L.-G. (2006). Importance and exposure in road network vulnerability analysis. *Transportation Research Part A: Policy and Practice*, *40*(7), 537–560.

Jespersen-Groth, J., Potthoff, D., Clausen, J., Huisman, D., Kroon, L., Maróti, G., & Nielsen, M. N. (2009). Disruption Management in Passenger Railway Transportation. *Robust and Online Large-Scale Optimization*, 399–421.

Ji, T., Fu, K., Self, N., Lu, C.-T., & Ramakrishnan, N. (2018). Multi-Task Learning for Transit Service Disruption Detection. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

Jiang, R., Lu, Q.-C., & Peng, Z.-R. (2018). A station-based rail transit network vulnerability measure considering land use dependency. *Journal of Transport Geography*, *66*, 10–18.

Kang, H., & Imbens, G. (2016). *Peer Encouragement Designs in Causal Inference with Partial Interference and Identification of Local Average Network Effects*. *arXiv preprint arXiv:1609.04464*.

Kim, K., Oh, K., Lee, Y. K., Kim, S., & Jung, J.-Y. (2014). An analysis on movement patterns between zones using smart card data in subway networks. *International Journal of Geographical Information Science*, *28*(9), 1781–1801.

Kind, A., Stoecklin, M., & Dimitropoulos, X. (2009). Histogram-based traffic anomaly detection. *IEEE Transactions on Network and Service Management*, *6*(2), 110–121.

Kruegel, C., Mutz, D., Robertson, W., & Valeur, F. (2013). Bayesian event classification for intrusion detection. *19th Annual Computer Security Applications Conference, 2003. Proceedings.*

Kuang, K., Li, L., Geng, Z., Xu, L., Zhang, K., Liao, B., Huang, H., Ding, P., Miao, W., & Jiang, Z. (2020). Causal Inference. *Engineering*, *6*(3), 253–263.

Kusakabe, T., & Asakura, Y. (2014). Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies*, *46*, 179–191.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. *Econometric Evaluation of Labour Market Policies*, 43–58.

Lee, S., & Hickman, M. D. (2011). Travel Pattern Analysis Using Smart Card Data of Regular Users. *Transportation Research Board 90th Annual Meeting*.

Leng, N., De Martinis, V., Corman, F. (2018). Agent-based simulation approach for disruption management in rail schedule. *Proceedings of the 14th Conference on Advanced Systems in Public Transport (CASPT)*, Brisbane, Australia.

Li, W., Peng, Q., Wen, C., Wang, P., Lessan, J., & Xu, X. (2020). Joint optimization of delay-recovery and energy-saving in a metro system: A case study from China. *Energy*, *202*, 117699.

Liao, Y., & Vemuri, V. Rao. (2002). Use of K-Nearest Neighbor classifier for intrusion detection. *Computers & Security*, *21*(5), 439–448.

Liu, L., Hudgens, M. G., & Becker-Dreps, S. (2016). On inverse probability-weighted estimators in the presence of interference. *Biometrika*, *103*(4), 829–842.

London Datastore. (2018). *Data Quality Standards*. https://data.london.gov.uk/about/data-quality-standards/

Lord, E., Willems, M., Lapointe, F.-J., & Makarenkov, V. (2017). Using the stability of objects to determine the number of clusters in datasets. *Information Sciences*, *393*, 29–46.

Lu, Q. C. (2018). Modeling network resilience of rail transit under operational incidents. *Transportation Research Part A: Policy and Practice*, 117, 227-237.

M'cleod, L., Vecsler, R., Shi, Y., Levitskaya, E., Kulkarni, S., Malinchik, S., & Sobolevsky, S. (2017). Vulnerability of Transportation Networks: The New York City Subway System under Simultaneous Disruptive Events. *Procedia Computer Science*, *119*, 42–50.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F., & Liu, J. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, *36*, 1–12.

Ma, Z., Ferreira, L., & Mesbah, M. (2014). Measuring Service Reliability Using Automatic Vehicle Location Data. *Mathematical Problems in Engineering*, *2014*, 1–12.

Mahmassani, H. S., Haas, C., Zhou, S., & Peterman, J. (1999). *Evaluation of Incident Detection Methodologies*. University of Texas at Austin. Center for Transportation Research.

Malandri, C., Fonzone, A., & Cats, O. (2018). Recovery time and propagation effects of passenger transport disruptions. *Physica A: Statistical Mechanics and its Applications*, *505*, 7-17.

Marra, A. D., & Corman, F. (2020). From Delay to Disruption: Impact of Service Degradation on Public Transport Networks. *Transportation Research Record: Journal of the Transportation Research Board*, *2674*(10), 886–897.

Mass Transit Railway. (2019). *KEEP MOVING Annual Report 2019*. https://www.annualreports.com/HostedData/AnnualReports/PDF/OTC_MTRJF_2019.pdf

Mattsson, L.-G., & Jenelius, E. (2015). Vulnerability and resilience of transport systems – A discussion of recent research. *Transportation Research Part A: Policy and Practice*, *81*, 16–34.

Melo, P. C., Harris, N. G., Graham, D. J., Anderson, R. J., & Barron, A. (2011). Determinants of Delay Incident Occurrence in Urban Metros. *Transportation Research Record: Journal of the Transportation Research Board*, *2216*(1), 10–18.

Mesbah, M., Currie, G., Lennon, C., & Northcott, T. (2012). Spatial and temporal visualization of transit operations performance data at a network level. *Journal of Transport Geography*, *25*, 15–26.

Metropolitan Transportation Authority. (2019). *NYC Subway Action Plan*. http://www.mtamovingforward.com/files/NYC_Subway_Action_Plan.pdf

Morency, C., Trépanier, M., & Agard, B. (2007). Measuring transit use variability with smart-card data. *Transport Policy*, *14*(3), 193–203.

Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies*, *24*, 9–18.

Nian, G., Chen, F., Li, Z., Zhu, Y., & Sun, D. (Jian). (2019). Evaluating the alignment of new metro line considering network vulnerability with passenger ridership. *Transportmetrica A: Transport Science*, *15*(2), 1402–1418.

O'Kelly, M. E. (2015). Network Hub Structure and Resilience. *Networks and Spatial Economics*, *15*(2), 235–251.

Park, J. Y., Kim, D.-J., & Lim, Y. (2008). Use of Smart Card Data to Define Public Transit Use in Seoul, South Korea. *Transportation Research Record: Journal of the Transportation Research Board*, *2063*(1), 3–9.

Patcha, A., & Park, J.-M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Computer Networks*, *51*(12), 3448–3470.

Paulsen, M., Rasmussen, T.K., Anker Nielsen, O. (2018). Modelling railway-induced passenger delays in multi-modal public transport networks. *Proceedings of the 14th Conference on Advanced Systems in Public Transport (CASPT)*, Brisbane, Australia.

Pearl, J. (2000). *Casuality : models, reasoning, and inference*. Cambridge University Press.

Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, *10*(4), 339–348.

Pelletier, M.-P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*(4), 557–568.

Pnevmatikou, A., & Karlaftis, M. (2011). *Demand changes from metro line closures*. European Transport Conference, Glasgow, Scotland.

Reggiani, A., Nijkamp, P., & Lanzi, D. (2015). Transport resilience and vulnerability: The role of connectivity. *Transportation Research Part A: Policy and Practice*, *81*, 4–15.

Rigdon, J., & Hudgens, M. G. (2015). Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine*, *34*(6), 924–935.

Riter, S., & McCoy, J. (1977). Automatic vehicle location—An overview. *IEEE Transactions on Vehicular Technology*, *26*(1), 7–11.

Riveiro, M., Lebram, M., & Elmer, M. (2017). Anomaly Detection for Road Traffic: A Visual Analytics Framework. *IEEE Transactions on Intelligent Transportation Systems*, *18*(8), 2260–2270.

Rjabovs, A., & Palacin, R. (2017). The influence of system design-related factors on the safety performance of metro drivers. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, *231*(3), 317–328.

Rodríguez-Núñez, E., & García-Palomares, J. C. (2014). Measuring the vulnerability of public transport networks. *Journal of Transport Geography*, *35*, 50–63.

Rosenbaum, P. R. (1989). Optimal Matching for Observational Studies. *Journal of the American Statistical Association*, *84*(408), 1024–1032.

Rosenbaum, P. R. (2002). *Observational studies*. Springer.

Rosenbaum, P. R. (2007). Interference Between Units in Randomized Experiments. *Journal of the American Statistical Association*, *102*(477), 191–200.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*, *39*(1), 33–38.

