THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學
Pao Yue-kong Library
包玉剛圖書館

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

Pao Yue-kong Library, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

http://www.lib.polyu.edu.hk

UNDERSTANDING AND ANALYSIS OF BICYCLE TRAVEL AND SAFETY

HONGLIANG DING

PhD

The Hong Kong Polytechnic University

2022

The Hong Kong Polytechnic University
Department of Civil and Environmental Engineering

UNDERSTANDING AND ANALYSIS OF BICYCLE TRAVEL AND SAFETY

HONGLIANG DING

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

August 2022

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

(Signed)

_____

DING Hongliang (Name of student)

Abstract

Cycling has received more and more attention in urban and transport planning in recent years. As an active transport mode, cycling does not only relieve traffic congestion and reduce vehicle emissions, but also improves the well-being of society. Despite the benefits for health and environment, bicyclists are vulnerable to injuries and mortalities in road crashes. It is crucial to identify the influencing factors that affect the bicycle crash risk. Therefore, effective countermeasures can be implemented to improve overall bicycle safety.

In this study, effects of policy interventions on bicycle travel and safety are examined, based on comprehensive traffic and crash data. For example, policy interventions including low emission zone, congestion charging scheme, and public bicycle rental scheme are considered. The propensity score matching method is applied to account for the effects of confounding factors like built environment and population socio-demographics. Results indicate that bicycle travel increases remarkably after the implementation of low emission zone, especially for short and intermediate bicycle trips. However, bicycle crash frequencies also increase after the introduction of congestion charging and public bicycle rental schemes.

On the other hand, association between built environment, population socio-demographics, road network configuration, traffic characteristics, and bicycle crash frequency at zonal level is measured, with which the bicycle crash exposure is accounted. For example, bicycle usage data from the public bicycle rental system is used to estimate the bicycle crash exposure. In addition, a weighted shortest path approach is proposed to estimate the bicycle distance travelled, with which the configuration of cycle lane network and safety perception of bicyclists are considered. Results indicate that bicycle crash frequency model that incorporates bicycle distance travelled as exposure is superior to those using bicycle time travelled and bicycle trip frequency as exposure. Furthermore, factors including land use, bicycle infrastructure, population density, gender, age, median household income, and weather condition are found to affect bicycle crash frequency, after controlling for the effects of unobserved heterogeneity and spatial correlation.

Last but not least, advanced statistical and deep learning models are developed to resolve the prevalent problems in safety analysis. For example, a multivariate Poisson-lognormal regression model is developed to account for the correlation between the frequencies of different bicycle crash types. Furthermore, imbalanced crash data and boundary crash problems are resolved using the deep learning approaches including augmented variational autoencoder and crash feature-based allocation methods. Results indicate that crash frequency models developed using the aforementioned approaches have better prediction performances. More importantly, more influencing factors can be identified.

To sum up, findings of this study can enhance the understanding on the roles of environmental, physical, social, and political factors in bicycle travel and safety. This should shed light on the optimal urban planning, engineering design, and transport policy that can promote bicycle travel and improve bicycle safety in the long run.

(464 words)

Publications arising from the thesis

Referred Journal:

1. **Ding, H**., Sze, N.N*., Li, H., Guo, Y., 2020. Roles of infrastructure and land use in bicycle crash exposure and frequency: a case study using Greater London bike sharing data. *Accident Analysis & Prevention*, 144, 105652.

2. **Ding, H**., Sze, N.N*., Li, H., Guo, Y., 2021. Effect of London cycle hire scheme on bicycle safety. *Travel Behaviour and Society*, 22,227-235.

3. **Ding, H**., Sze, N.N*., Guo, Y., Li, H., 2021. Role of exposure in bicycle safety analysis: Effect of cycle path choice. *Accident Analysis & Prevention*, 153, 106014.

4. **Ding, H**., Lu, Y., Sze, N.N*., Chen, T., Guo, Y., Lin, Q., 2022. A Deep Generative Approach for Crash Frequency Model with Heterogeneous Imbalanced Data. *Analytic Methods in Accident Research*, 34, 100212.

5. **Ding, H**., Sze, N.N*., 2022. Effects of road network characteristics on bicycle safety: A multivariate Poisson-lognormal model. *Multimodal Transportation*, 1 (2), 100020.

6. **Ding, H**., Guo, Y., Sze, N.N*., 2022. Effect of the Ultra-Low Emission Zone on the usage of the London Cycle Hire Scheme. *Transportation Letters: The International Journal of Transportation Research*, in press.

7. **Ding, H**., Lu, Y., Sze, N.N*., Li, H., 2022. Effect of dockless bike-sharing scheme on the demand for dock-based bike-sharing at disaggregate level using a deep learning approach. *Transportation Research Part A: Policy and Practice*, 166, 150-163.

8. **Ding, H**., Lu, Y., Sze, N.N*., Antoniou, C., Guo, Y., 2023. A crash feature-based allocation method for boundary crash problem in spatial analysis of bicycle crashes. *Analytic Methods in Accident Research*, 37, 100251.

Other journal publications

1. **Ding, H**., Sze, N.N*., Li, H., Guo, Y., 2021. Affected area and residual period of London Congestion Charging scheme on road safety. *Transport Policy*, 100,120-128.

2. **Ding, H**., Li, H., Sze, N.N*., 2022. Effects of the abolishment of London Western Charging Zone on traffic flow and vehicle emissions. *International Journal of Sustainable Transportation*, 6 (16), 558-569.

3. Lu, Y., **Ding, H**., Ji, S., Sze, N. N., He, Z*., 2021. Dual attentive graph neural network for metro passenger flow prediction. *Neural Computing and Applications*, 33(20), 13417-13431.

4. Li, H*., Zhang, Z., Sze, N. N., Hu, H., **Ding, H**., 2021. Safety effects of law enforcement cameras at non-signalized crosswalks: A case study in China. *Accident Analysis & Prevention*, 156, 106124.

5. Chen, T., Lu, Y., Fu, X., Sze, N. N*., **Ding, H**., 2022. A resampling approach to disaggregate analysis of bus-involved crashes using panel data with excessive zeros. *Accident Analysis & Prevention*, 164, 106496.

Conference paper

1. **Ding, H**., Sze, N.N*., Wang, S., 2019. Market Analysis of Electric Vehicle Sharing Using Market Segmentation Approach. Paper presented at the 24th International Conference of Hong Kong Society for Transportation Studies, 14-16 December, Hong Kong.

2. **Ding, H**., Sze, N.N*., Lu, Y., 2022. A deep learning approach for boundary crash effect in spatial bicycle crash analysis. Paper submitted for possible presentation at the 26th International Conference of Hong Kong Society for Transportation Studies, 12-13 December, Hong Kong.

Acknowledgements

First and foremost, I would like to express my most sincere gratitude to my chief supervisor, Dr. Nang-ngai Sze, for his unreserved guidance and supervision. The priceless time he spent with me on this research is gratefully appreciated. I have gained the once-in-a-lifetime chance of PhD study and much knowledge from him. He shows much patience and kindness throughout my postgraduate study. He not only cares about students' health and well-being but also always provides valuable suggestions and help when facing difficulties in our research or daily life. He is an excellent supervisor, a friend, a family, and a life mentor to me. I have learned a lot from his guidance, wisdom, and expertise, which will be greatly important for my future career.

I would like to express my sincere thanks to my co-supervisor, Prof. Anthony Chen, for his rigorous attitude toward academic research, encouraging me to pursue excellence, and providing professional advice and support. He is meticulous in scientific research and always cares about the student's future development.

Also, I would like to thank Prof. Yanyong Guo and Haojie Li (from the Southeast University, Nanjing, China) for their constant support and help in my academic research and personal life. Whenever I faced troubles, they were always there to enlighten me and offer help with a better solution. To me, they are not only respected researchers and supervisors but also good friends and families. It has been my great luck to be under their supervision of them. Also, I would like to thank Prof. Constantinos Antoniou (from the Technical University of Munich) for his valuable guidance and critical comments on my research work.

Next, I would like to thank my groupmates in the transportation team of the Hong Kong Polytechnic University: Dr. Junbiao Su, Dr. Tiantian Chen, Dr. Dianchen Zhu, Ms. Manman Zhu, Mr. Penlin Song, for their kind assistants and supports in the past few years. I also wish to thank my collaborators, Yuhuan Lu (currently at the University of Macau), Qinhai Lin (currently at the Sun Yat-sen University), Shunchao Wang (currently at the Southeast University), and Jingfeng Ma (currently at the Southeast University).

And of course, I would like to thank my friends in Hong Kong for our friendship, which has been unforgettable memory in my PhD study.

Last but not least, I would like to thank my family: mom (Fenglian Du), dad (Yiyou Ding), and sister (Hongli Ding). They always provide me with unwavering love and encourage me to fight for my pursuits. Thanks, my dearest parents, this journey would not have been finished without your support and helps. Thanks for your companions and loves all the time. Love you forever!

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background

Sustainable urban development has struggled with the problem of car dependency. Air pollution, climate change, traffic congestion, unsafe roads, and poor physical health are just a few of the issues it causes (Ruiz-Padillo et al., 2018, Johnson and Silveria, 2014). Therefore, it is paramount to promote alternative modes of transportation, such as public transit, cycling, and walking. Cycling has been increasingly marketed as a sustainable mode of transportation. It improves overall social well-being in addition to reducing traffic congestion and emissions associated with traffic (Li et al., 2019, Guo et al., 2018b). Throughout the world, numerous cycling-friendly transport policies have been implemented in recent years. As an example, residents of Greater London have suggested that they were inspired to start cycling by the London Cycle Hire (LCH) programme, which was launched in July 2010. (ITV, 2014). Cycle Superhighways, designed to offer cyclists safer, faster, and more direct journeys through the city, were implemented from outer London into and across central London in company with the launch of the cycle hire program. Transport for London (TfL, 2018) reported that between 2015 and 2017, the average number of daily bicycle trips in London increased by 3.9%. Specifically, approximately 25% of bicycle trips occurred in Central London. In certain areas of Central London, bicycles constituted a notably high proportion of commuter trips. In 2016, 65% of commuters at Torrington Place travelled by bicycle (55% at Tooley Street and 48% at Southwark Bridge).

Although the advantages of cycling are well-documented, bicycle safety could be a major concern. The same road infrastructure and facilities that are used by cars, buses, and trucks must frequently be shared by cyclists. Bicyclists are indeed more susceptible to severe injury and fatality on the road since they are not protected by their vehicles (Davis and Pless, 2001). According to the World Health Organization (WHO), over one-third of road traffic fatalities involve pedestrians and cyclists (WHO, 2018). As an illustration, the European Union reported that approximately 8% of road fatalities involved bicyclists, whereas in the Netherlands, this figure reached 24%. (Lajunen et al., 2016). According to

the National Highway Traffic Safety Administration (NHTSA), between 2007 and 2017, there were approximately 8,028 fatalities involving bicyclists, representing an increase of approximately 11.70 % (NHTSA, 2019). Despite the fact that road safety in the United Kingdom has improved significantly, with fatalities decreasing by 49% from 2000 to 2012, to a total of 1,637. The total number of cyclists killed or seriously injured increased by 21% during the same period, evidencing an inverse trend (Talbot et al., 2014). In addition to public health, road accidents can result in financial losses (Esiyok et al., 2005). The property damage at the scene of an accident, as well as the high costs of emergency treatment and medical care, could push a family into poverty, especially in developing countries (Hijar et al., 2003). For instance, bicycle accident victims incur 20-fold higher medical expenses than patients treated and released from the emergency department (Gaither et al., 2018).

In recent years, researchers and practitioners have focused on the long-term improvement of bicycle safety. Understanding the effects of environmental, physical, social, and political factors on bicycle travel and crashes is instrumental for developing a safer cycling environment. Nonetheless, a number of concerns and issues remain unclear. The subsequent section introduces the research gaps in the existing literature that inspired this thesis and, in turn, defines the study objectives.

1.2 Motivations

This thesis aims to evaluate bicycle travel and safety from the policy interventions, built environment, population characteristics, and modelling issues. It is of great importance to assess the effect of policy interventions, built environment, and population characteristics on bicycle travel. Since the bicycle travel demand could affect bicycle safety. It is crucial to identify the relationship between contributory factors, bicycle travel and bicycle safety. Last but not least, the crash frequency model issues should be addressed to well assess the effects of risk factors on the occurrence of bicycle crashes.

This thesis is first motivated by the causal relationship between policy intervention, bicycle travel, and bicycle safety. Numerous studies have evaluated the effects of

contributing factors on bicycle travel (McNeil et al., 2018; Garca-Palomares et al., 2012; Li et al., 2018; 2019; Santos and Shaffer, 2004). Population characteristics, the built environment, and bicycle infrastructure are among the influencing factors. Several studies have found, however, that policy interventions can also influence bicycle travel. Public bike-sharing systems (Li et al., 2019; Midgley, 2011; Fishman et al., 2014), bicycle highway infrastructures (Li et al., 2018), mass transit systems (Gu et al., 2019; Bakó et al., 2020), and congestion charging schemes (Santos and Shaffer, 2004) were among the policy interventions considered. However, the effects of traffic emission interventions were rarely studied. In fact, such interventions can also promote the transition to greener modes of transportation. Moreover, policy interventions can indirectly affect road safety by influencing other factors, such as traffic volume, which plays a crucial role in the occurrence of crashes. For instance, Elvik and Vaa (2004) stated that when traffic volume is increased by 100%, the number of crashes would increase by 80% and 20% for injury and fatal crashes, respectively. As a consequence, one of the pertinent questions that arises is whether the number of bicycle crashes will rise as a result of the implementation of policy interventions, considering more bicyclists are on the roads?