Rossi, R., Gastaldi, M., Gecchele, G., & Barbaro, V. (2015). Fuzzy Logic-based Incident Detection System using Loop Detectors Data. *Transportation Research Procedia*, *10*, 266–275.

Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection*. Wiley.

Rubin, D. B. (1973a). The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biometrics*, *29*(1), 185–203.

Rubin, D. B. (1973b). Matching to Remove Bias in Observational Studies. *Biometrics*, *29*(1), 159.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics*, *2*(1), 1.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, *6*(1), 34–58.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, *25*(3), 279–292.

Rubin, D. B., & Thomas, N. (2000). Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates. *Journal of the American Statistical Association*, *95*(450), 573–585.

Rubin, G. J., Brewin, C. R., Greenberg, N., Simpson, J., & Wessely, S. (2005). Psychological and behavioural reactions to the bombings in London on 7 July 2005: cross sectional survey of a representative sample of Londoners. *BMJ*, *331*(7517), 606.

Santhosh, K. K., Dogra, D. P., & Roy, P. P. (2020). Anomaly Detection in Road Traffic Using Visual Surveillance. *ACM Computing Surveys*, *53*(6), 1–26.

Shelat, S., & Cats, O. (2017, June). Measuring spill-over effects of disruptions in public transport networks. In 2017 5th IEEE international conference on models and technologies for intelligent transportation systems (MT-ITS) (pp. 756-761). IEEE.

Silva, R., Kang, S. M., & Airoldi, E. M. (2015). Predicting traffic volumes and estimating the effects of shocks in massive transportation systems. *Proceedings of the National Academy of Sciences*, *112*(18), 5643–5648.

Smith, Bivens, Embrechts, Palagiri, & Szymanski. (2002). Clustering approaches for anomaly based intrusion detection. *Proceedings of Intelligent Engineering Systems through Artificial Neural Networks*, *12*, 579–584.

Smith, H. L. (1997). 6. Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies. *Sociological Methodology*, *27*(1), 325–353.

Sobel, M. E. (2006). What Do Randomized Studies of Housing Mobility Demonstrate? *Journal of the American Statistical Association*, *101*(476), 1398–1407.

Sodemann, A. A., Ross, M. P., & Borghetti, B. J. (2012). A Review of Anomaly Detection in Automated Surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part c (Applications and Reviews)*, *42*(6), 1257–1272.

Srinivasan, S., & Ferreira, J. (2002). Travel behavior at the household level: understanding linkages with residential choice. *Transportation Research Part D: Transport and Environment*, *7*(3), 225–242.

Steenbruggen, J., Tranos, E., & Rietveld, P. (2016). Traffic incidents in motorways: An empirical proposal for incident detection using data from mobile phone operators. *Journal of Transport Geography*, *54*, 81–90.

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, *25*(1), 1–21.

Sun, D. (Jian), & Guan, S. (2016). Measuring vulnerability of urban metro network from line operation perspective. *Transportation Research Part A: Policy and Practice*, *94*, 348–359.

Sun, D., Zhao, Y., & Lu, Q.-C. (2015). Vulnerability Analysis of Urban Rail Transit Networks: A Case Study of Shanghai, China. *Sustainability*, *7*(6), 6919–6936.

Sun, H., Wu, J., Wu, L., Yan, X., & Gao, Z. (2016). Estimating the influence of common disruptions on urban rail transit networks. *Transportation Research Part A: Policy and Practice*, *94*, 62–75.

Sun, L., Huang, Y., Chen, Y., & Yao, L. (2018). Vulnerability assessment of urban rail transit based on multi-static weighted method in Beijing, China. *Transportation Research Part A: Policy and Practice*, *108*, 12–24.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309–317.

Tirachini, A., Godachevich, J., Cats, O., Muñoz, J. C., & Soza-Parra, J. (2021). Headway variability in public transport: a review of metrics, determinants, effects for quality of service and control strategies. *Transport Reviews*, 1–25.

Toledo, T., Cats, O., Burghout, W., & Koutsopoulos, H. N. (2010). Mesoscopic simulation for transit operations. *Transportation Research Part C: Emerging Technologies*, *18*(6), 896–908.

Tonnelier, E., Baskiotis, N., Guigue, V., & Gallinari, P. (2018). Anomaly detection in smart card logs and distant evaluation with Twitter: a robust framework. *Neurocomputing*, *298*, 109–121.

Transport for London. (2017). *London Underground Performance Report Period 1-13 2016/17*. http://content.tfl.gov.uk/lu-performance-report-period-13-2016-17.pdf

Transport for London. (2019). *Underground services performance*. https://tfl.gov.uk/corporate/publications-and-reports/underground-services-performance

Trépanier, M., Morency, C., & Agard, B. (2009). Calculation of Transit Performance Measures Using Smartcard Data. *Journal of Public Transportation*, *12*(1), 79–96.

UITP- International Union of Public Transport. (2018). *WORLD METRO FIGURES 2018*. http://www.uitp.org/sites/default/files/cck-focus-papers files/Statistics%20Brief%20-%20World%20metro%20figures%202018V4_WEB.pdf

Utsunomiya, M., Attanucci, J., & Wilson, N. (2006). Potential Uses of Transit Smart Card Registration and Transaction Data to Improve Transit Planning. *Transportation Research Record: Journal of the Transportation Research Board*, *1971*(1), 118–126.

van der Laan, M. J. (2014). Causal Inference for a Population of Causally Connected Units. *Journal of Causal Inference*, *2*(1).

VanderWeele, T. J., Tchetgen Tchetgen, E. J., & Halloran, M. E. (2012). Components of the Indirect Effect in Vaccine Trials: Identification of Contagion and Infectiousness Effects. *Epidemiology*, *23*(5), 751–761.

Verbitsky-Savitz, N., & Raudenbush, S. W. (2012). Causal Inference Under Interference in Spatial Settings: A Case Study Evaluating Community Policing Program in Chicago. *Epidemiologic Methods*, *1*(1).

Vic Barnett, & Tolewis. (1994). *Outliers in statistical data*. Wiley.

Wan, X., Li, Q., Yuan, J., & Schonfeld, P. M. (2015). Metro passenger behaviors and their relations to metro incident involvement. *Accident Analysis & Prevention*, *82*, 90–100.

Wei, X., Yang, Z., Liu, Y., Wei, D., Jia, L., & Li, Y. (2019). Railway track fastener defect detection based on image processing and deep learning techniques: A comparative study. *Engineering Applications of Artificial Intelligence*, *80*, 66–81.

Welankiwar, A., Sherekar, S., Bhagat, A. P., & Khodke, P. A. (2018). Fault Detection in Railway Tracks Using Artificial Neural Networks. *2018 International Conference on Research in Intelligent and Computing in Engineering (RICE)*.

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*, 79–82.

Wong, W.-K., Moore, A., Cooper, G., & Wagner, M. (2002). Rule-Based Anomaly Pattern Detection for Detecting Disease Outbreaks. *AAAI/IAAI*, 217–223.

Yan, Y., Liu, Z., & Bie, Y. (2016). Performance Evaluation of Bus Routes Using Automatic Vehicle Location Data. *Journal of Transportation Engineering*, *142*(8), 04016029.

Yang, X., Chen, A., Wu, J., Gao, Z., & Tang, T. (2018). An energy-efficient rescheduling approach under delay perturbations for metro systems. *Transportmetrica B: Transport Dynamics*, *7*(1), 386–400.

Yang, Y., Liu, Y., Zhou, M., Li, F., & Sun, C. (2015). Robustness assessment of urban rail transit based on complex network theory: A case study of the Beijing Subway. *Safety Science*, *79*, 149–162.

Yap, M. D., Cats, O., van Oort, N., & Hoogendoorn, S. P. (2017). A robust transfer inference algorithm for public transport journeys during disruptions. *Transportation Research Procedia*, *27*, 1042–1049.

Yap, M., & Cats, O. (2020). Predicting disruptions and their passenger delay impacts for public transport stops. *Transportation*, *48*, 1703–1731.

Yap, M., Cats, O., Törnquist Krasemann, J., van Oort, N., & Hoogendoorn, S. (2021). Quantification and control of disruption propagation in multi-level public transport networks. *International Journal of Transportation Science and Technology*.

Ye, Q., & Kim, H. (2019). Assessing network vulnerability of heavy rail systems with the impact of partial node failures. *Transportation*, *46*(5), 1591–1614.

Yin, H., Han, B., Li, D., & wang, Y. (2016). Evaluating Disruption in Rail Transit Network: A Case Study of Beijing Subway. *Procedia Engineering*, *137*, 49–58.

Yu, J., Stettler, M. E. J., Angeloudis, P., Hu, S., & Chen, X. (Michael). (2020). Urban network-wide traffic speed estimation with massive ride-sourcing GPS traces. *Transportation Research Part C: Emerging Technologies*, *112*, 136–152.

Zhang, D., Du, F., Huang, H., Zhang, F., Ayyub, B. M., & Beer, M. (2018a). Resiliency assessment of urban rail transit networks: Shanghai metro as an example. *Safety Science*, *106*, 230–243.

Zhang, J., Wang, S., & Wang, X. (2018b). Comparison analysis on vulnerability of metro networks based on complex network. *Physica A: Statistical Mechanics and Its Applications*, *496*, 72–78.

Zhang, J., Xu, X., Hong, L., Wang, S., & Fei, Q. (2011). Networked analysis of the Shanghai subway network, in China. *Physica A: Statistical Mechanics and Its Applications*, *390*(23-24), 4562–4570.

Zhang, N., Graham, D. J., Hörcher, D., & Bansal, P. (2021a). A causal inference approach to measure the vulnerability of urban metro systems. *Transportation*.

Zhang, S., Chen, X., & Wu, F. (2010). Analysis and application of automated data collection system in subway. *6th Advanced Forum on Transportation of China (AFTC 2010)*, 216–220.

Zhang, X., Deng, Y., Li, Q., Skitmore, M., & Zhou, Z. (2016). An incident database for improving metro safety: The case of shanghai. *Safety Science*, *84*, 88–96.

Zhang, X., Zheng, Y., Zhao, Z., Liu, Y., Blumenstein, M., & Li, J. (2021b). Deep learning detection of anomalous patterns from bus trajectories for traffic insight analysis. *Knowledge-Based Systems*, *217*.

Zhu, L. (2019). *Spatio-temporal traffic anomaly detection for urban networks* (Doctoral dissertation, Imperial College London).

Zhu, L., Guo, F., Krishnan, R., & Polak, J. W. (2018). A Deep Learning Approach for Traffic Incident Detection in Urban Networks. *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*.

Zhu, S., Masud, H., Xiong, C., Yang, Z., Pan, Y., & Zhang, L. (2017a). Travel Behavior Reactions to Transit Service Disruptions. *Transportation Research Record: Journal of the Transportation Research Board*, *2649*(1), 79–88.

Zhu, Y., & Goverde, R. M. P. (2019). Dynamic Passenger Assignment for Major Railway Disruptions Considering Information Interventions. *Networks and Spatial Economics*, *19*(4), 1249–1279.

Zhu, Y., Koutsopoulos, H. N., & Wilson, N. H. M. (2017b). A probabilistic Passenger-to-Train Assignment Model based on automated data. *Transportation Research Part B: Methodological*, *104*, 522–542.

Zulfiqar, O., Chang, Y.-C., Chen, P.-H., Fu, K., Lu, C.-T., Solnick, D., & Li, Y. (2020). RISECURE: Metro Incidents And Threat Detection Using Social Media. *2020*

IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 531–535.

Appendix A
Supplementary Material: Chapter 4

A.1 Robustness check against percentile of headway deviations in simulation

As mentioned in Section 4.4.2, the percentile of observed headway deviations, which we use to construct undisrupted observations in the simulation, affects the composition of the synthetic samples and detection accuracy. The higher the percentile is, the more overlap will be between the disrupted and regular headway deviations, potentially leading to lower detection accuracy. To evaluate the robustness of our GMM-based model in terms of this concern, we test for the following percentiles: 96, 97, 98, 99. The proportion of simulated disruptions in the sample data is kept at 5%.

Table A.1: Robustness check against the choice of deviation percentile in simulation

Percentile choice	5% Disruptions			
	Precision	Recall	Accuracy	Optimal threshold
96 percentile	1.000	0.945	0.997	0.994
97 percentile	1.000	0.942	0.997	0.989
98 percentile	1.000	0.937	0.996	0.968
99 percentile	1.000	0.935	0.995	0.956

Table A.1 summarises the changing trend of performance measures. Although the overall accuracy and recall rate continue to drop when the percentile increases, their minimum values are still over 0.99 and 0.93, respectively. Such results indicate that the proposed detection model is robust, even if the simulated deviations are generated from different percentile of empirical data.

A.2 Sensitivity analysis of detection accuracy

Besides the mentioned example station (1) and time interval (7:30-8:00) in Section 5, we select another suburban station (2) of the studied metro line and an additional evening-peak interval (17:00-17:30) to make the validation more comprehensive. In validation, two types of disruptions are considered. The first type is minor disruption that follows log-normal distribution and ranges between 1.5 to 8 minutes. The second type also follows log-normal distribution but includes a wider range of headway deviations, consisting of both minor interruptions and more severe interruptions of few hours. The empirical headway deviation distribution is truncated at the 95th percentile, when simulating undisrupted observations. Under different combinations of stations, disruption types and disruption rates, the performance measures of detection are presented in Table A.2.

Table A.2: Results of sensitivity analysis for two example stations (1000 runs)

Disruption type	Station (Northbound)	Time	Disruption rate	Performance measures			
				Ave. precision	Ave. recall	Ave. F-score	Ave. accuracy
Minor interruptions (< 8 minutes)	Example 1	7:30-8:00	1%	1	0.9441	0.9678	0.9995
			2%	1	0.9453	0.9752	0.9989
			5%	1	0.9465	0.9720	0.9974
		17:00-17:30	1%	1	0.9999	1	1
			2%	1	0.9999	0.9999	1
			5%	1	0.9998	0.9999	1
	Example 2	7:30-8:00	1%	1	0.9033	0.9393	0.9992
			2%	1	0.9074	0.9425	0.9983
			5%	1	0.9157	0.9543	0.9960
		17:00-17:30	1%	1	0.9823	0.9899	0.9998
			2%	1	0.9839	0.9910	0.9997
			5%	1	0.9870	0.9934	0.9994
Mixed interruptions (minor - few hours)	Example 1	7:30-8:00	1%	1	0.9934	0.9963	0.9999
			2%	1	0.9931	0.9962	0.9999
			5%	1	0.9928	0.9960	0.9999
		17:00-17:30	1%	1	1	1	1
			2%	1	0.9999	1	1
			5%	1	0.9999	0.9999	0.9999
	Example 2	7:30-8:00	1%	1	0.9917	0.9950	0.9999
			2%	1	0.9916	0.9960	0.9999

		5%	1	0.9916	0.9956	0.9996
	17:00-17:30	1%	1	0.9973	0.9985	1
		2%	1	0.9973	0.9984	0.9999
		5%	1	0.9972	0.9984	0.9999

We find that the detection performance varies slightly in each platform-interval, due to the different composition of headway deviation data. Overall, the detection is effective for both minor and mixed disruptions, under any given disruption rate.

Appendix B Supplementary Material: Chapter 5

B.1 Balance improvements under different matching methods

Table B.1: Means of confounding factors before and after matching

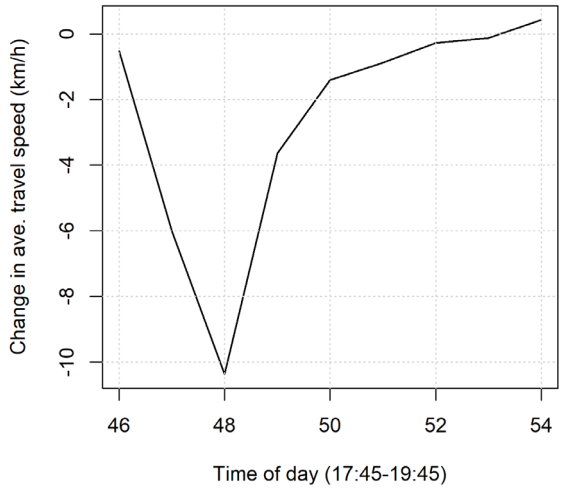
Confounding factors	Mean value		Balance before matching	t-Test p-value	Mean value			Balance after matching*	t-Test* p-value	Improvement * %
	Treatment	Control before matching			Control after matching					
					Subclassification	NN=2, W	**NN=1, W**			
Past disruptions	2.030	1.631	0.399	0.006	1.835	1.882	**1.950**	0.080*	0.681*	79.9*
Time 1 (7:30-10:30)	0.333	0.333	0	1	0.333	0.333	**0.333**	0*	1*	0*
Time 3 (16:30-19:30)	0.323	0.323	0	1	0.323	0.323	**0.323**	0*	1*	0*
Temperature	20.444	21.179	-0.735	0.008	21.046	20.575	**20.531**	0.113*	0.7128	84.6*
Rain	0.153	0.067	0.086	0.121	0.078	0.093	**0.132**	0.021*	0.784*	75.6*
Overground	0.394	0.394	0	1	0.394	0.394	**0.394**	0*	1*	0*
Pre-15 min entry ridership	1249.394	1206.966	42.428	0.769	1213.667	1226.384	**1231.980**	17.414*	0.906*	59.0*
Overground*Wind speed	4.687	5.241	-0.554	0.560	5.016	4.350	**4.479**	0.208*	0.893*	62.5*

Note: * denotes the results responding to nearest neighbour matching 1:1, with replacement.
NN - nearest neighbour matching, W – with replacement.