To better quantify the likelihood of bicycle crash involvement and interpret the risk posed by various entities, it is necessary to measure crash exposure. In previous studies, population or population density was frequently used as a proxy for exposure at the macroscopic scale, particularly for active transportation modes such as walking and bicycling (Cottrill and Thakuriah, 2010; Siddiqui et al., 2012; Lee et al., 2015a; Wang et al., 2017; Sze et al., 2019). Additionally, some studies have utilised total bicycle track length as a proxy for bicycle crash exposure (Wei and Lovegrove, 2013; Siddiqui et al., 2012). They do not, however, account for the variation in bicycle travel between individuals. Alternately, several studies have adopted bicycle count (Miranda-Moreno et al., 2011; Blaizot et al., 2013; Guo et al., 2018a; Nordback et al., 2013), bicycle time travelled (BTT), and bicycle distance travelled (BDT) (Mindell et al., 2012; Blizot et al., 2013; Poulos et al., 2015) as the exposure measures for bicycle crash frequency models. On the basis of comprehensive traffic count data, annual average traffic flow (AADT) and vehicle kilometre travelled (VKT) can be used to estimate the exposure (Pei et al., 2012). However, data regarding bicycle counts are rarely available. On the basis of self-

reported data, bicycle crash exposure may be measured using retrospective and prospective approaches. However, they are subject to self-selection bias. Moreover, an extensive household travel survey can be costly and time-consuming. Due to the pressing need for research to advance the estimation of exposure in the bicycle safety analysis, this subject is of considerable interest to us in this case.

The ultimate purpose of this thesis relates to advanced statistical and deep learning models for safety analysis. Three modelling issues, namely correlations between various crash types, excessive zero observations, and the boundary crash problem, will be explored.

Considering the prevalence of zonal safety analysis, numerous crash models have been devised to investigate the relationship between crash frequency and potential influencing factors. Several studies have examined the possible correlation between counts of different crash types (Lee et al., 2015b; Ma et al., 2008; Pei et al., 2016; Tunaru, 2002; Park and Lord, 2007; Yasmin and Eluru, 2016; Zhan et al., 2015; Zhao et al., 2018). However, the majority of them focused on crashes involving only motor vehicles and the correlation between crashes of varying degrees of severity. In fact, the effects of potential factors on the frequency of bicycle crashes can vary depending on the type of collision (Guo et al., 2018a; Ma et al., 2008; Park and Lord, 2007). For instance, bicycle infrastructure is more sensitive to bicycle-only crashes than accidents involving bicycle-vehicle (De Rome et al., 2014; Teschke et al., 2014; Beck et al., 2016). Therefore, bicycle crash frequency models need to incorporate multivariate correlations.

A second concern relates to the excessive zero observations in the safety analysis. Crash is a rare occurrence, as is commonly understood. This gives rise to the issue of the unbalanced crash and non-crash cases during the development of crash frequency models (Abdel-Aty et al., 2004). Prior studies indicated that imbalanced crash data could contribute to bias in parameter estimation and inadequate model fit (Miaou, 1994, Shankar et al., 1997). In addition, it can adversely impact the identification of crash explanatory factors (Pei et al., 2016, Yu et al., 2020, Cai et al., 2020, Washington et al., 2011). In this context, statistical and data-driven approaches were widely utilised (Lee and Mannering, 2002, Shankar et al., 2003, Huang et al., 2008, Chen et al., 2018,

Malyshkina and Mannering, 2010; Yang et al., 2018; Cai et al., 2020). Nevertheless, there are drawbacks associated with these approaches as well. For instance, statistical approaches frequently encounter sample size, missing data, and data inconsistency problems. For data-driven approaches, improved performance is possible. However, challenges associated with variable correlations, training stability, robustness, and adaptability should be addressed. More importantly, the abovementioned data-driven approaches cannot handle complicated structure data since all data variables are assumed to be real-valued. Consequently, the purpose of this thesis is to revisit such a prevalent problem in the safety analysis. The finding should shed light on the development of bicycle crash frequency models for researchers and practitioners.

In the conventional safety analysis, traffic and crash data are frequently aggregated at census tracts, street blocks, and traffic analysis zones, which are commonly delineated by roads and other physical entities (Lovegrove and Sayed, 2006; Quddus, 2008; Siddiqui and Abdel-Aty, 2012; Abdel-Aty et al., 2011; Dong et al., 2014, 2015). Therefore, a considerable portion of crashes would occur at or close to the boundaries of geographical units, notably when the roads are used to delineate various units. These crashes are also referred to as boundary crashes (Siddiqui and Abdel-Aty, 2012; Wang et al., 2012; Lee et al., 2014). In general, a boundary crash is assigned to a geographical unit based on spatial proximity, regardless of its correlation with the unit's environmental, traffic, or population characteristics. Therefore, the results of parameter estimation of the crash frequency model would be biased. In prior research, mathematical techniques such as half-and-half (Sun, 2009; Wei, 2010), collision density ratio (Cui et al., 2015), multiple membership multilevel modelling approach (Park et al., 2020, 2022), and iterative method (Zhai et al., 2018) were adopted. However, these approaches do not consider the individual crash characteristics (e.g., injury severity), which might be correlated with environmental, traffic, and road user characteristics of the corresponding geographical unit. To develop an effective bicycle crash frequency model for the relationship between risk factors and bicycle safety, it is necessary to comprehend the effects of boundary crashes.

1.3 Objectives

In response to the existing concerns elaborated in Section 1.2, the objective of this thesis can be given as follows:

(1) Effects of policy interventions on bicycle travel and safety

- To investigate the causal relationship between policy interventions, bicycle travel, and bicycle safety utilising designed empirical studies. Specifically, an advanced causal inference tool, the propensity score matching method (PSM), will be applied to account for the confounding effects of factors such as built environment and population socio-demographics.

(2) Effects of built environment and population characteristics on bicycle travel and safety

- To estimate the association between built environment, population characteristics, and bicycle crash frequency, with the bicycle crash exposure is accounted.

- To develop a weighted shortest path approach for modelling bicycle routing choices and estimating bicycle distance travelled (BDT).

- To investigate the role of bicycle crash exposure in bicycle safety analysis, accounting for the effects of built environment and population characteristics.

(3) Advanced statistical and deep learning methods for safety analysis

- To examine the association between bicycle crash frequency and possible explanatory factors, with which the correlation between different bicycle crash types is considered.

- To propose a deep learning approach to address the issue of excessive zero observations in the safety analysis.

- To propose a deep learning approach for allocating boundary crashes to develop effective crash frequency models.

It is expected that the findings of this thesis will benefit transport operators in decision-making regarding the management of bicyclists. In addition, it can strengthen the prevailing understanding of bicycle travel and safety and provide valuable insights into relevant countermeasures, such as bicycle infrastructures, traffic management and control, and education and enforcement strategies, which can enhance the safety culture and awareness of bicyclists. Resultantly, bicycle safety can be bolstered in the long run.

1.4 Thesis organization

The remainder of the thesis is organized as follows, and **Figure 1.1** outlines the interconnections of the chapters. First, the background and literature reviews are summarized. Then, the effect of policy intervention, built environment and population characteristics on bicycle travel and safety is investigated. Afterwards, modelling issues, including multivariate correlations, excessive zero observation and boundary crash problem, will be addressed using advanced technologies. The specific contents are as follows:

Chapter 2 reviews the literature on various aspects of bicycle travel and safety studies, including factors influencing bicycle travel, factors impacting bicycle safety, and analytic methodologies for bicycle travel and safety.

Chapter 3 introduces the cross-sectional and causal inference model applied in the following chapters.

Chapter 4 assesses the effects of policy intervention on bicycle travel. The London Ultra-Low Emission Zone (ULEZ), a form of policy intervention, is taken into consideration. In particular, the effects of the ULEZ on bike-sharing usage will be evaluated, encapsulating overall usage, usage by trip duration, and usage by trip destination.

Chapter 5 explores the effects of policy interventions on bicycle safety. Consideration is given to policy interventions, including the London congestion charging scheme (LCC) and the London cycling hiring system (LCH). Specifically, the effects of the LCC and LCH on the frequency and severity of bicycle crashes could be illustrated.

Chapter 6 focuses on the association between built environment, population characteristics and bicycle travel (i.e., bicycle crash exposure). For instance, ridership data (frequency and duration) from the London cycling hiring system (LCH) are employed to estimate the bicycle crash exposure. Moreover, separate bicycle crash frequency models would be developed for different seasons, i.e. from May to October (warm season) and November to April (cold season), factoring in the behaviour of bicyclists in different weather conditions.

Chapter 7 seeks to propose a weighted shortest path method, with which the configuration of the cycle lane network and safety perception of bicyclists are considered. Thus, bicycle routing will be modelled, and the bicycle distance travelled (BDT) for each trip can be estimated based on origin and destination data. Furthermore, it would be investigated what roles the three exposure measures, i.e. bicycle trips, bicycle time travelled (BTT), and BDT, play in the analysis of bicycle safety, with account for the effects of built environment and population characteristics.

Chapter 8 measures the relationships between bicycle crash frequency and possible risk factors, with which the correlation between bicycle-vehicle and bicycle-bicycle crashes are considered. The effects of road network characteristics, including road network connectivity and accessibility, on bicycle crash frequencies are also considered, in addition to population demographics, household characteristics, built environments, and traffic characteristics.

Chapter 9 proposes a deep learning approach, augmented variational autoencoder, to address the issue of excessive zero observations for crash frequency models by generating synthetic crash data. A conventional data synthesis technique, synthetic minority

oversampling technique-nominal continuous, is also considered for comparison with model prediction and factor interpretations, respectively, to assess the performance of the proposed approach.

Chapter 10 proposes a deep learning approach, crash feature-based allocation method, to resolve the boundary crash problem for macro-level crash frequency. Again, two conventional boundary crash allocation methods, including half-and-half and iterative assignment approaches, are considered for comparison from model prediction and factor interpretations, respectively.

Chapter 11 concludes the thesis with a summary of the findings, implications, limitations, and future research directions.

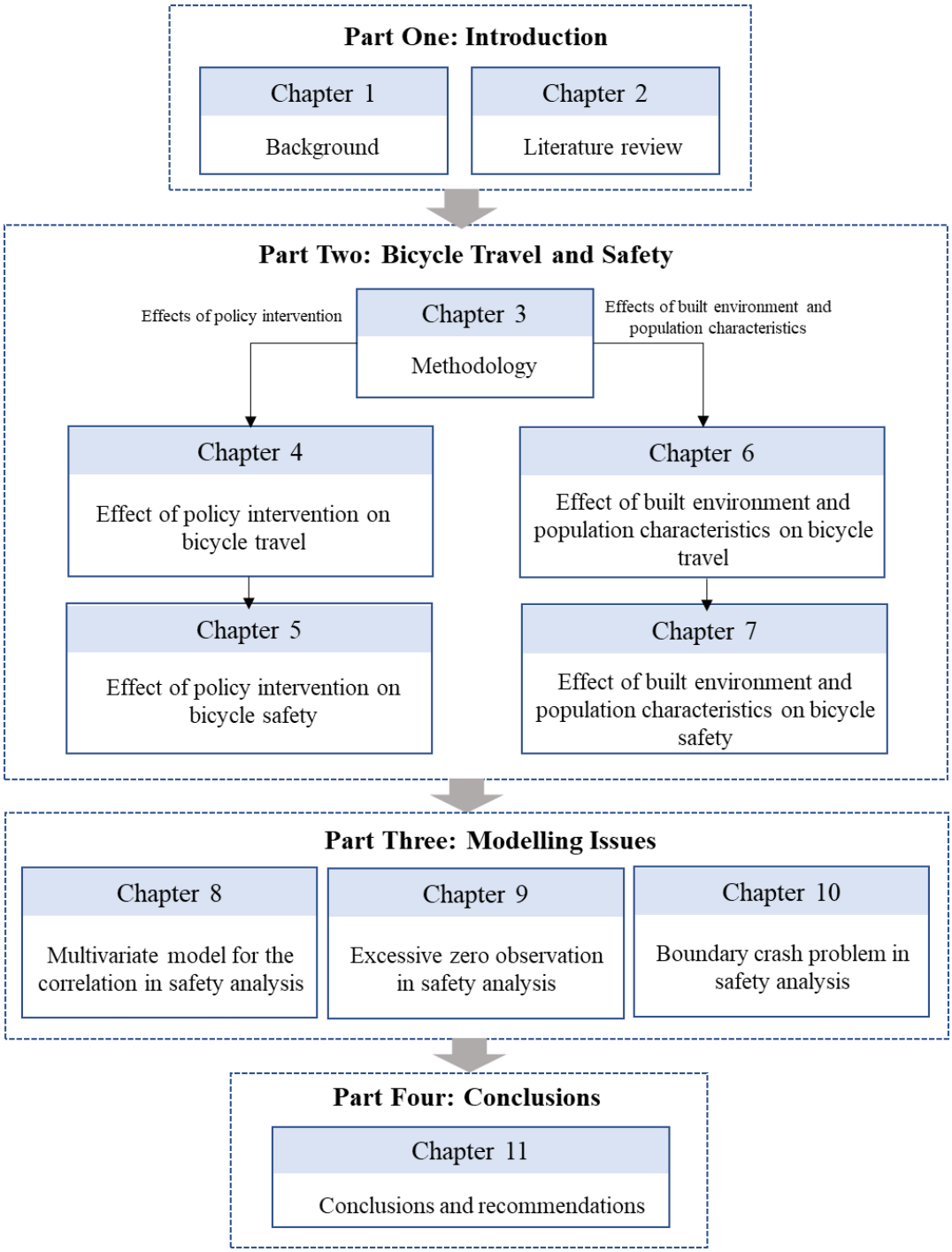


Figure 1.1 Structure of the thesis

Chapter 2 Literature review

2.1 Factors affecting bicycle travel

Bicycle travel can be characterized by population characteristics, built environment, bicycle infrastructure, and policy strategy. (Trapp et al., 2011; Campbell et al., 2016; García-Palomares et al., 2012; Gutiérrez et al., 2020; Li et al., 2018; Gu et al., 2019; Bakó et al., 2020).