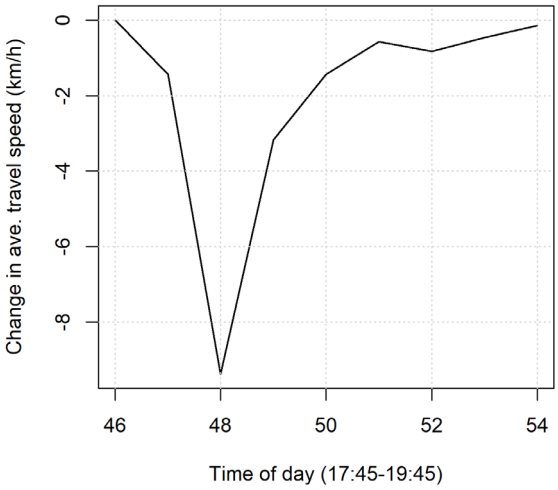
Appendix C

Supplementary Material: Chapter 6

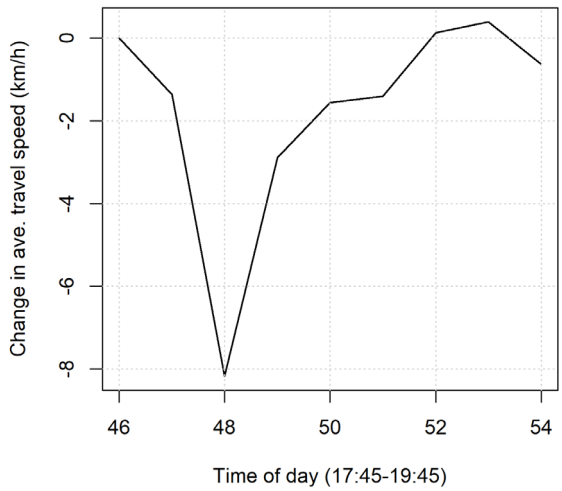
C.1 Estimated spillover effects on average travel speed over time



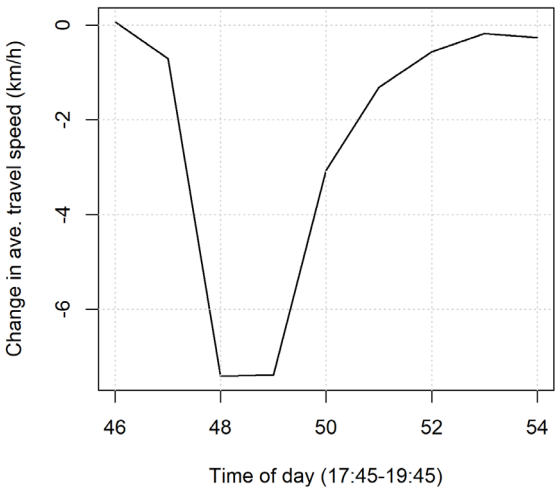
(1) Heng Fa Chuen



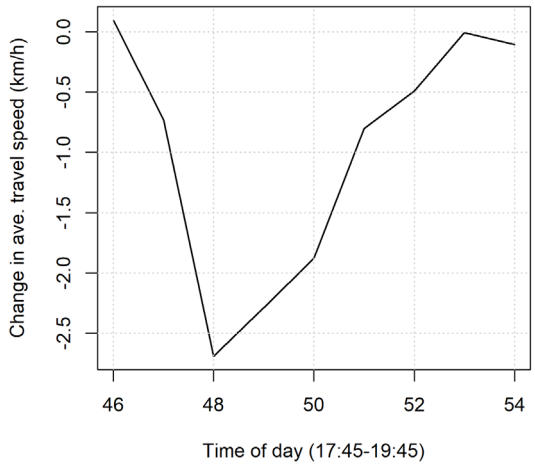
(2) Shau Kei Wan



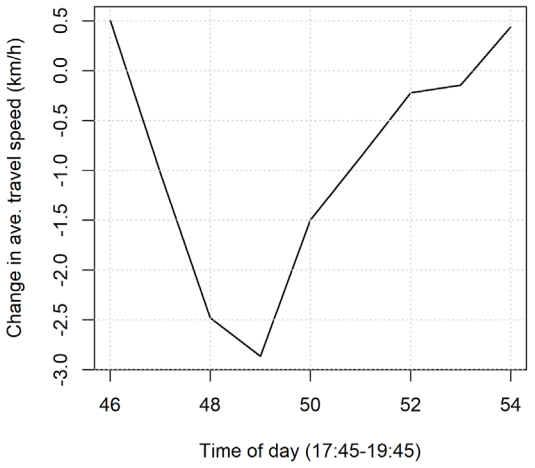
(3) Sai Wan Ho



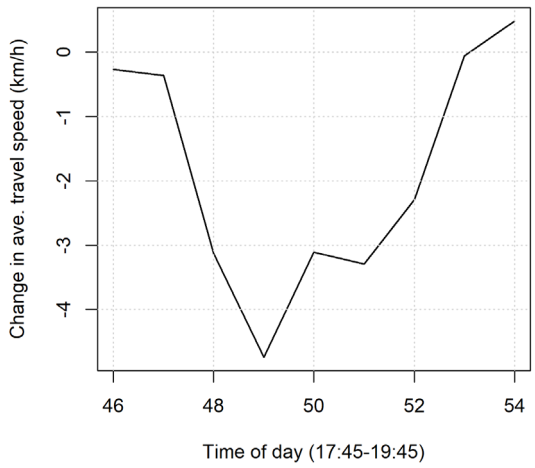
(4) Tai Koo



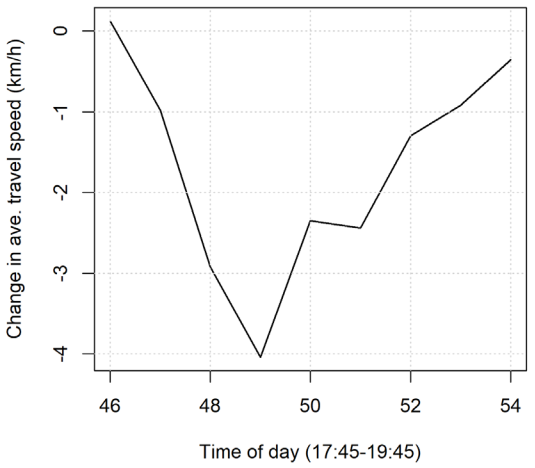
(5) Quarry Bay



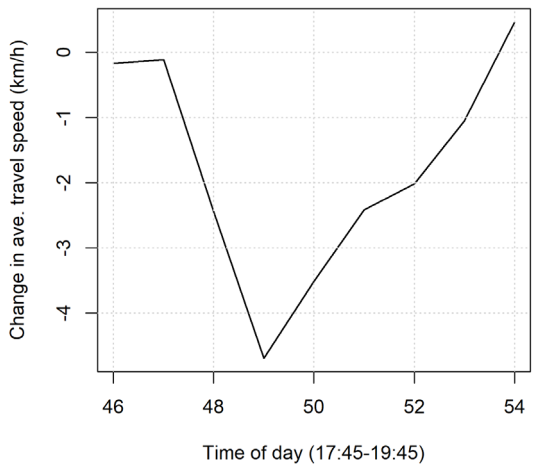
(6) North Point



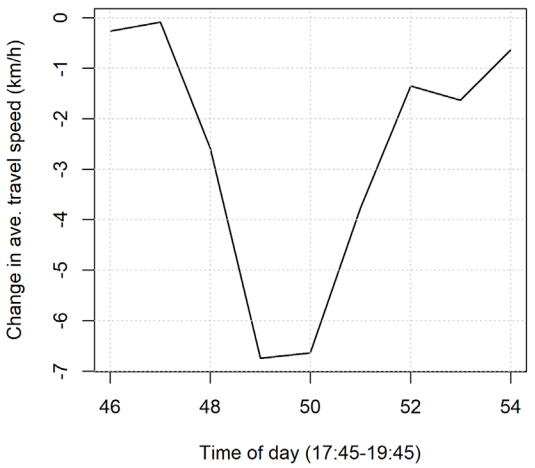
(7) Fortress Hill



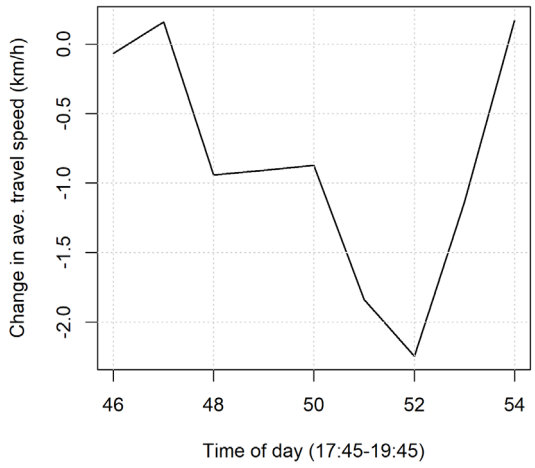
(8) Tin Hau



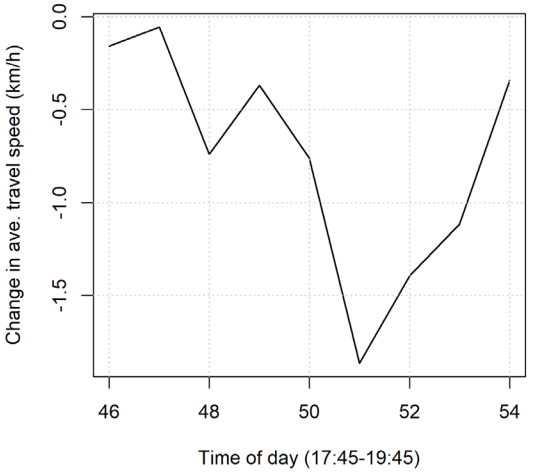
(9) Causeway Bay



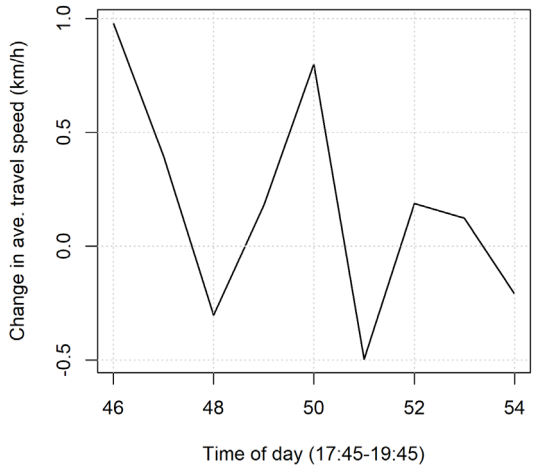
(10) Wan Chai



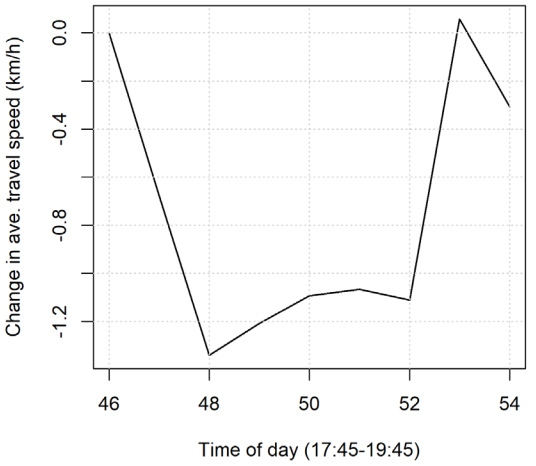
(11) Admiralty



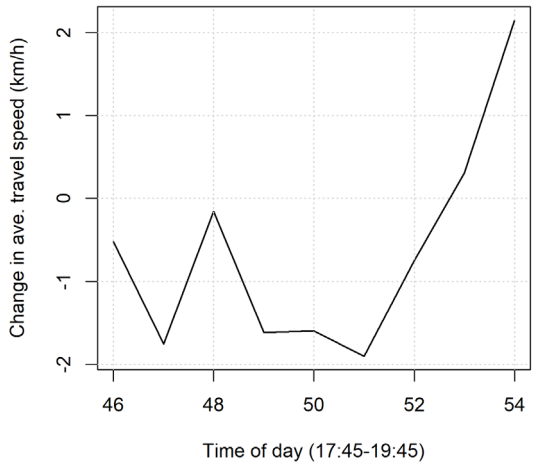
(12) Central