Population and socio-demographics are essential factors for bicycle travel. However, effects can be varied considerably with geographical regions. For example, male bicycle usage is higher than female in the United Kingdom and United States (Heinen, et al., 2010; Trapp et al., 2011; Li et al., 2019). This could be attributed to the difference in safety perception between males and females (Handy, 2011; Heesch et al., 2012). However, there is no significant difference in bicycle usage between males and females in the Netherlands (Prati et al., 2019). In addition, there is no consistent finding on association between age and bicycle usage (Heinen et al., 2010). In general, bike usage of adolescents, children, and the elderly is less than that of their counterparts because they are more dependent (Campbell et al., 2016; Rixey, 2013; Wang et al., 2019a). In the United Kingdom and the United States, bicycle usage drops remarkably starting from middle age. In contrast, bicycle usage drops only starting from age 70 in the Netherlands (Götschi et al., 2015; Pucher et al., 2011). Furthermore, household income and education level can also affect bicycle usage. For instance, low-income populations are less willing to cycle in the United Kingdom and Australia, despite the equitable bicycle infrastructure (Heesch et al., 2012). Differently, low-income populations have been found to cycle more or no less than higher-income populations in the United States (Pucher et al., 2011).

As for the effects of the built environment, previous studies indicated that bicycle usage is associated with land use. To be specific, bike usage in industrial and commercial areas is generally higher than that in other areas (García-Palomares et al., 2012; Gutiérrez et al., 2020; Faghih-Imani et al., 2014; Zhang et al., 2017; Kim et al., 2012), even that they can be modified by time and weather conditions (Campbell et al., 2016; Gebhart and

Noland, 2014). For instance, a study by Kim et al. (2012) found that bicycle usage in commercial areas is around 15 times higher than that in residential areas in fine weather conditions. On the other hand, bicycle travel in the green area is also higher than that in other areas (Li et al., 2018, 2019; Kim et al., 2012). For instance, more than 31% of cycle hire journeys are made for leisure purposes in London (TfL, 2015). In addition, city size and terrain can also affect bicycle travel (Eren and Uz, 2020). Although many studies have indicated that the relationship between development density and bicycle demand was linear, a non-linear relationship was revealed in Australia (Boulange et al., 2017; Kerr et al., 2016).

Furthermore, road network characteristics include road density (Zhang et al., 2017; García-Palomares et al., 2012), connectivity (Cervero et al., 2009), intersection density (Ding et al., 2021a), and road geometry design (Sener et al., 2009; Chen et al., 2017; Casello and Usyukov., 2014) can all affect the bicycle travel. Also, bike usage is sensitive to the design, development and management of bicycle infrastructure. For example, the presence of bicycle lanes (Li et al., 2018; Romanillos et al., 2018), cycle superhighways (Ding et al., 2020), and configuration (i.e., a separation between bicycles and motor vehicles, speed limit, road signs and markings, etc.) (Sener et al., 2009) of bicycle network can affect the safety perception and route choice of bicyclists, and therefore the bicycle usage. Zhang et al. (2017) employed a multiple linear regression model to examine the possible influencing factors to bicycle usage in Zhongshan, China. Results indicated that bike usage is positively associated with the length of bike lanes while negatively associated with the distance to the city centre. Li et al. (2018) explored the effects of the Cycle Superhighway on the use of the London cycle hiring system (LCH). The results suggested a significant increase of 27.1% in the average ridership of the LCH within the affected area. In addition, density (Li et al., 2019; Ding et al., 2021a) and capacity (Faghih-Imani et al., 2014) of docking stations can also affect bicycle usage. Moreover, accessibilities of public transport, including buses (Li et al., 2018; Ding et al., 2021a) and rail transit (Gu et al., 2019; Bakó et al., 2020; Eren and Uz., 2021), are also positively associated with the bike usage. For instance, a higher ridership is often observed for bike docking stations located in the area with high public transit demand since cycles as a feeder mode for public transport (Jäppinen et al., 2013).

A few studies also have evaluated the effects of policy intervention on bicycle usage. For instance, Li et al. (2019) explored the effects of dockless bike-sharing systems on London cycle hire (LCH) usage. The results suggested a significant reduction in the average weekly usage of the LCH dicking station caused by the dockless bike-sharing system. Also, the operation of the mass transit system can significantly affect bicycle travel (Gu et al., 2019; Bakó et al., 2020). In addition, several studies found that congestion charging scheme which aims to reduce motor vehicles also has positive effects on bike usage (Santos and Shaffer., 2004). In recent years, traffic emission intervention, low emission zone, has been implemented worldwide. Previous studies on low emission zone mainly focused on air quality, human health, and car ownership (Wolff, 2014; Gehrsitz, 2017; Margaryan, 2021; Browne et al., 2005; Ellison et al., 2013). For instance, Gehrsitz (2017) and Margaryan (2021) found a favourable accumulative effect of low emission zone on human health. In accordance with the vehicle emission model established by Transport for London (TfL), concentrations of PM_{10} and NO_X were reduced by 2% and 4%, respectively, within the low emission zone (Kelly and Kelly, 2009). Last but not least, low emission zones can shift the transportation mode to "greener" vehicles, especially in highly developed urban cities like London (Peters et al., 2021; Ellison et al., 2013; Ding et al., 2022a). However, effects of low emission zone on bike usage are rarely investigated. Indeed, private car users may shift to cycling when travelling within/into low emission zone to avoid the high toll. **Table 2.1** Summarizes the contributory factors to bicycle travel.

Table 2.1 Some examples of contributory factors to bicycle travel

Categories	Influencing factors	Reference
Population and socio-demographics	Gender, age, household income	Heinen et al., 2010; Trapp et al., 2011; Li et al., 2019; Campbell et al., 2016; Rixey et al., 2013; Heesch et al., 2012; Pucher et al., 2010

Built environment	Land use, weather conditions, city size and terrain	García-Palomares et al., 2012; Gutiérrez et al., 2020; Faghih-Imani et al., 2014; Zhang et al., 2017; Kim et al., 2012
Road network characteristics	Road density, connectivity, intersection density, road geometry design, bicycle infrastructures, public transport	Zhang et al., 2017; García-Palomares et al., 2012; Cervero et al., 2009; Ding et al., 2021a; Sener et al., 2009; Chen et al., 2017; Casello and Usyukov., 2014; Li et al., 2018
Policy interventions	Dockless bike-sharing system, mass transit system	Li et al., 2019; Gu et al., 2019; Bakó et al., 2020

2.2 Factors affecting bicycle safety

Bicycle is a popular transport mode for short-distance trips, both commuting and leisure travel, especially for people who do not have access to a private car, e.g. adolescents, children and the elderly (Lajunen et al., 2016; Vanparijs et al., 2015). Many studies have been carried out to identify the possible risk factors for bicycle crashes. Factors considered are environmental, traffic attributes, and population and household characteristics (Siddiqui et al., 2012; Wei and Lovegrove, 2013; Chen, 2015; Pulugurtha and Thakur, 2015; Guo et al., 2018a, b).

For the environmental factors, land use, built environment and road infrastructures can affect bicycle safety. For example, a study by Chen (2015) indicated that mixed land use could increase the bicycle crash risk. In particular, the likelihood of bicycle crashes in industrial and commercial areas is higher than that of other land uses. It could be attributed to the conflicts between motor vehicles, bicycles and pedestrians (Narayanamoorthy et al., 2013). In addition, environmental factors, including landscape and weather conditions, can also affect the level of service and safety of bicyclists (Vanparijs et al., 2015; Xing et al., 2019; Zhai et al., 2019a; Fournier et al., 2017; El-Assi

et al., 2017). For the effect of traffic management and control attributes, increase in bicycle crash frequency is found to be associated with the increase in the density of intersections (Siddiqui et al., 2012; Wei and Lovegrove, 2013; Pulugurtha and Thakur, 2015), traffic signal (de Geus et al., 2012; Chen, 2015), presence of cycle lanes (Reynolds et al., 2009; Hamann and Peek-Asa, 2013; Wei and Lovegrove, 2013; Chen et al., 2016), and presence of on-street parking (Wei and Lovegrove, 2013; Vandenbulcke et al., 2014). However, the findings abovementioned are not consistent and vary across different studies. For instance, Chen et al. (2012) suggested that the presence of cycle lanes did not lead to additional bicycle crashes but a possible increase in bicycle activities.

Personal demographic, socioeconomics, household characteristics and population profiles all affect the bicycle crash frequency. In particular, bicycle crash involvement rates of adolescents, children and the elderly are higher than that of other bicyclists. Additionally, their involvement rates in single bicycle crashes are particularly high (Rodgers, 1995; Tin Tin et al., 2010; Siddiqui et al., 2012; Ghekiere et al., 2014). Lack of sufficient skills and non-compliance with relevant guidelines are correlated to the high accident rates of adolescents and children (Mandic et al., 2018; Chong et al., 2017). For older bicyclists, the elevated crash rate could be attributed to the degradation of cognitive performance and mobility (Noland and Quddus, 2004; Vanparijs et al., 2015). For the effect of gender, studies indicated that the fatality rate of the male cyclist is higher than that of the female counterpart (Rodgers, 1995; Beck et al., 2007; Mindell et al., 2012; Wei and Lovegrove, 2013; Vanparijs et al., 2015; Guo et al., 2018b). This might be because male bicyclists are generally more aggressive and have a tendency to violate traffic controls. For the socioeconomics and household characteristics, previous studies indicated that household income could affect bicycle ownership, travel behaviour and, therefore, bicycle crash involvement (Siddiqui et al., 2012; Guo et al., 2018a). **Table 2.2** Summarizes the contributory factors to bicycle safety.

Table 2.2 Some examples of contributory factors to bicycle safety

Categories	Influencing factors	Reference
Population and socio-demographics	Population, gender, age, household income	Rodgers, 1995; Tin Tin et al., 2010; Ghekiere et al., 2014; Noland and Quddus, 2004; Vanparijs et al., 2015; Siddiqui et al., 2012; Guo et al., 2018a
Built environment	Land use, weather conditions, landscape, traffic flows	Chen, 2015; Vanparijs et al., 2015; Xing et al., 2019; Zhai et al., 2019a; Fournier et al., 2017; El-Assi et al., 2017
Road infrastructure	Intersection density, traffic signal, cycle lanes, parking	Siddiqui et al., 2012; Wei and Lovegrove, 2013; Pulugurtha and Thakur, 2015; de Geus et al., 2012

For the association measure between bicycle crash frequency and possible influencing factors, it is necessary to consider the exposure to facilitate the accurate assessment and effective comparison. For example, cycling activities can vary across different built environments and road infrastructures. A study by Chipman et al. (1993) has warned that different exposure measures could lead to different estimation results. In the literature, population or population density were often used to proxy the exposure at the macroscopic level, especially for active transportation modes like pedestrian and bicycle (Cottrill and Thakuriah, 2010; Siddiqui et al., 2012; Wang et al., 2017; Sze et al., 2019). Also, some studies have applied the total bicycle track length to proxy the bicycle crash exposure (Wei and Lovegrove, 2013; Siddiqui et al., 2012). However, these studies did not account for the differences in traffic flow between different roads and cycling activities between different population groups. To get rid of this, some studies adopted bicycle trips (Miranda-Moreno et al., 2011; Guo et al., 2018b), vehicular traffic volume (Beck et al., 2007; Hamann and Peek-Asa, 2013; Wei and Lovegrove, 2013), bicycle time travelled (BTT) and bicycle distance travelled (BDT) (Mindell et al., 2012; Poulos et al., 2015) as the exposure measure in bicycle crash analysis. Unlike the vehicle crash analysis, automated bicycle counts are often unavailable to estimate bicycle exposure. To measure the bicycle exposure, a possible way is to investigate the travel behaviour (in terms of

bicycle trip, BTT and BDT) of a specific bicyclist group using the questionnaire survey (Poulos et al., 2015). However, accuracies of the survey data, especially for time and distance travelled, are subject to recall bias. Several studies have estimated bicycle exposure using actual bike counts (Guo et al., 2018; El-Esawey et al., 2015). For instance, Guo et al (2018) adopted more than 810,000 hourly volumes (covers more than 70% of Vancouver's bike network) to estimate bike safety exposure. **Table 2.3** shows the data collection methods of bicycle exposures in the previous studies.

Table 2.3 Some examples of bicycle exposure measures

Reference	Design	Methodology	Exposure
Rodgers et al., 1995	Retrospective	Telephone questionnaire	BTT
Aultman and Kaltenecker, 1999	Retrospective	Questionnaire and map	BDT
Thornley et al., 2008	Retrospective	Questionnaire	BTT
Vandenbulcke et al., 2009	Retrospective	National travel survey	BTT and BDT
Bacchieri et al., 2010	Retrospective	Face to face interview	BTT
Lusk et al., 2011	Retrospective	Automated traffic counts	BDT
Tin Tin et al., 2010	Retrospective	National travel survey	BTT
Blazizot et al., 2013	Retrospective	Regional household travel survey	Bicycle trips; BTT; BDT
Hoffman et al., 2010	Prospective	Online questionnaire	BDT
Johnson et al., 2010	Prospective	Video camera	BTT
De Geus et al., 2012	Prospective	Online questionnaire	Bicycle trips; BTT; BDT
Sayed et al., 2013	Prospective	Video camera	Bicycle trips

In contrast, bicycle trips, origin and destination data are more reliable. However, exposure measures also are limited to bicycle trips and bicycle time travelled. Therefore, it may be possible to estimate the bicycle distance travelled based on the shortest path between the origin and destination of each trip (Zacharias, 2005; Pucher and Buehler, 2006; Larsen and El-Geneidy, 2011). For the route choice decision of motor vehicle drivers, common

influencing factors are monetary cost, travel time and reliability. However, for the route choice decision of bicyclists, some other factors including road environment and level of service should also be considered (Ehrgott et al., 2012; Yang and Mesbah, 2013; Chen et al., 2017; Sener et al., 2009). For example, previous studies indicated that bicyclists tend to choose routes with fewer traffic signals and stop signs to avoid frequent stop-and-go (Heinen et al., 2010; Menghini et al., 2010; Stinson and Bhat, 2003). In addition, bicyclists tend to avoid interactions with pedestrians and motor vehicles by choosing routes with fewer crosswalks and roadside parking (Stinson and Bhat, 2003; Yang and Mesbah, 2013). Furthermore, road geometric designs, including gradient, crossfall and road surface condition, are also associated with the bicycle route choice (Sener et al., 2009; Chen et al., 2017; Casello and Usyukov, 2014).