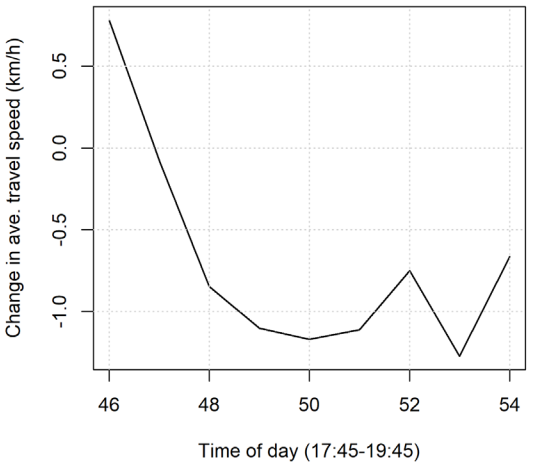
(13) Sheung Wan



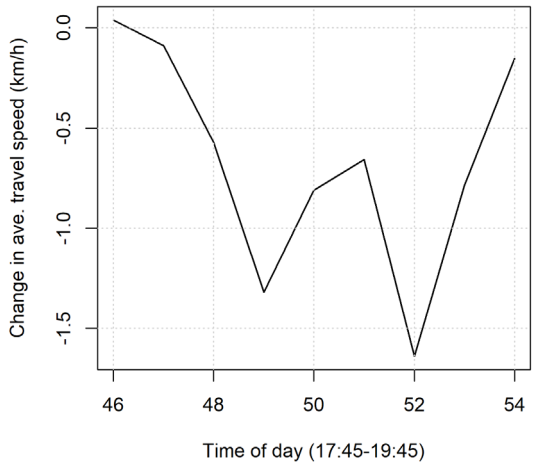
(14) Sai Ying Pun



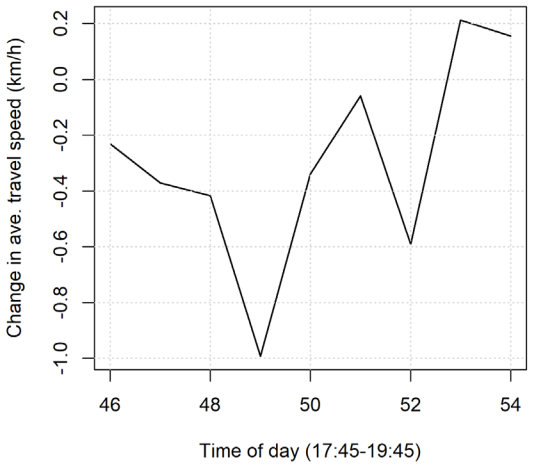
(15) HKU



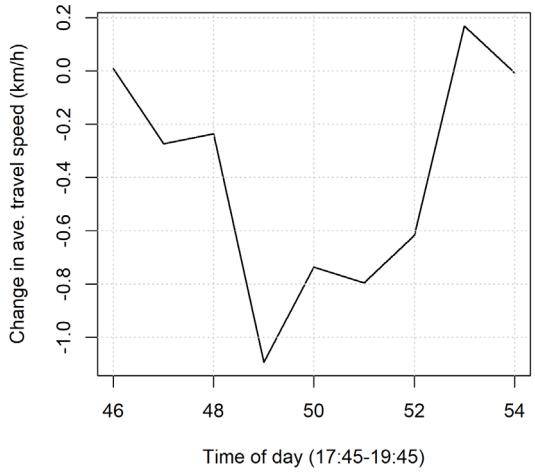
(16) Kennedy Town



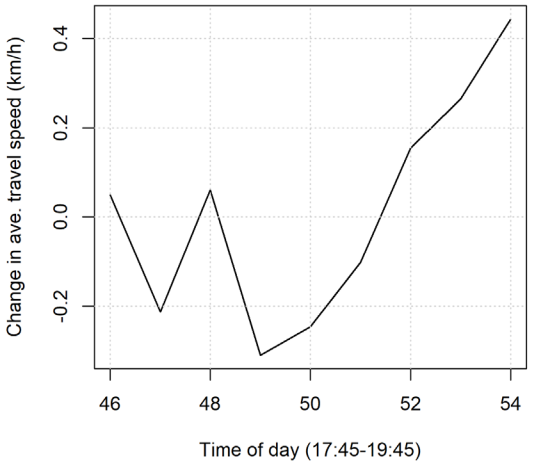
(17) Tsim Sha Tsui



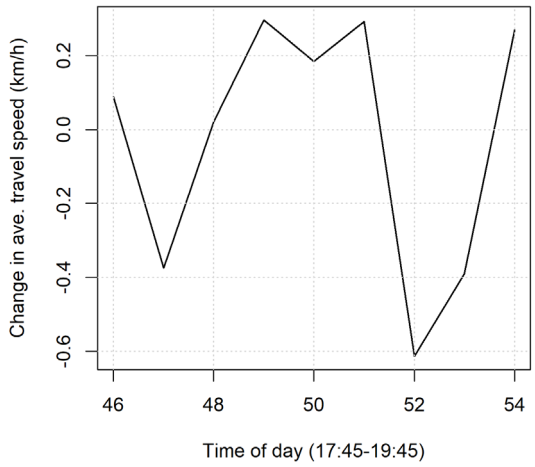
(18) Jordan



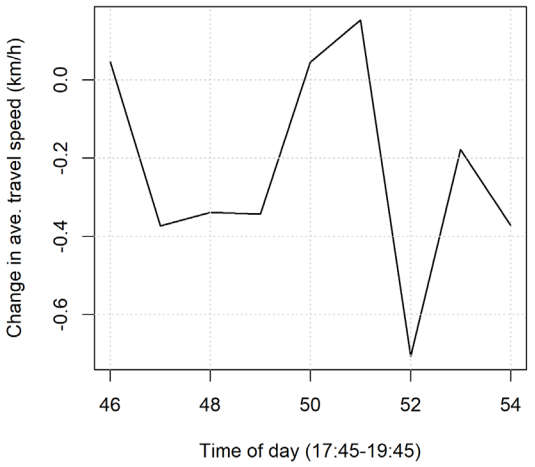
(19) Yau Ma Tei



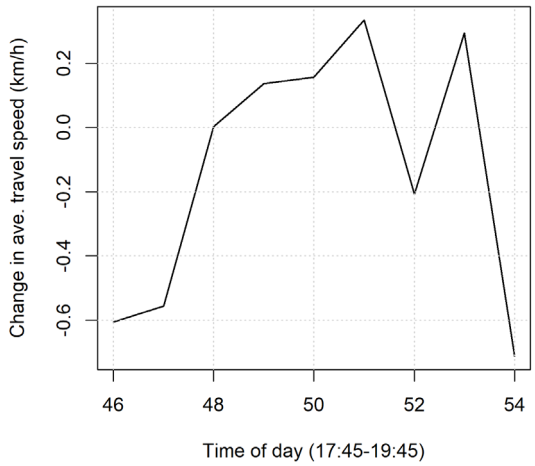
(20) Mong Kok



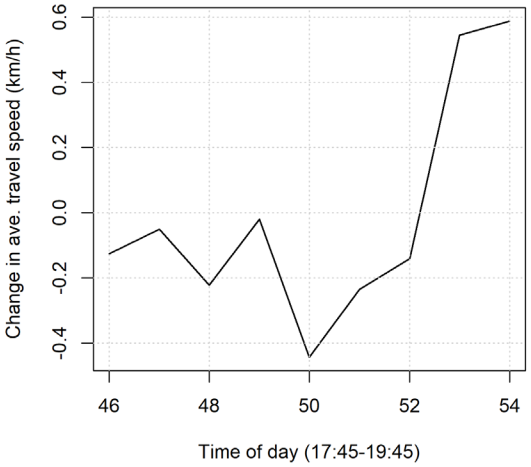
(21) Prince Edward



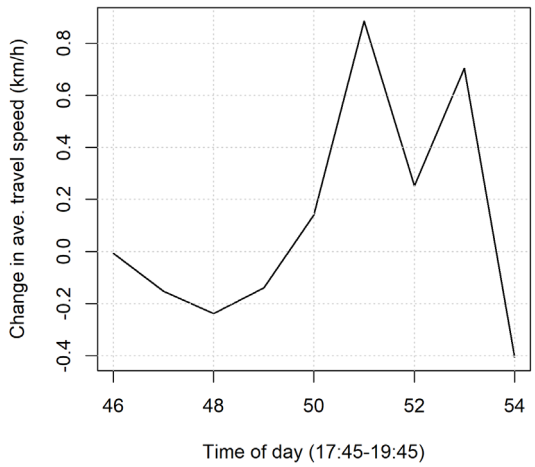
(22) Sham Shui Po



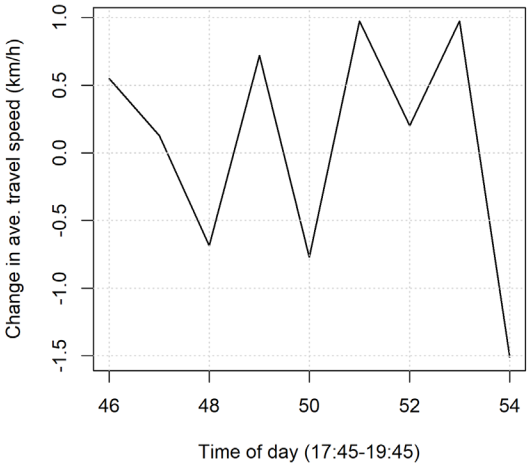
(23) Cheung Sha Wan



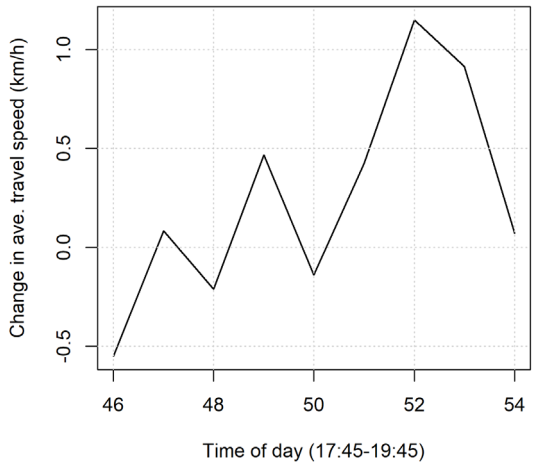
(24) Lai Chi Kok



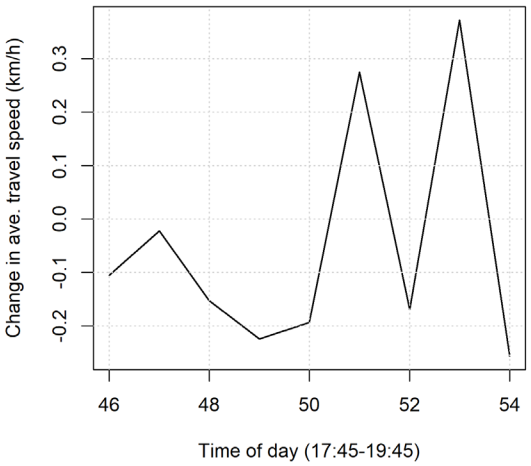
(25) Mei Foo



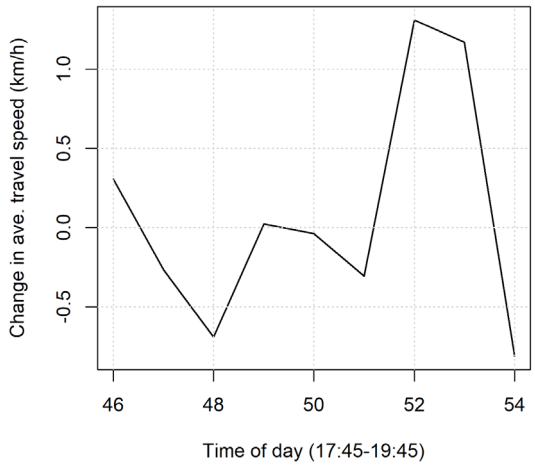
(26) Lai King



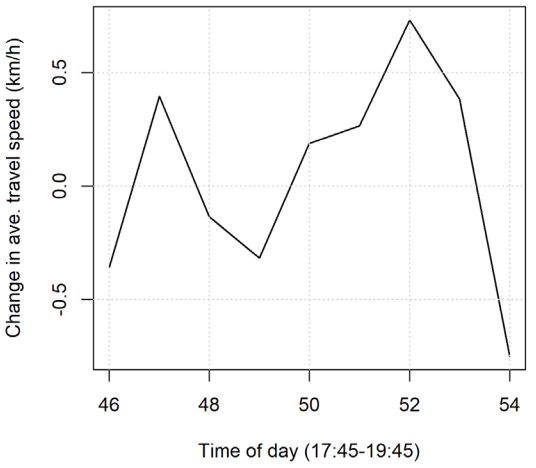
(27) Kwai Fong



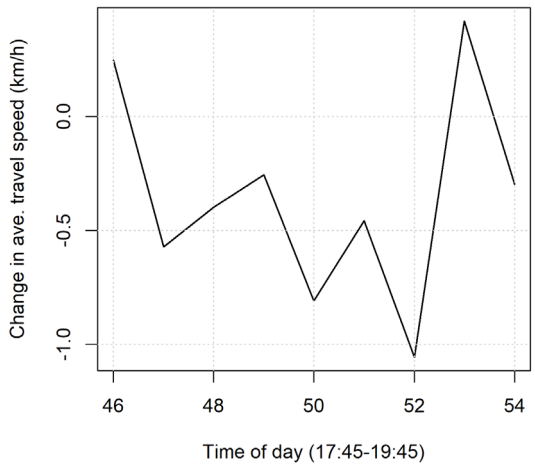
(28) Kwai Hing



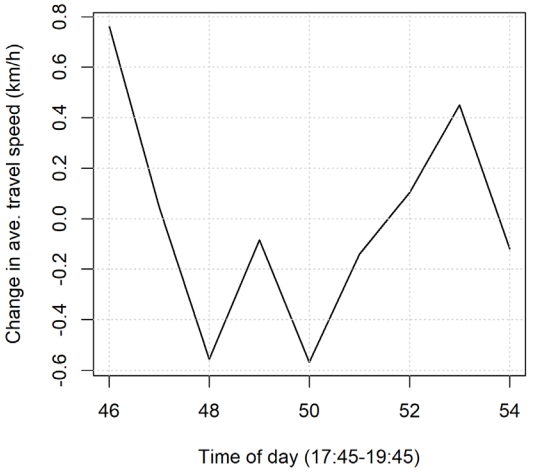
(29) Tai Wo Hau



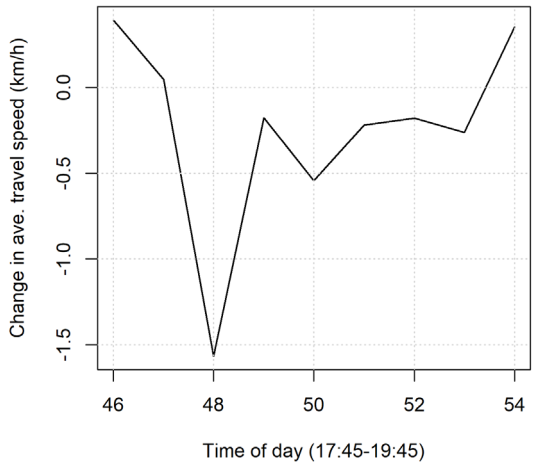
(30) Tsuen Wan



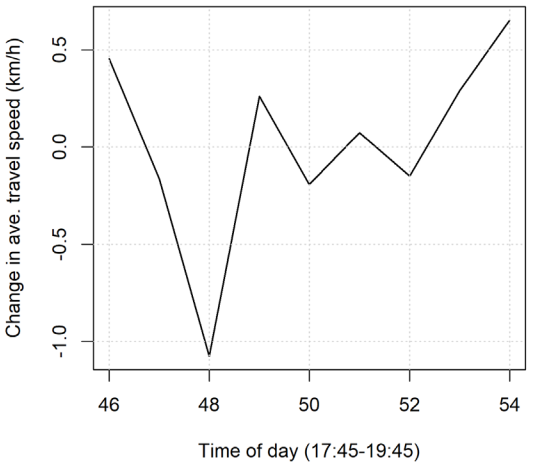
(31) Tiu Keng Leng



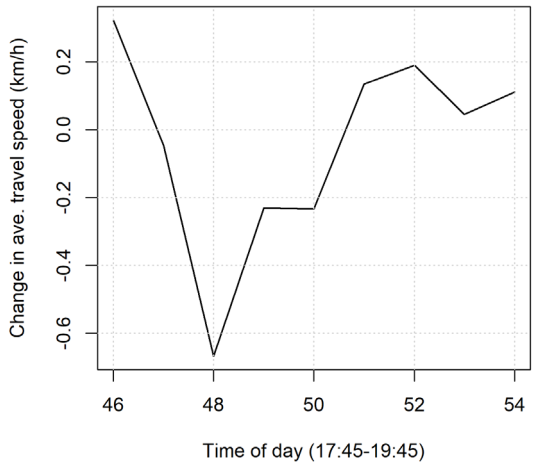
(32) Yau Tong



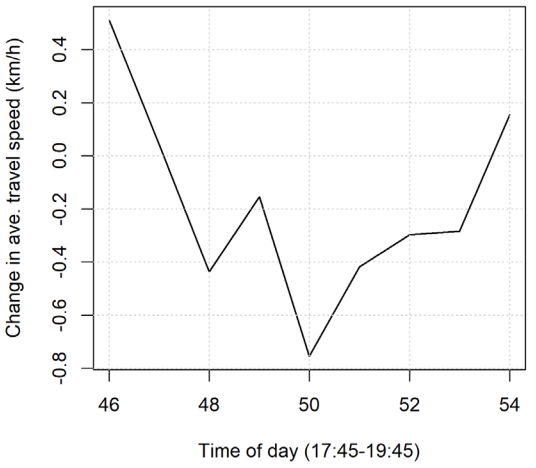
(33) Lam Tin



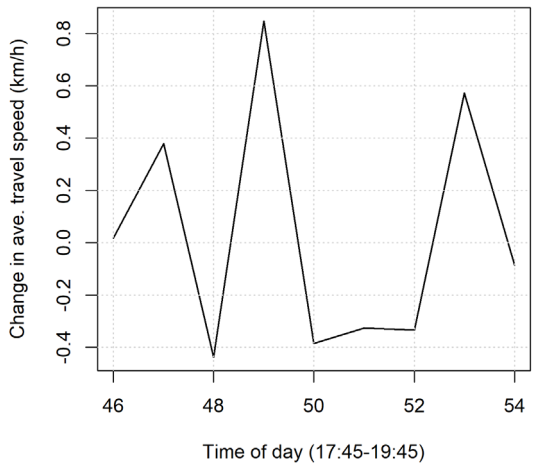
(34) Kwun Tong



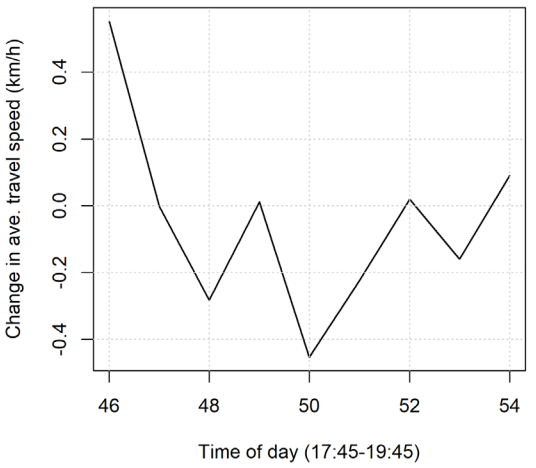
(35) Ngau Tau Kok



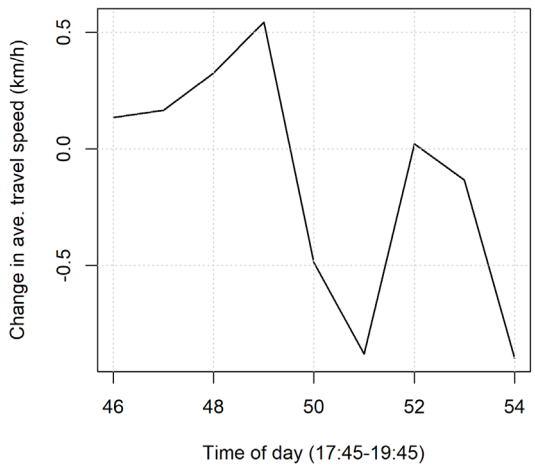