Nevertheless, the perceived safety risk can play an important role, as much as distance and time, in the bicycle route choice (Hopkinson and Wardman, 1996; Broach et al., 2012; Ehrgott et al., 2012). Possible factors that may affect the perceived safety risk of bicyclists are vehicular traffic flow and speed (Menghini et al., 2010; González et al., 2016). For instance, bicyclists tend to ride on roads with less vehicular traffic and lower speed limits (Sener et al., 2009). In addition, the presence of bicycle infrastructures and facilities, including cycle lanes, cycle tracks, intersection crossing markings and corner refuge islands, is associated with the increase in bicycle use (Barnes et al., 2006; Sener et al., 2009; Deliali et al., 2020). **Figure 2.1** depicts the typical bicycle facilities, including (a) segregated cycle lane, (b) designated cycle lane, (c) shared bus and cycle lane, and (d) shared cycle lane and footpath. Several studies were conducted to examine the relationship between bicycle facilities and bicycle route choice (Broach et al., 2012). Results indicated that bicyclists generally prefer segregated cycle lanes to designated cycle lanes. The shared cycle lanes are the least preferred choice (Jensen, 2007; Winters and Teschke, 2010). Moreover, directness and connectivity of the bicycle infrastructures can also affect bicycle use. It is necessary to provide a direct and uninterrupted route for bicyclists to reach their desired destinations (Stinson and Bath, 2003). Last but not least, the presence of protected intersections can improve the safety perception of bicyclists since the vehicular traffic is physically separated from the bicycles (Deliali et al., 2020).

(a) Segregated cycle lane

(Source: http://www.walkbikecupertino.org/new_wbc/index.php/2019/02/20/separated-bicycle-lanes-coming-to-mcclellan-road/)

(b) Designated cycle lane

(Source: https://ourhamilton.co.nz/on-the-move/council-take-steps-to-improve-cycle-lane-safety/

(c) Shared bus and cycle lane

(Source: https://future-economics.com/2019/03/24/bike-bus-lanes-can-i-interest-you-in-a-time-share/)

(d) Shared cycle lane and footpath

(Source: https://www.brooklynpaper.com/breaking-away-city-panel-green-lights-protected-pulaski-bike-lane/)

Figure 2.1 Illustrations of typical bicycle facilities

In addition to the factors mentioned above, effects of policy interventions on bicycle safety also should be investigated. Policy interventions can affect the bicycle exposure and, in turn, contribute to the occurrence of bicycle crashes. In the literature, a considerable body of studies have been conducted to evaluate the effects of policy interventions on road safety (Hyatt et al., 2009; Quddus, 2008; Noland et al., 2008; Li et al., 2017; Jones et al., 2008; Lord et al., 2005). Results suggested that policy interventions can indirectly affect road safety by affecting traffic flows, speed and other factors. In particular, traffic volume is the most crucial factor affecting road safety in both the short- and long run. For instance, the number of crashes could be significantly increased with the traffic volume if other environmental conditions remain unchanged (Golob and Recker, 2001; Martin, 2002; Dixit et al., 2011; Lord et al., 2005). As mentioned in Chapter 2.1, policy interventions, including congestion charging and public bicycle rental schemes, can stimulate the bicycle usage. For instance, 49% of LCH users are encouraged by the scheme to start cycling in London (ITV, 2014). However, their effects on bicycle safety are rarely examined.

With regard to the congestion charging scheme, not only the favourable effects on vehicular speed, traffic flow and vehicle emission, but also the safety influences could be revealed after the introduction of congestion charging (TfL, 2005). Congestion charging can effectively relieve the traffic congestion by reducing the overall traffic volume, shortening the travel time and increasing vehicular speed. This could in turn affect road safety levels (Xie and Olszewski, 2011; Lord et al., 2005). Studies indicated that motor vehicle crashes were reduced after the congestion charging scheme was introduced in London (Green et al., 2016; Quddus, 2008; Noland et al., 2008). However, the number of bicycle casualties increased (13.3%) simultaneously (Li et al., 2012). Yet, it was not well studied whether such an increase was attributed to the increase in bicycle trips or other factors like traffic volume, vehicle mix and vehicular speed.

A few studies have attempted the safety effects of the cycle hire scheme, and their findings are controversial. For example, the presence of cycle hire scheme is found to be associated with reduced risk of bicycle injuries. The likelihoods of fatal and severe injuries of bike-share users are lower than that of other bicyclists (Fishman and Schepers,

2016; Fishman and Schepers, 2018). In contrast, road users tend to consider the bicycle unsafe in general, considering the bicycles' vulnerability, instability and invisibility. Hence, safety concern is an issue that hinders the adoption of the cycle hire scheme (Nikitas et al., 2014; Sun, 2018; Hess and Schubert, 2019). Nevertheless, rigorous analysis of bicycle crash risk associated with bike-sharing is crucial to decision-makers regarding the introduction and expansion of the cycle hire scheme.

2.3 Analytic methods for bicycle travel and safety

2.3.1 Bike demand prediction model

The spatial granularity for bike demand prediction can generally be stratified into three classes, macroscopic level (Ermagun et al., 2018; Giot and Cherrier, 2014), mesoscopic level (Zhou, 2015; Bao et al., 2017; Liu et al., 2016), and microscopic level (Faghih-Imani et al., 2014; Rixey., 2013; Li et al., 2015; Yang et al., 2016). The macroscopic and mesoscopic studies estimate average bicycle demand for the whole city and clusters of docking stations or bicyclist groups. In the microscopic studies, time-series trends of bicycle demand at individual docking stations are modelled. Although the bicycle demand prediction at the finer spatial granularity level can facilitate the fleet management of a dock-based bike-sharing system, microscopic prediction at the station level is more challenging, as compared to macroscopic and mesoscopic models, considering that the bicycle demand is highly dynamic and context-dependent. For the time horizon, bicycle demand prediction models at different time scales, including hourly (Gao and Lee, 2019; Faghih-Imani et al., 2014), daily (Wang et al., 2016), weekly (Schneider et al., 2009; Sohrabi et al., 2020), and monthly (Rixey, 2013), are established. Accurate real-time bicycle demand prediction at a smaller time scale is crucial for dynamic bicycle re-positioning and balancing (Lin et al., 2019; Feng et al., 2018).

In recent years, machine learning approaches, including decision tree (VE and Cho, 2020), support vector machine (Sathishkumar et al., 2020), Bayesian network (Froehlich et al., 2009), and neural network (Xu et al., 2018a) models, were applied to predict the bicycle demand. In addition, to account for the effects of temporal, spatial and semantic

correlations in bicycle demand, deep learning approaches, including convolutional neural networks (CNNs) (Ruffieux et al., 2017), recurrent neural networks (RNNs) (Chen et al., 2019; Chen et al., 2020a; Pan et al., 2019; Ljubenkov et al., 2020), graph convolutional networks (GCNs) (Kim et al., 2019; Ke et al., 2021), and their variants (Lin et al., 2018; Yang et al., 2020; Wang and Kim, 2018; Yang et al., 2020), were applied. Although superior predictive performance can be achieved, especially for real-time demand forecast, causal inferences of bicycle demand and its influencing factors are limited (Karlaftis and Vlahogianni, 2011; Yang et al., 2020).

In contrast, parametric and non-parametric statistical models were often adopted to predict the macroscopic and mesoscopic level bicycle demand at a greater time scale, with account for the effects of influencing factors (Corcoran et al., 2014; Rudloff and Lackner, 2014; Rixey, 2013; Kaltenbrunner et al., 2010). For example, count data models, including Poisson and negative binomial regression, were applied to model the seasonal and weather effects on the demand for dock-based bike sharing (Corcoran et al., 2014; Rudloff and Lackner, 2014). To account for the non-stationary temporal variations in the bicycle demand modelling and forecast, time-series models including autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), and their variants were adopted (Kaltenbrunner et al., 2010). However, policy interventions are complicated and impact bicycle travel indirectly. The causal relationship is sometimes too indirect to estimate. In this case, the conventional statistical approaches mentioned above may not be applicable.

2.3.2 Crash frequency model

(1) Conventional models

Assessing the effects of possible risk factors on crash frequency, a great number of crash frequency models have been adopted in the safety literature. For instance, Poisson and Negative Binomial regression approaches were commonly used (Yao and Loo, 2016; Wong et al., 2007; Turner et al., 2006). To allow for parsimonious specification, a panel mixed negative binomial model (PMNB) was proposed (Bhowmik et al., 2019). PMNB

has good model fit as conventional multivariate negative binomial model. It should be noted that the crash frequency models mentioned above are global models, variable from these models are forced to have the same effect on all units and zones (Amoh-Gyimah et al., 2016). Indeed, some of the many factors affecting the crash occurrence are not observable, or the necessary data may not be able to collect. If these unobservable factors are correlated to the observed factors, bias estimation and inference could be drawn. Thus, random parameter models were proposed to account for unobserved heterogeneities. Compared with the fix-parameter model, the random-parameter models allow the coefficients of exogenous variables to vary across the individual observations. Although the random-parameter models can achieve better goodness-of-fit, the estimation computation is often more complex due to the specific distribution of concerned variables.

(2) Multivariate models

Crash data can be divided into different categories by mode type and crash severity. For instance, motorized and non-motorized modes are two of the most common categories for crash classification by involved modes. Crash severity can be categorized into three severity levels, namely fatal crash, severe injury crash, and slight injury crash, in accordance with the degree of injury of the most seriously injured person in a crash. Previous studies have proved that the effects of possible factors on crash frequencies can also vary with collision types (Guo et al., 2018a; Ma et al., 2008; Park and Lord, 2007). Therefore, it is worth modelling the crash frequency by different collision types (Guo et al., 2018a; Pei et al., 2016). Separate modelling of univariate crash frequency lacks a comprehensive understanding of crash occurrence since the correlation between different collision types is not considered. Such correlation could result in biased parameter estimation (Ye et al., 2009; Ma et al., 2008). To this end, multivariate models, such as the multivariate Poison lognormal model, were developed to simultaneously model the crash frequencies of different types, with which a generalized correlation structure is allowed, and over-dispersion is accounted for (Park and Lord, 2007; El-Basyouny and Sayed, 2009). However, the correlation between bicycle crash frequencies of different types was rarely considered in the literature. Previous studies mainly focused on the crashes

involving motor vehicles only, and the correlation between crashes of different severity levels (Ma et al., 2008; Pei et al., 2016; Tunaru, 2002; Park and Lord, 2007; Yasmin and Eluru, 2016; Zhan et al., 2015; Zhao et al., 2018).

(3) Models for excessive zero observations

Crash frequency model is often subject to excessive zero observation because of the rare nature of crashes. To resolve the excess zero problem, one possible way is to accumulate more crash cases by aggregating the data with respect to time (e.g., months and years) and space (e.g., census tracts, traffic analysis zones and counties) (Lord and Mannering, 2010; Mannering and Bhat, 2014; Zeng et al., 2017). On the other hand, alternative statistical approaches, including zero-inflated Poisson, zero-inflated negative binomial, zero-state Markov switching count-data models, and panel data mixed logit models, could be applied to model the crash occurrence with excessive zero observations (Lee and Mannering, 2002; Shankar et al., 2003; Huang et al., 2008; Chen et al., 2018; Malyshkina and Mannering, 2010). In addition, Pei et al. (2016) attempted the problem of imbalanced crash data by reproducing sets of balanced crash and non-crash cases using a mathematical simulation approach – bootstrap resampling. Results indicated that standard errors of the crash frequency model estimated using the bootstrapping approach could be reduced. Hence, the precision of parameter estimation could be enhanced. However, conventional statistical approaches are often subject to sample size, missing data, and data inconsistency problems (Mannering, 2018; Mannering et al., 2020).

In recent years, there has been an increasing interest in applying data-driven approaches in road safety analysis, particularly when dealing with complicated data structures in crash frequency and severity models (Mannering et al., 2020). For instance, several studies have applied under-sampling and under-reporting techniques to resolve the problem of unbalanced crash data, with which the excessive zero crash cases are removed (Wang et al., 2019a; Wang et al., 2019b; Yamamoto and Shankar, 2004; Yamamoto et al., 2008). However, under-sampling technique can also result in information loss because of the elimination of non-crash cases. It would then bias the parameter estimation (Yang et al., 2018; Cai et al., 2020; Johnson and Khoshgoftaar, 2019). Alternately, it is possible

to resolve the problem of unbalanced crash data using over-sampling (increasing the number of crash cases) techniques, e.g., synthetic minority over-sampling technique and generative adversarial network method (Goodfellow et al., 2014; Yuan et al., 2019; Basso et al., 2018; Li et al., 2020). Compared with other data generators like governing equations for physics-based models, synthetic data generators are more appropriate for the crash frequency models, where the relationship between outcome, environment, traffic, and behavioural variables is complicated. It may not be straightforward for causal analysis with empirical data (Yu et al., 2020; Goodfellow et al., 2014; Cai et al., 2020).

Although advanced data mining, artificial intelligent, machine learning and neural network methods can model the training data very well, some may be subject to overfitting problems when there are too many parameters (Chawla et al., 2004; Yu et al., 2020; Schlögl et al., 2019). This can bias the parameter estimation of variables in crash occurrence and severity analyses (Mannering et al., 2020). Also, problems including unobserved heterogeneity, temporal instability and spatial dependency should be considered (Mannering, 2018). For instance, the data generation process of the synthetic minority over-sampling technique method cannot capture the correlations between explanatory variables (He and Garcia, 2009; Cai et al., 2020). Loss in model accuracy may occur in the training process of the generative adversarial network method. To this end, an alternative approach – variational autoencoder framework was proposed. This method can improve the data augmentation and compression performances and achieve stable learning accuracy in the learning process (Yang et al., 2017; Razavi et al., 2019; Walker et al., 2017). In addition, such methods can regularize the encoding distribution in the training process and ensure that the latent space is continuous for sample reconstruction (Boquet et al., 2020), and enhance the data generation performance (Islam et al., 2021). Nevertheless, it should be noted that a critical assumption of the abovementioned synthetic methods is that all variables in the data should be real-valued. They are not capable of handling categorical and nominal data.