(36) Kowloon Bay



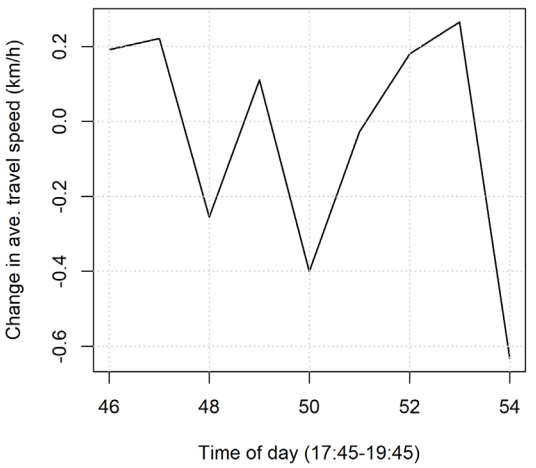
(37) Choi Hung



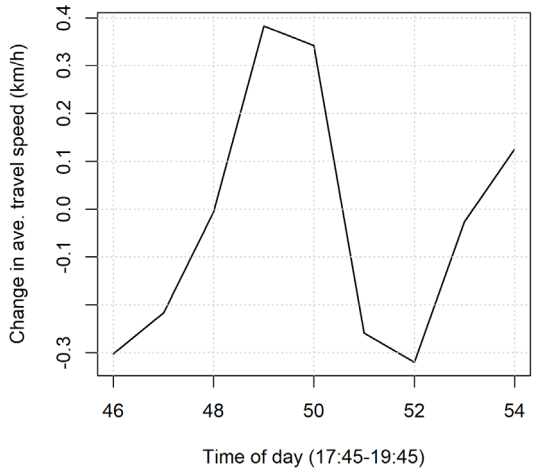
(38) Diamond Hill



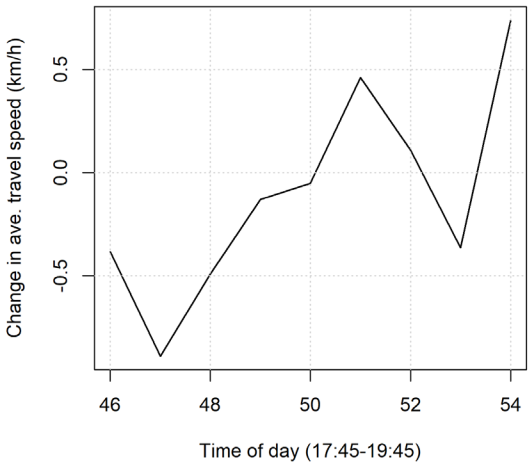
(39) Wong Tai Sin



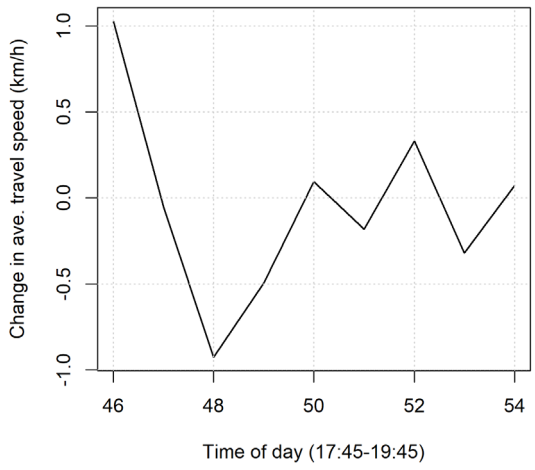
(40) Lok Fu



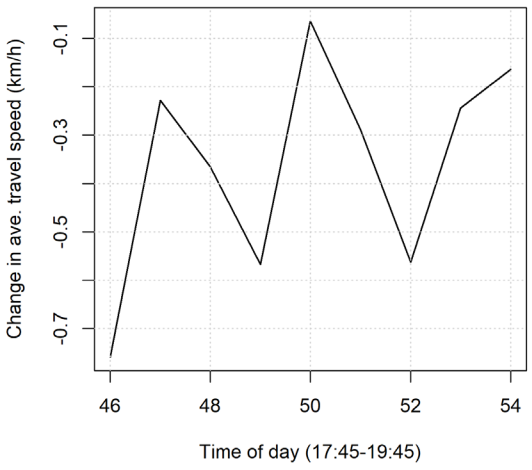
(41) Kowloon Tong



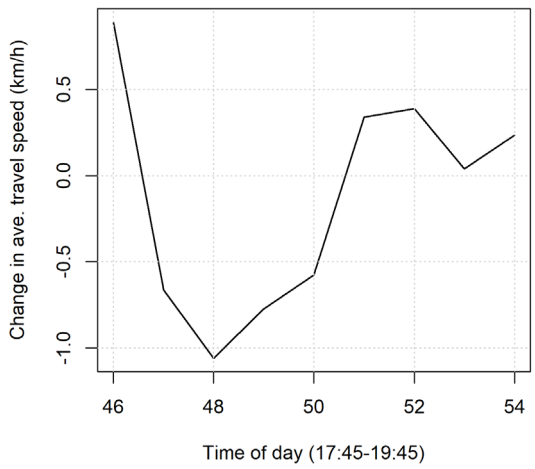
(42) Shek Kip Mei



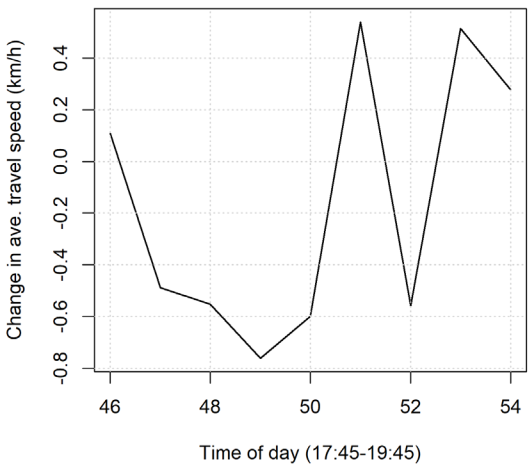
(43) Ho Man Tin



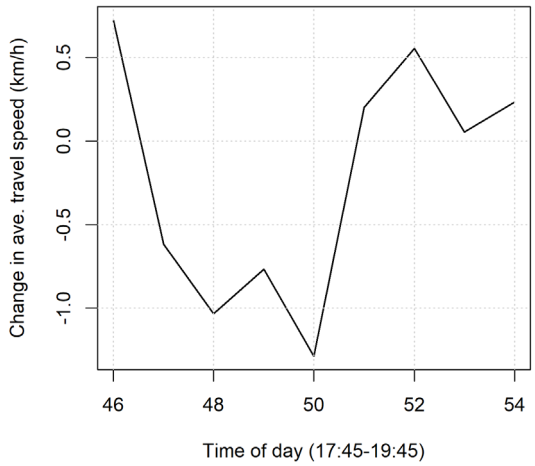
(44) Whampoa



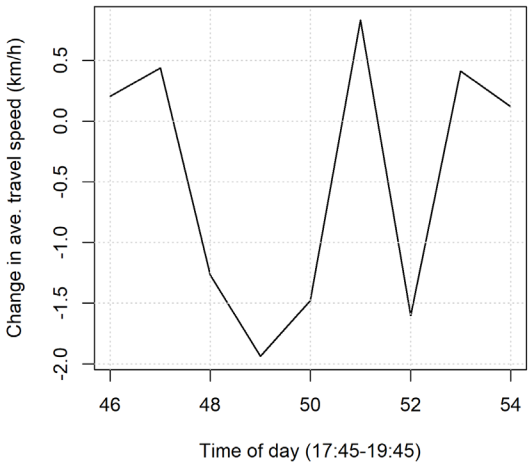
(45) Tseung Kwan O



(46) Hang Hau



(47) Po Lam



(48) LOHAS Park

Figure C.1: The spillover effects of the example disruption on other 48 stations in the MTR