(4) Models for the boundary crash allocation

One possible way to compensate for the boundary effect is to aggregate the boundary crashes into the neighbourhoods (Cui et al., 2015; Zhai et al., 2018; Siddiqui and Abdel-Aty, 2012). Before that, one critical issue is the identification of possible boundary crashes. The manual inspection is time-consuming and requires massive human resources (Cui et al., 2015). A better solution is to recognize the boundary crashes by constructing a buffer zone along the regional boundary. Crashes in the buffer zones are known as "boundary crashes" and "interior crashes" otherwise. In previous studies, width of the buffer zone ranged from 200 feet to 350 feet, depending on the configuration and scale of geographical units (Ivan et al., 2006; Siddiqui and Abdel-Aty, 2012; Zhai et al., 2018). To this end, mathematical approaches like the curve slope method (Siddiqui and Abdel-Aty, 2012) and entropy-based histogram threshold method (Cui et al., 2015) were adopted to estimate the optimal buffer zone width.

Then, it is crucial to assign the boundary crashes correctly to respective geographical units (Siddiqui and Abdel-Aty, 2012; Wang et al., 2012; Lee et al., 2014). For example, the half-and-half approach was adopted for boundary crash allocation (Sun, 2009; Wei, 2010). Boundary crashes were evenly allocated to all neighbouring geographical units, regardless of the ratios of crash and exposure of the units. Indeed, the spatial distribution of crashes around boundaries should be differential, and the neighbouring zone also hardly has an equal effect on the boundary crashes. To improve the prediction performance, ratios of the metrics, including road length, vehicle kilometre, and crash density in neighbouring geographical units, were adopted for boundary crash allocation (Wei, 2010; Cui et al., 2015). However, such methods fail to account for the crash mechanism and potential risk factors fully. The occurrence of crashes is quite complicated that associated with many influencing factors, including the population and household characteristics, land use, built environment, and traffic attributes (Siddiqui and Abdel-Aty, 2012; Wei and Lovegrove, 2013; Abdel-aty et al., 2011; Dong et al., 2014, 2015; Mannering, 2018; Mannering et al., 2016). It might be reasonable to allocate the boundary crashes based on the crash predisposing agents. Therefore, an iterative approach was proposed to allocate the boundary crashes based on crash predisposing agents, i.e., expected crashes (Zhai et al., 2018). Furthermore, a weighting factor, which is correlated with macro-level environmental, traffic, and road user characteristics of each

geographical unit, can be adopted for boundary crash allocation (Wang and Huang, 2016; Lee et al., 2017; Wang et al., 2017; Cai et al., 2018). Last but not least, the multiple membership multilevel modelling approach can be adopted by simultaneously correlating the weights with the characteristics of multiple geographical units (Park et al., 2020, 2022). Despite that the methods mentioned above can address the boundary crashes very well, without taking into account the features of individual crashes, all of them failed to achieve the individual level crash assignment.

2.4 Concluding remarks

This chapter demonstrates the results of the literature survey on bicycle travel and safety studies. There are several research gaps identified in the literature, which are listed as follows:

(1) Previous studies have revealed the effects of population characteristics, built environment, bicycle infrastructure, and policy intervention on bicycle travel. As for the policy interventions, results indicated that interventions like bicycle hiring schemes, congestion charging zone, mass transit systems, and cycle superhighways could significantly affect bicycle usage. However, the causal link between policy interventions and bicycle safety was rarely investigated. In addition, various engineering measures and policy strategies have been initiated to mitigate the hazardous effects of vehicle emissions. These interventions also are expected to influence bicycle travel. To the best of our knowledge, its effects on bicycle usage have not yet been revealed.

(2) To better quantify the potential of bicycle crash involvement and interpret the risk of different entities, it is necessary to measure the crash exposure. In previous studies, bicycle exposures adopted were bicycle flow counts, bicycle trips, bicycle time travelled (BTT), and bicycle distance travelled (BDT), which were measured using retrospective and prospective surveys. Regardless of the sampling framework and survey design, data may be subject to recall and selection biases. In addition, an extensive household travel survey can be expensive and time-consuming. Some recent studies proposed to estimate bicycle exposure using actual bike counts. Therefore, a possible approach for estimating

bicycle exposure is proposed in this thesis using a detailed transaction record of a public bicycle rental system. Although this system covered most bicycle trips, exposure measures are limited to bicycle trips and BTT. Therefore, bicycle routing choice should be considered, and the BDT can be estimated based on the origin and destination data of each trip.

(3) Prior studies have identified the environment, traffic and road user factors that affect the risk of bicycle-related crashes. However, it is rare that difference in their effects on the risk amongst different bicycle crash types is investigated. Indeed, bicycle crashes could present different collision types due to the difference in the built environments and traffic features. Furthermore, there is a possible correlation between counts of different crash types. Therefore, the shared unobserved factors across collision types should be considered when modelling.

(4) Although advanced statistical methods can be applied to model the zero-inflated crash data, they are not plausible to resolve the imbalanced data problem. Alternatively, machine learning approaches can be applied to balance the crash dataset, including synthetic minority over-sampling technique, generative adversarial network, bootstrap resampling, and random under-sampling methods. Nevertheless, these data synthesis approaches also have deficiencies, including correlations between variables (synthetic minority over-sampling technique method), training stability (generative adversarial network), robustness (random under-sampling) and flexibility (bootstrap resampling). In light of these weaknesses in previous research in this area, this thesis seeks to contribute to the research literature by developing an advanced synthetic approach to the problem of unbalanced crash data in the safety analysis.

(5) In preceding studies, mathematical approaches like half-and-half, collision density ratio, iterative method, and multiple membership multilevel modelling approach were adopted to compensate for the boundary crash problem. However, these approaches did not consider the individual crash characteristics, which should correlate with the corresponding geographical unit's environmental, traffic, and road user characteristics. Association between crash frequency and influencing factors could be modified by

covariates like injury severity and collision mode. Hence, it is necessary to account for the crash characteristics when allocating the boundary crashes.

Chapter 3 Methodology

This chapter introduces formulations of the cross-sectional models and causal inference model applied in this thesis. Also, various assessment criteria for model performance are presented.

3.1 Cross-sectional model

3.1.1 Poisson regression model

Poisson regression method is often applied to model the crash frequency because of the random and non-negative nature of crash data. The mean and variance of Poisson distribution are assumed to be equal (Yao and Loo, 2016; Wong et al., 2007; Turner et al., 2006). Probability of having y crash in the i^{th} unit and period t can be written as,

$$P(\mathrm{y}_{it}|\mu_{it}) = \frac{\exp(-\mu_{it})(\mu_{it})^{y_{it}}}{\mathrm{y}_{it}!} \qquad i, t = 0,1,2 \cdots n \tag{3.1}$$

Where $E(\mathrm{y}_{it}) = \mu_{it}$ be the expected number of crashes.

Also, a generalized linear model with a Poisson distribution is given as,

$$\ln(\mu_{it}) = \beta_0 + x_{it}^T \cdot \beta \tag{3.2}$$

Where x_{it} is the column vector of exogenous variables corresponding to the i^{th} unit at period t, β is a vector of parameters, and β_0 is the constant term.

3.1.2 Negative binomial regression model

Various factors, including data clustering and misspecification of the model, can lead to over-dispersion (i.e., the variance is greater than the mean). Over-dispersion is primarily

due to the nature of crash data, which are subject to Bernoulli trails. When over-dispersion exists, the negative binomial regression approach (NB) should be used (Mannering et al., 2016; Wong et al., 2007; Lord and Mannering, 2010).

The negative binomial regression model is also known as the Poisson-gamma model, which can be derived by incorporating an error term that follows the gamma distribution into the probability density function. The functional form of the negative binomial regression model is given by,

$$\ln(\mu_{it}) = \beta_0 + {X_{it}}^T \cdot \beta + \varepsilon_i \tag{3.3}$$

Where ε_i is the Gamma distributed error with mean 1 and variance α. Thus, the probability density function of the negative binomial regression model can be given by the following equation,

$$P(y_{it}) = \frac{\Gamma(y_{it}+\alpha^{-1})}{y_{it}\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_{it}}\right)^{\alpha^{-1}}\left(\frac{\mu_{it}}{\alpha^{-1}+\mu_{it}}\right)^{y_{it}} \tag{3.4}$$

Where the $\Gamma(\cdot)$ is subjected to the Gamma distribution and the α is the over-dispersion parameter.

3.1.3 Correlated random parameter model

Since not all possible influencing factors are available, the unobserved heterogeneity could bias the parameter estimations and model fit. Previous studies have proved that the conventional negative binomial regression and Poisson regression models failed to deal with the unobserved heterogeneity (Mannering and Bhat, 2014). To account for the effect of unobserved heterogeneity, random parameter models were developed for crash prediction models (Mannering et al., 2016). Unlike the fixed-parameter model, random parameter models allow factors to varying across the study entities (Train, 2009). Thus, parameter β of Equation (3.3) is assumed to be randomly distributed with,

$$\beta_i = \beta + \varphi_i \tag{3.5}$$

Where φ_i is normally distributed with mean of 1 and variance of σ^2.

Therefore, the probability of crash occurrence in Equation (3.4) is calculated as:

$$P(\mathrm{y}_{it}) = \int P(\mathrm{y}_{it}|\beta)\, f(\beta) d\beta \tag{3.6}$$

It should be noted that the above formulations assume that the random parameters are independent of each other. However, there may be possible correlations between random parameters. Therefore, correlated parameter approach was proposed (Caliendo et al., 2019; Venkataraman et al., 2011; Venkataraman et al., 2013; Saeed et al., 2019; Meng et al., 2021). To be specific, the random parameter is assumed to follow a multivariate normal distribution given by,

$$\beta_i = b + V w_i \tag{3.7}$$

$$b = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_I \end{bmatrix} \tag{3.8}$$

$$V = \begin{bmatrix} (\sigma_1)^2 & & \\ \vdots & \ddots & \\ \sigma_{j,1} & \cdots & (\sigma_j)^2 \end{bmatrix} \tag{3.9}$$

Where b denotes the mean vector, V denotes the variance-covariance matrix, j is the number of random parameters, and w_i is the randomly and independently distributed uncorrelated vector.

3.1.4 Multivariate Poisson-lognormal model

To better understand the calculation of the multivariate Poisson-lognormal model, the conventional univariate Poisson-lognormal model is first introduced as follows,

Let Y_i denote the number of crashes at entity i ($i = 1,2,\cdots,I$), where Y_i follows a Poisson distribution with parameter θ_i,

$$Y_i \sim Poisson(\theta_i) \tag{3.10}$$

To account for over-dispersion, an error term ε_i that follows normal distribution is added to the regression equation as,

$$\ln(\theta_i) = \beta_0 + \sum_{j=1}^{J} \beta_j X_{ji} + \varepsilon_i \tag{3.11}$$

$$\varepsilon_i \sim Normal\ (0, \sigma_\varepsilon) \tag{3.12}$$

Where X_{ji} is the value of j^{th} explanatory variable for entity i; β_0 is the model intercept, β_j are the coefficient for the explanatory variables; σ_ε denotes the extra Poisson Variance.

For the multivariate Poisson-lognormal model, crash count can be stratified into K classes. Let Y_i^k denote the number of crashes at entity i ($i = 1,2,\cdots,N$) of crash type k ($k = 1,2,\cdots,K$). Then, equation (3.10) can be modified as,

$$\mathrm{y}_i^k \sim Poisson(\theta_i^k) \tag{3.13}$$

Also, probability of y_i^k is given by,

$$Pr\{y_i^k | \theta_i^k\} = e - \theta_i^k \frac{\theta_i^{k y_i^k}}{y_i^k!} \tag{3.14}$$

$$\theta_i^k = \mu_i^k e^{\varepsilon_i^k} \tag{3.15}$$

$$\ln(\theta_i^k) = \beta_0^k + \sum_{j=1}^{J} \beta_j^k X_{ji} + \varepsilon_i^k, j = 1,2,3,\cdots,J \qquad (3.16)$$

Where ε_i^k denotes the normally distributed multivariate error with $\varepsilon_i^k \sim N_k(0, \sum)$,

$$\varepsilon_i = \begin{pmatrix} \varepsilon_i^1 \\ \vdots \\ \varepsilon_i^k \end{pmatrix} \qquad (3.17)$$

$$\sum = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{11} \\ \vdots & \vdots & \vdots \\ \sigma_{K1} & \cdots & \sigma_{KK} \end{pmatrix} \qquad (3.18)$$

Where the diagonal parameter σ_{kk} is the variance of ε_i^k, and the off-diagonal parameter $\sigma_{k,l}(k \neq l)$ represents the covariance of ε_i^k and ε_i^l respectively.

In this thesis, the proposed multivariate Poisson-lognormal regression model will be solved using the full Bayesian inference. For the full Bayesian estimates, priors for the model parameters, including coefficients and covariance matrix for the error term, should be specified. To this end, prior for the parameter would be given by Norm (0, 10^4) and that for the error term would be given by a Wishart prior, respectively (Guo et al., 2019). Then, the Markov Chain Monte Carlo (MCMC) simulation would be applied to estimate the posterior distribution of parameters using the WinBUGS software. To assess the model convergence, three common approaches: (1) two separate chains with different initial values; (2) Brooks-Gelamn-Rubin (BGR) value being less than 1.2 (Brooks and Gelman, 1998); and (3) MCMC trace plots of the parameters, would be adopted. Specifically, the first 40,000 iterations would be excluded in the burn-in period, and the subsequent 60,000 iterations would be used for parameter estimation.

3.1.5 Assessment of the model performance

To make comparisons of the performances between different models, two indicators, Akaike information criterion (AIC) and Bayesian information criterion (BIC), were commonly applied for evaluating the goodness-of-fit of the random parameter models (Washington et al., 2011; Hilbe, 2011). AIC and BIC can be written as,

$$AIC = -2\ln(L) + 2k \tag{3.19}$$

$$BIC = \ln(n)k - 2\ln(L) \tag{3.20}$$

Where L is the maximized value of the likelihood function; n is the number of observations and k is the number of parameters considered respectively.

The deviance of information criterion (DIC) is introduced for model assessment and comparison for Bayesian inference (Spiegelhalter et al., 2014), which is calculated as,

$$DIC = 2\overline{D(\theta)} - D\overline{(\theta)} \tag{3.21}$$

Where $\overline{D(\theta)}$ is the mean of the posterior deviance $\mathrm{D}(\theta)$, $D\overline{(\theta)}$ is the deviance at the mean of posterior parameters. The deviance $\mathrm{D}(\theta)$ of the model at the values of the parameter θ is calculated by,

$$\mathrm{D}(\theta) = -2\log(\mathrm{P}(\hat{y}|\theta)) \tag{3.22}$$

3.2 Causal inference model

For the policy intervention evaluation, it is crucial to estimate the outcome of the same entity if the "policy intervention" had not been implemented. For randomized control trials like clinical experiments, direct comparison between treated and perfect "control" units may be plausible. Sufficient control over all possible confounding factors can be achieved. However, it may not be the case for empirical studies (Wood et al., 2015; Li et al., 2018). To this end, an empirical-based treatment-control effectiveness evaluation

method – propensity score matching method (PSM) – can be adopted. In the PSM framework, a control group similar to the treated group, considering a set of observed covariates that affect the conditional probability of receiving policy intervention, would be identified. Thus, similarity between treated and control groups can be addressed, while the bias attributed to all observed confounding factors can be eliminated (Li et al., 2019). PSM approach is more efficient than the conventional treatment-control matching method since a single performance metric – propensity score – can be used to proxy the effects of all observed covariates that affect the conditional probability of "policy intervention". Therefore, the difference in the outcomes between treated and control groups is mainly attributed to the policy intervention.

Let $y_i(D_i)$ denotes the outcome (e.g., bicycle usage) of unit i, where $i = 1, \cdots N$ and N is the total number of analysis units. D_i is treatment indicator, with D_i = 1 if unit i receives the "policy intervention" and D_i = 0 otherwise.

Then, the effect of policy intervention on unit *i* can be given by,

$$\delta_i = y_i\,(1) - y_i\,(0) \tag{3.23}$$

Hence, parameter of interest is average treatment effect (ATT) of all treated units. It can be given by,

$$\delta_{ATT} = E(\delta|D = 1) = E(Y(1)|D = 1) - E(Y(0)|D = 1) \tag{3.24}$$

To guarantee the validity and reliability of PSM, three assumptions given as follows must hold true (Rosenbaum and Rubin, 1983):

- Assumption 1: Stable Unit Treatment Value Assumption (SUTVA)

SUTVA requires that the policy intervention does not have any effect on the units other than the treated units.

- Assumption 2: Conditional Independence Assumption (CIA)

CIA assumes that probability of the outcome is independent of the policy intervention, and all observed factors are controlled for. It can be described as,

$$(Y(1), Y(0)) \perp T|X \tag{3.25}$$

- Assumption 3: Common Support Condition (CSC)

CSC is also known as overlap assumption. It ascertains that there is a sufficient overlap for the characteristics of treated and control units for matching. It can be given by,

$$0 < P(T = 1|X) < 1 \tag{3.26}$$

To implement PSM approach, propensity score of every unit is first calculated using the conventional discrete outcome approaches including logit and Probit models (Smith, 1997; Guo et al., 2018a). An early study indicated that there was no significant difference in the estimation results between the two models (Smith, 1997). In this study, logit model is adopted to calculate the propensity score and is specified as follow,

$$P(T = 1|X) = \frac{EXP(\alpha+\beta' X)}{1+EXP(\alpha+\beta' X)} \tag{3.27}$$

Where α is the intercept and β' is the vector of parameters for covariates *X*.

After estimating the propensity score, a control group is constructed for each treated unit. In conventional studies, multiple matching algorithms are considered for assessment purpose. In this thesis, four common matching algorithms: (1) K-nearest neighbours matching; (2) caliper and radius matching; (3) kernel and local linear matching; and (4) stratification and interval matching, are adopted for the construction of control groups (Heinrich et al., 2010).

Finally, effect of policy intervention can be estimated by comparing the difference in the outcomes between treated group and corresponding control group. In this thesis, the effect will be estimated using the software package Psmatch2 of STATA (Leuven and Sianesi, 2003).

Chapter 4 Effect of policy intervention on bicycle travel

4.1 Introduction

Previous studies have evaluated the effects of policy intervention on bicycle usage. Policy interventions include bike-sharing system, mass transit system, cycle highway infrastructure, and congestion charging schemes were considered. However, effects of traffic emission interventions on bicycle usage are rarely investigated.

Traffic emission, as one of the major contributors of greenhouse gas (GHG) emission, has been an alarming problem in sustainable urban development. Traffic emissions can pose severe chronic health issues and thus increase the risks of morbidity and mortality of drivers, commuters, and people living near the roadways (Zhong and Bushell, 2017). Various engineering measures and policy strategies have been initiated to mitigate the hazardous effects of vehicle emissions. For example, Ultra-Low Emission Zone (ULEZ) was introduced in London to encourage commuters to switch to green transportation modes, including public transportation, cycling, and walking. In accordance with a report published by the Greater London Authority (GLA) (2019), CO_2, NO_2 and NO_x emissions were reduced by 6%, 37%, and 35%, respectively, after the introduction of ULEZ.

ULEZ was introduced in London in April 2019. The boundary of ULEZ is the same as that of the central London Congestion Charging (LCC) scheme. The total area of ULEZ is 21 km^2 (equivalent to 1.3% of the total area of Greater London). There are 25 Middle Super Output Areas (MSOAs) and about 260,000 residents (0.44% of the total population of the Greater London) in ULEZ. MSOA is the basic geographical unit established for population census in the United Kingdom. The average population of an MSOA is around 10,392. **Figure 4.1** illustrates the boundary of ULEZ. Driver or owner of a motor vehicle that does not meet the relevant emission standards would be charged when entering ULEZ (TfL, 2019). **Table 4.1** summarizes the emission limits and charges of ULEZ. The charge is £12.50 per day for light vehicles including private cars, motorcycles, and light goods vehicles (with gross vehicle weight within 3,500 kg), and £100 per day for heavy vehicles

including medium and heavy goods vehicles (with gross vehicle weight more than 3,500 kg), and buses (with gross vehicle weight more than 5,000 kg) respectively.

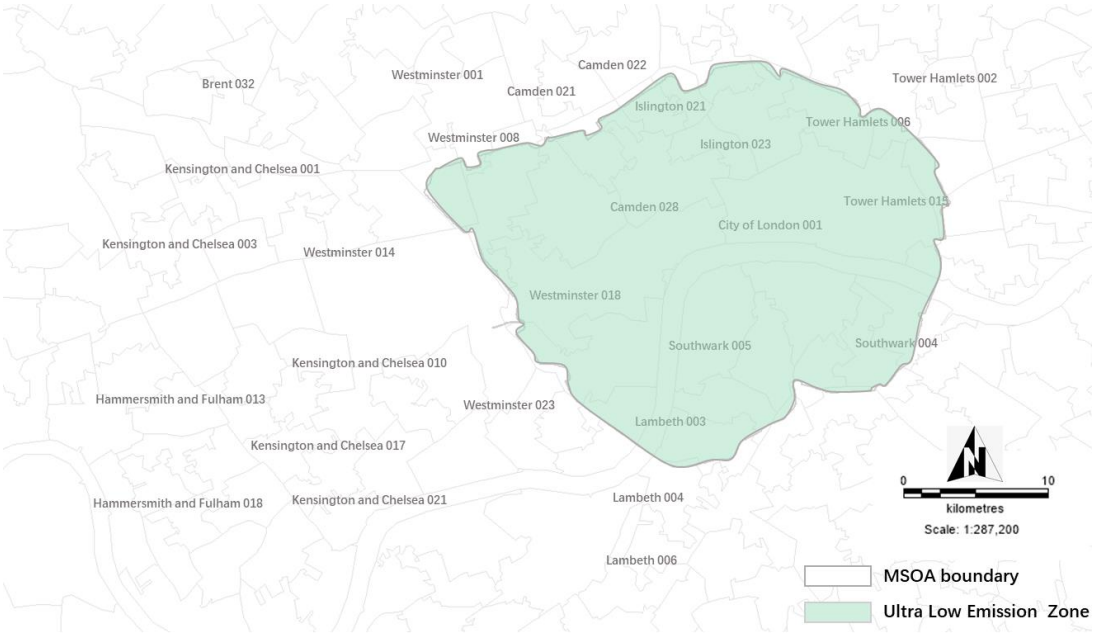


Figure 4.2 Illustration of Ultra-Low Emission Zone

Table 4.1 Emission limits and fees of ULEZ

Vehicle Class	Emission Limit		Fee/per day
Private cars	Petrol: Euro 4	CO: 1.0g/km THC: 0.10g/km NOx: 0.08g/km	£12.50
	Diesel: Euro 6	CO: 0.50g/km HC + NOx: 0.17g/km NOx: 0.08g/km PM: 0.005g/km PN: 6.0*10^11#/km	
Light goods vehicles (up to 3,500 kg)	Petrol: Euro 4	CO: 2.27g/km THC: 0.16g/km NOx: 0.11g/km	£12.50
	Diesel: Euro 6	CO: 0.740g/km HC + NOx: 0.215g/km	

		NOx: 0.125g/km PM: 0.0045g/km PN: 6.0*10^11#/km	
Motorcycles and mopeds	Euro 3 for NOx	NOx: 0.15g/km	£12.50
Medium and heavy goods vehicles (over 3,500 kg)	Euro 6 for NOx and PM	NOx: 1.2g/kWh PM: 0.01g/kWh	£100
Buses (over 5,000 kg)	Euro 6 for NOx and PM	WHSC Note 1 NOx: 0.4g/kWh PM: 0.01g/kWh WHTC Note 2 NOx: 0.46g/kWh PM: 0.01g/kWh	£100

Note 1: WHSC refers to World Harmonized Stationary Cycle

Note 2: WHTC refers to World Harmonized Transient Cycle

In addition to the mitigation of vehicle emissions, ULEZ can also relieve the traffic congestion in Central London (GLA, 2019). For instance, overall traffic entering Central London was reduced by 3% to 9% from May to September 2019, compared with the same period in 2018. Cycling, as a green transportation mode, not only helps to relieve traffic congestion and reduce vehicle emissions but also improves the overall social well-being (Li et al., 2019; Guo et al., 2018b; Huang et al., 2020). As such, it is hypothesized that private car users may shift to other transportation modes like public transportation and cycling to avoid the high charges of ULEZ. Therefore, it is possible that the bicycle demand would increase. According to a recent London survey, 65% of respondents indicated that they would switch to other transportation modes because of ULEZ. In addition, 17% of mode shifts were cycling (Green Car Congress, 2020). This study contributes to the literature by evaluating the effect of the ULEZ on the usage of public bike sharing in London. To account for the possible confounding factors, the PSM approach is applied to establish the appropriate control group for each treatment unit (i.e., docking station within ULEZ).

The remainder of this chapter is structured as follows. The study design is described in Section 4.2. The estimation results of PSM are presented and discussed in Section 4.3 and Section 4.4. Section 4.5 concludes the study with a summary of findings and future research directions.

4.2 Study design

4.2.1 Covariates affecting bike sharing usage

Validity of the PSM model depends on the unconfoundedness assumption. Despite that the unconfoundedness assumption may not be assessed directly, the effects of confounding factors can be eliminated using the relevant covariates that affect the conditional probability of receiving treatment, i.e., ULEZ. Choice of covariates, which is often data-driven, may affect the reliability and accuracy of effectiveness evaluation. Hence, a rule of thumb is to include all observed covariates, regardless of their significance to the "treatment", that may affect the outcome. In contrast, one should note that the precision of estimation can be reduced when a covariate that is not relevant to the outcome is included (Brookhart et al., 2006). To this end, a stepwise approach can be adopted to select covariates.

In the proposed PSM model, the observation unit is bicycle docking station of the London cycle hiring system (LCH). LCH scheme was launched in central London in July 2010. There were 5,000 bicycles and 315 docking stations. By 2018, the number of bicycles and docking stations increased to 11,500 and 750, respectively (TfL, 2018). Locations of bicycle docking stations are shown in **Figure 4.2**. Outcome variable is bike sharing usage (i.e., number of borrow transactions per station in the study period). Information on bike sharing usage is obtained from Transport for London (TfL). Data includes borrow (origin) and return (destination) stations, start time, end time, and loan period are available. Study period is May 2019 to October 2019.

Since the inclusion criteria of ULEZ is not known, covariates are selected based on preceding studies on the association between public bike sharing demand and influencing

factors. For example, population density, socio-demographics (i.e., gender, age), and household characteristics can affect bicycle demand. Therefore, they should be considered in the matching process (Trapp et al., 2011; Li et al., 2019). Such information can be obtained from the Office for National Statics (ONS) database of the United Kingdom.

In addition, built environment, land use, and transport facilities can also affect bicycle demand (Trapp et al., 2011; Campbell et al., 2016; García-Palomares et al., 2012). Information on the proportion of different land use types, i.e., residential area, commercial and office area, industrial area, green area, and road area, can be obtained from the Greater London Authority (GLA) database. On the other hand, information on road network characteristics (e.g., road density, road type, and intersection density), traffic flow (annual average daily traffic (AADT)), transport facilities (e.g., bus stop, railway station, and Cycle Superhighway) can be obtained from Department for Transport (DfT) database. In the United Kingdom, urban roads are categorized into three classes: Class A roads, Class B roads, and minor roads. Class A roads refer to major arterials, Class B roads refer to minor arterials and collector roads, and minor roads refer to local streets, respectively.

The aforementioned population socio-demographic, land use, and transport facilities data are aggregated at MSOA level, where a bicycle docking station is located, using geographical information system (GIS) technique. For the bike sharing usage data, they represent that of bicycle docking station. **Table 4.2** summarizes the covariates considered in the proposed PSM model.

Table 4.2 Summary statistics of the sample

Factor	Attribute	Mean	S.D.	Min.	Max.
Number of observations = 699 (bicycle docking stations)					
Bike sharing usage	Number of transactions	7,118	5,153	492	48,533
Population density	Population per km^2	59.53	57.38	3.05	320.27
Gender	Proportion of male	0.52	0.03	0.46	0.58

Factor	Attribute	Mean	S.D.	Min.	Max.
	Proportion of female	0.48	0.03	0.39	0.55
Age	Proportion of age above 64	0.11	0.04	0.03	0.24
	Proportion of age below 16	0.28	0.06	0.15	0.47
Income	Annual average household income (€)	56,980	8748.8	38,500	75,500
Land use	Proportion of residential area	0.15	0.07	0.04	0.36
	Proportion of business and office area	0.24	0.14	0.01	0.50
	Proportion of green area	0.29	0.16	0.05	0.74
	Proportion of road, railway and footpath area	0.32	0.08	0.15	0.77
Road density	Class A road (km per km^2)	3.75	1.84	0.07	9.89
	Class B road (km per km^2)	0.77	0.85	0	5.13
	Minor road (km per km^2)	1.07	1.17	0	4.46
Intersection density	Intersection per km^2	0.39	0.35	0.02	2.54
Traffic flow	Annual average daily traffic	22,894	5,340	14,920	31,106
Cycle superhighway	Length of cycle superhighway (km)	0.37	0.52	0	1.87
Density of bus stop	Bus stop per km^2	0.03	0.03	0	0.14
Density of railway station	Railway station per km^2	0.01	0.03	0	0.20

4.2.2 Treatment and control groups

PSM method is recognized as a "data-hungry" approach. For each treated unit, a considerable number of control units are required to establish the matched control group. Therefore, sufficient overlap between treatment and control groups can be ascertained. In this study, a total of 699 bicycle docking stations are considered. Treated units (210

docking stations) refer to those within the ULEZ, and control units (489) refer to those outside the ULEZ. **Figure 4.2** illustrates the spatial distributions of treated and control units. It is worth noting that the average bike sharing usage of treated units (7,949.2) was remarkably higher than that of control units (6,727.8) before implementing the PSM approach. This could be attributed to the possible confounding factors like network connectivity since ULEZ is located in central London. Hence, more frequent cycling activities are expected (Quintero et al., 2013). Nevertheless, effects of confounding factors could be accommodated, at least partially, using the proposed PSM method.

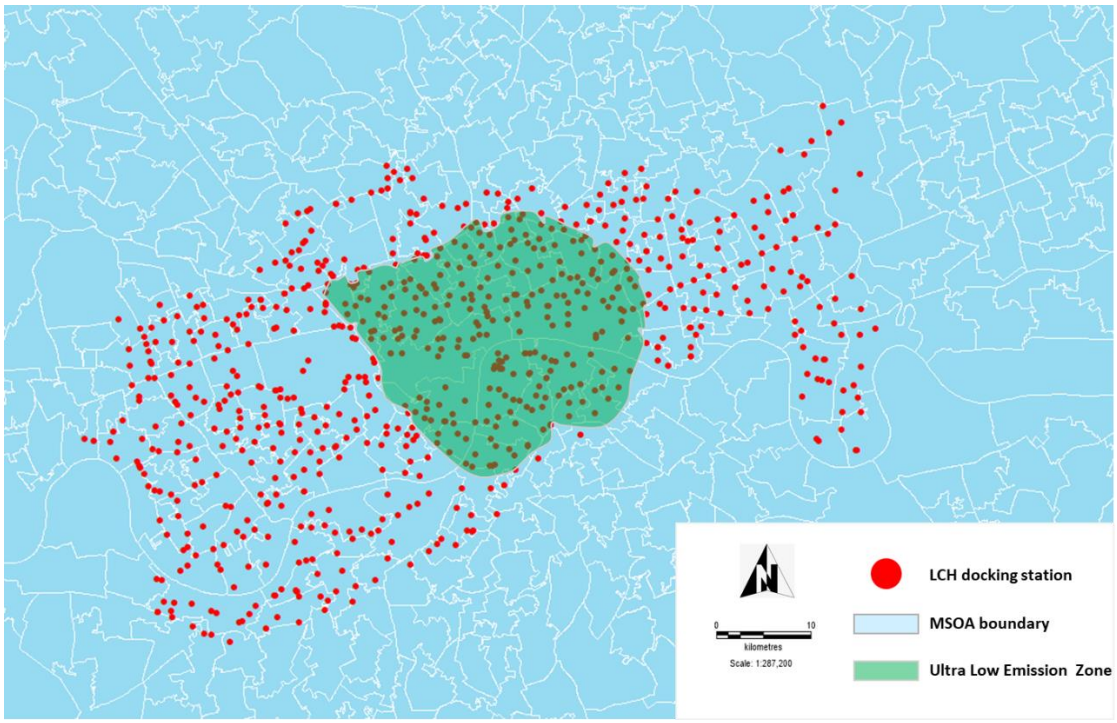


Figure 4.2 Distributions of docking stations

4.3 Estimation results of PSM

Prior to the estimation of the effect of ULEZ on bicycle demand, it is necessary to establish an appropriate control group for each treated unit using the PSM approach. Firstly, validity of the proposed PSM model would be assessed using the balancing test. A factor attribute of the treatment unit and corresponding control group is unbalanced when the t-statistic is significant. **Table 4.3** presents the result of balancing test. Results indicate that attributes including traffic flow (t = -14.04), residential area (-10.04), road

area (5.08), proportion of age above 64 (3.27), and proportion of age below 16 (-2.39) are unbalanced prior to propensity score matching. Despite that, bias attributed to the difference in attributes between treatment and control groups can be mitigated after matching. All covariates are balanced. Hence, conditional independence assumption (CIA) holds true.

Then, probability distributions of propensity score of treatment and control groups are established to testify the overlap assumption. Overlap area in the frequency distribution of propensity score indicates 'common support'. As shown in **Figure 4.3**, there is sufficient overlap for treatment and control groups. All treated and control units are within common support area. Hence, common support condition assumption holds true. In addition, propensity scores of treated units are higher (left skewed) than those of untreated units. Such phenomenon is reasonable as the propensity score implies the probability for a unit being treated.

Table 4.3 Results of balancing test for treatment and control groups

Covariate	Unmatched (U)/ Matched (M)	Mean		% reduction		t-test	
		Treatment	Control	% bias	bias	t-statistic	p-level
Income	U	57,023	56,962	0.7	-20.8	0.08	0.933
	M	57,023	56,950	0.9		0.32	0.930
Proportion of age above 64	U	0.116	0.105	26.8	92.4	3.27	0.001**
	M	0.116	0.116	2.0		0.21	0.830
Proportion of age below 16	U	0.278	0.290	-18.4	-2.3	-2.39	0.017*
	M	0.278	0.270	18.8		1.91	0.046
Log (Traffic flow)	U	4.275	4.379	-122.3	90.7	-14.04	0.001**
	M	4.275	4.282	-11.3		-1.41	0.158
Log (Population)	U	3.984	3.986	-1.8	-220.8	-0.21	0.830
	M	3.984	3.991	-5.9		-1.41	0.495
Proportion of residential area	U	0.111	0.164	-92.1	88.0	-10.04	0.001**
	M	0.111	0.105	11.0		1.38	0.167

Proportion of road area	U	0.346	0.314	45.8	70.3	5.08	0.001**
	M	0.346	0.356	-13.6		-0.95	0.344
Density of bus station	U	0.026	0.028	-8.2	94.5	-0.98	0.329
	M	0.026	0.026	-0.5		-0.05	0.959

** and ** denote statistical significance at the 5% and 1% levels respectively.*

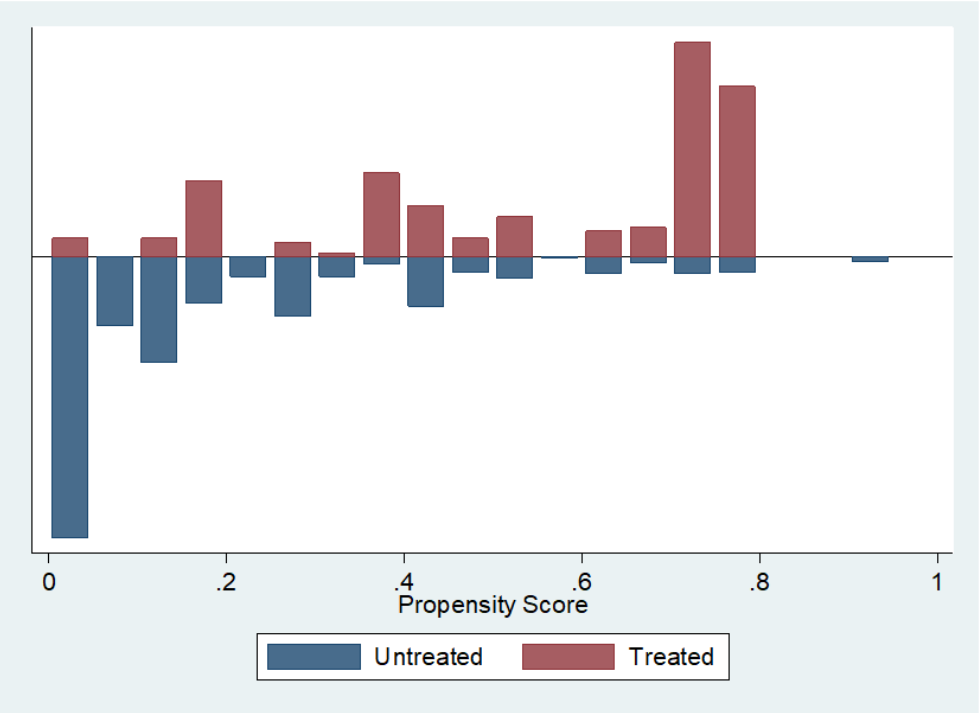


Figure 4.3 Results of overlap test

Multiple matching algorithms are considered to ascertain the robustness of estimation results. For instance, five typical matching algorithms, including K-nearest neighbours (K=1), K-nearest neighbours (K=3), K-nearest neighbours (K=5), kernel matching (bandwidth=0.05), and radius matching (caliper=0.05), are considered. As shown in **Table 4.4**, bicycle demand of treated units is 15.4% higher than that of control units before matching. Estimation results among the five matching algorithms are comparable (24.9% to 27.9% higher for the treated unit). This implies that the estimation results are independent of the matching algorithm. Indeed, robustness of the results could be assessed pragmatically using this approach (Caliendo and Kopeinig, 2005). In the

subsequent analysis, kernel matching algorithm (with the greatest t-statistic among the five matching algorithms) would be adopted to estimate the treatment effect.

Table 4.4 Effects of ULEZ on overall bike sharing usage

Matching algorithm	Treatment	Control	Difference	S.E.	t-statistic	% change
Unmatched	7,949.15	6,727.83	1,221.32	421.01	2.90	15.4%**
K-nearest neighbours matching (K=1)	7,949.15	5,791.19	2,157.96	1,026.39	2.10	27.1%*
K-nearest neighbours matching (K=3)	7,949.15	5,973.69	1,975.45	848.78	2.33	24.9%*
K-nearest neighbours matching (K=5)	7,949.15	5,882.37	2,066.79	811.18	2.55	26.0%*
Kernel matching	7,949.15	5,728.27	2,220.88	688.79	3.22	27.9%**
Radius matching	7,949.15	5,759.44	2,189.71	686.94	3.19	27.5%**

** and ** denote statistical significance at the 5% and 1% levels respectively.*

4.4 Effects on bike sharing usage

4.4.1 Effect on bike sharing usage by trip duration

Two membership subscription options are available for the LCH scheme: (i) 2 pounds for 24-hour access; and (ii) 90 pounds for 365-day access. All journeys up to 30 minutes are free of charge within the access period. The charge would increase by 2 pounds for each additional 30 minutes. Therefore, it would be crucial to evaluate the variation in the effects of ULEZ on bicycle demand by trip duration. As shown in **Table 4.5**, estimation is stratified into three: (i) trip shorter than 15 minutes, (ii) trip from 15 to 30 minutes, and (iii) trip longer than 30 minutes. Results indicate that there are 25.3% and 28.8% increases in short (within 15 minutes) and intermediate (15 to 30 minutes) bicycle trips after the implementation of ULEZ. This indicates that the introduction of ULEZ can stimulate bicycle demand. Hence, shift to green transportation mode can be promoted (Peters et al.,

2021; Ellison et al., 2013). It could be because of the behavioural changes of residents within ULEZ (avoid emission charges by shifting to cycling) and the expansion of bicycle infrastructures, i.e., Cycle Superhighways (Ding et al., 2021a, b). For instance, eight Cycle Superhighways have been introduced in London since 2008. This should provide safer, faster and more direct routes for cyclists travelling in central London (Li et al., 2018). Also, commuters might not prefer long bicycle journeys (Li et al., 2019). Hence, there is no significant change in the demand for long (more than 30 minutes) bicycle trips.

Table 4.5 Effects of ULEZ on bike sharing usage by trip duration

Trip duration	Matching	Treatment	Control	Difference	S.E.	t-statistic	% change
Shorter than 15 minutes	Unmatched	4,978.97	4,258.26	720.71	255.69	2.82	25.3%**
	Matched	4,978.97	3,718.54	1,260.43	421.37	2.99	
15 to 30 minutes	Unmatched	2,162.06	1,854.78	307.28	140.82	2.18	28.8%**
	Matched	2,162.06	1,540.34	621.72	230.37	2.70	
Longer than 30 minutes	Unmatched	808.11	614.78	193.33	111.51	1.73	41.9% (IS)
	Matched	808.11	469.38	338.74	178.69	1.90	

** and ** denote statistical significance at the 5% and 1% levels respectively; IS denotes insignificant.*

4.4.2 Effect on bike sharing usage by trip destination

Effects of ULEZ on bike sharing usage by trip destination are also assessed. Hence, estimation is stratified into two: (i) return (destination) station within ULEZ; and (ii) return station outside ULEZ. As shown in **Table 4.6**, an increase (44.8%) in bicycle demand is significant for journeys ended within ULEZ. This could be attributed to the frequent travel activities in the area (García-Palomares et al., 2012; Gutiérrez et al., 2020; Faghih-Imani et al., 2014). For example, commercial, office, shopping and green areas constitute 57% of the land area of ULEZ. The increase in bicycle demand is incremental for the journeys that ended outside ULEZ (16.1%), although it is insignificant. Findings

should be indicative to the optimal bicycle relocation strategy that can enhance the level of service of the bike-sharing system.

Table 4.6 Effects of ULEZ on bike sharing usage by trip destination

Trip destination	Matching	Treatment	Control	Difference	S.E.	t-statistic	% change
Within ULEZ	Unmatched	3,278.77	2,038.86	1,239.90	201.71	6.15	44.8%**
	Matched	3,278.76	1,808.55	1,470.22	324.68	4.53	
Outside ULEZ	Unmatched	4,670.38	4,688.97	-18.58	248.12	-0.07	16.1% (IS)
	Matched	4,670.38	3,919.72	750.66	412.55	1.82	

** and ** denote statistical significance at the 5% and 1% levels respectively; IS denotes insignificant.*

4.4.3 Effect on trip duration by trip destination

Effects of ULEZ on bicycle trip duration are also estimated. As shown in **Table 4.7**, no significant change is found for bicycle trip duration after implementing ULEZ, regardless of trip destination. It could be because bicycle trip duration is more closely related to the attributes like bicycle network characteristics (i.e., connectivity), real-time weather and traffic conditions (Li et al., 2018; Jappinen et al., 2013). Relevant information is, however, not available in the current study. Therefore, it is worth exploring the moderating effects of traffic control and management on the association between ULEZ and bicycle trip duration when more comprehensive information is available in future research.

Table 4.7 Effects of ULEZ on bicycle trip duration (minute) by trip destination

Trip destination	Matching	Treatment	Control	Difference	S.E.	t-statistic	% change
Overall	Unmatched	19.37	20.51	-1.14	0.43	-2.62	-0.03% (IS)
	Matched	19.37	20.00	-0.63	0.74	-0.87	
	Unmatched	20.48	21.16	-0.67	0.61	-1.11	1.4% (IS)

Within ULEZ	Matched	20.48	20.19	0.29	1.01	0.29	
Outside ULEZ	Unmatched	18.98	20.94	-1.96	0.45	-4.39	-7.2% (IS)
	Matched	18.98	20.34	-1.37	0.75	-1.82	

** and ** denote statistical significance at the 5% and 1% levels respectively; IS denotes insignificant.*

4.5 Concluding remarks

In this study, effects of ULEZ on bicycle demand (in terms of bike sharing usage) are examined using PSM approach. For each treatment unit, a set of control group is established considering the covariates including population density, socio-demographics, land use, and transport facilities. Results indicate that bicycle demand significantly increase after the introduction of ULEZ. In particular, increases in short (within 15 minutes) and intermediate (15 to 30 minutes) bicycle trips are more remarkable, compared to long bicycle trips (more than 30 minutes). In addition, results also indicate that number of bicycle trips ended within ULEZ increase remarkably. However, no significant change can be found for the number of bicycle trips ended outside ULEZ and bicycle trip duration.

Findings of this study can be subjected to the effects of unobserved factors that are not considered in the proposed PSM model because of the availability of required information. For instance, seasonal variation and weather conditions can also affect the cyclists' travel behaviour. It is worth exploring the moderation effect of other confounding factors on bicycle demand when more comprehensive data are available in future study. Furthermore, data adopted are transaction records of the LCH scheme. This can be limited to the penetration rate of commuters in the study area. Hence, it is also worth investigating the effect of ULEZ on bicycle route choice and cycling distance in future study.

Chapter 5 Effect of policy intervention on bicycle safety

5.1 Introduction

In recent years, many policy interventions have been introduced to promote cycling around the world. Take London cycle hiring (LCH) program as an example, residents in London suggested that they were inspired by the LCH that was launched in July 2010 to start cycling (ITV, 2014). In 2010, there were 5,000 bicycles and 315 docking stations for the LCH program. By 2018, the number of bicycles increased to 11,500, and the number of docking stations increased to 750, respectively. The location of LCH docking stations is shown in **Figure 5.1**. Previous studies on cycle hire schemes mainly focused on travel behaviour, transport mode share and environmental benefits (Li et al., 2019; Fishman et al., 2014; Zhang et al., 2017; Campbell et al., 2016). It was rare that the effect of the cycle hire scheme on bicycle safety was investigated. Bicycle safety is an important metric affecting bicycle network planning and design. Indeed, bicyclists are vulnerable to road injuries compared with motor vehicle occupants (Nikitas et al., 2014). We hypothesize that the overall bicycle crash may increase after the introduction of the LCH program since there are more new bicyclists on the roads. For instance, 49% of LCH users were encouraged by the scheme to start cycling in London (ITV,2014). Hence, it is of essence to evaluate the effect of the LCH scheme on bicycle safety.

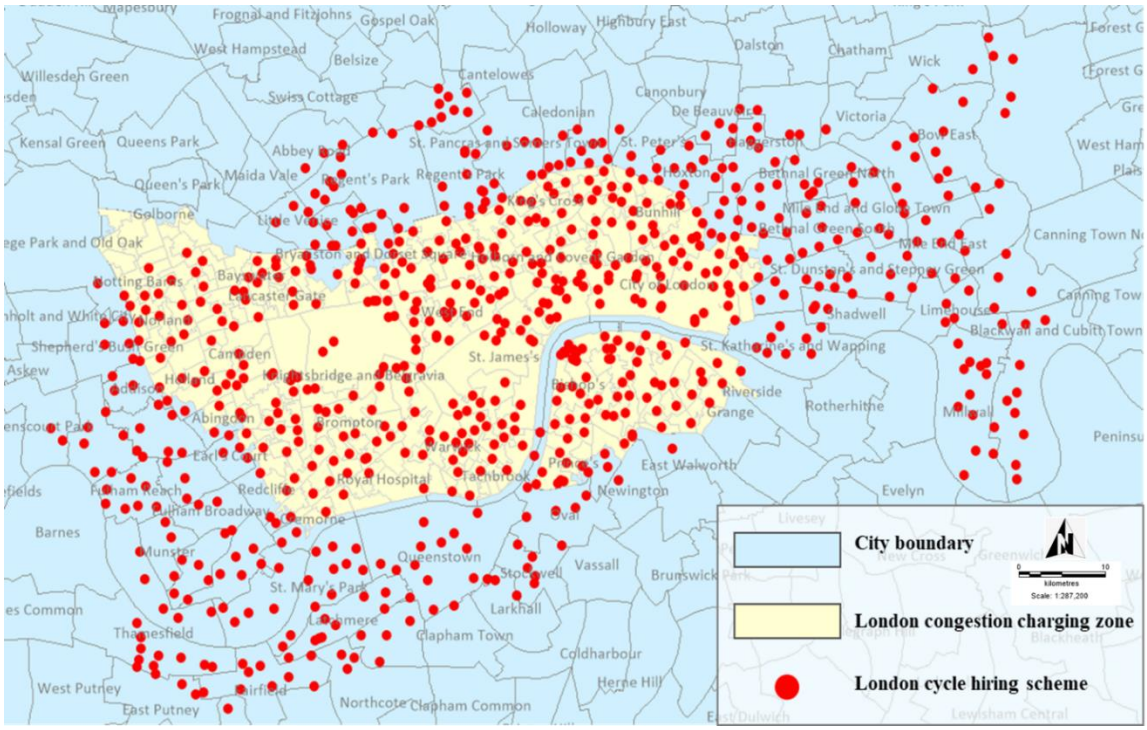


Figure 5.1 Locations of bicycle docking stations in London

On the other hand, some of the LCH docking stations are in the congestion charging area. London Congestion Charging scheme (LCC) was introduced in February 2003. LCC covered an area of 21 km^2 (also shown in **Figure 5.1**) and accounted for about 1.3% of the total area of Greater London. A few studies indicated that congestion charging was associated with reducing motor vehicle crashes but increasing bicycle crashes (Li et al., 2012). Therefore, it would be interesting to examine the role and impact of multiple policies on bicycle safety.

In this study, the Propensity Score Matching (PSM) method is applied to evaluate the influences of policy interventions (i.e. cycle hire and congestion charging schemes) on bicycle crashes, with which the effects of confounding factors are accounted using a systematically established 'control' group. Findings of this study are indicative to the decision making of transport planners that can improve the design of bicycle network and enhance the overall bicycle safety.

This chapter is organized as follows. Section 5.2 illustrates the study design, including data collection and selection of treated and control groups. Section 5.3 and Section 5.4

presents the analysis results and discussions respectively. Finally, concluding remarks and study limitations are summarized in Section 5.5.

5.2 Study design

5.2.1 Covariates affecting bicycle safety

As mentioned in chapter 4, validity of PSM largely depends on the unconfoundedness assumption. Unfortunately, level of confoundedness is not assessable. To avoid the violation of unconfoundedness assumption, all observed covariates, regardless of their significance to the "treatment", that may affect the outcome should be considered when calculating the propensity score. In this study, as the outcome is bicycle crash frequency, possible factors contributing to bicycle safety will be considered to achieve optimal precision and minimize the bias when estimating the propensity score (Brookhart et al., 2006).

In this study, observation unit is Lower Super Output Area (LSOA). LSOA is the primary unit of population census, home affairs administration and election in the United Kingdom. Each LSOA has a population of 1,500 on average. One of the 'interventions' under investigation is the LCC scheme, which is in force from 7:00 am to 6:00 pm on weekdays. Hence, bicycle crashes that occurred in the evenings and on the weekends would be excluded in the subsequent analysis. Bicycle crash data is obtained from the Department for Transport (DfT) dataset. It provides information on crash location, casualty age, gender and vehicle type of every bicycle crash involving personal injury.

In this study, covariates are primarily derived from those revealed in conventional bicycle crash prediction models. Hence, the possible covariates, including population characteristics, built environment, and transport infrastructures, are considered (Li et al., 2012; Wang et al., 2017; Guo et al., 2018b; Sze et al., 2019; Guo et al., 2019). Information on population demographic and socioeconomic characteristics (i.e. genders, age, and household income) are obtained from the Office for National Statics (ONS) database.

Land use (i.e. residential, commercial, green area and transport infrastructure) data is obtained from the Greater London Authority (GLA) database and transport network data (i.e. Class A road, Class B road and minor road lengths, traffic volume, bicycle flow and bus stop, etc.) is obtained from the DfT database. **Table 5.1** summarizes the covariates considered in the proposed PSM model.

Table 5.1 Summary statistics of the sample

Factor	Attribute	Mean	S.D.	Min.	Max.
Number of observations = 333 (LSOA)					
Bicycle crash frequency	Total bicycle crash	3.06	6.01	0	132
	Killed and severely injured crash	0.45	1.10	0	21
	Slightly injured crash	2.61	5.16	0	111
Population density	Population per km^2	13.06	5.98	0.62	49.85
Gender	Proportion of male	0.50	0.03	0.40	0.63
	Proportion of female	0.50	0.03	0.37	0.60
Age	Proportion of age above 64	0.09	0.04	0.02	0.21
	Proportion of age below 16	0.16	0.05	0.03	0.33
Income	Annual average household income (€)	50,626	18,444	26,140	153,420
Land use	Proportion of residential area	23.50	12.09	2.29	202.59
	Proportion of business and office area	27.58	49.24	0.48	1,041
	Proportion of green area	70.75	92.11	4.39	1,291
	Proportion of road, railway and footpath area	49.54	49.44	7.46	672.11
Road density	Class A road (km per km^2)	4.29	3.01	0	18.21
	Class B road (km per km^2)	0.60	1.44	0	13.40
	Minor road (km per km^2)	0.75	1.27	0	6.60
Traffic flow	Annual average daily traffic	16,110	11,847	42.5	108,828

Factor	Attribute	Mean	S.D.	Min.	Max.
Bicycle flow	Annual average daily bicycle flow	825	787	0	5,458
Density of bus stop	Bus stop per km^2	0.04	0.03	0	0.22
Cycle superhighway	Length of Cycle Superhighway (km)	1.41	1.45	0	6.22

5.2.2 Treatment and control groups

333 LSOAs are considered in this study. As shown in **Table 5.2**, LCC was imposed in 33 LSOAs, and LCH was introduced in 132 LSOAs, respectively. Since PSM is a 'data-hungry' approach that a large sample of treated and control units is required, as shown in **Table 5.2**, 201 LSOAs that have no LCH nor LCC are considered to ensure sufficient overlap (Wood and Donnell, 2017). To increase the sample size, two-year data (i.e. 2011 and 2012) are used. Therefore, a total number of analysis unit is 666. This study will evaluate the safety effect of LCH only (Analysis I) and marginal safety effect of LCC on LCH (Analysis II). For Analysis I, treated units refer to those with LCH only, and control units refer to those with neither LCH nor LCC imposed, respectively. For Analysis II, treated units refer to those with both LCH and LCC and control units refer to those with LCH only, respectively. This justifies the Stable Unit Treatment Value assumption (SUTVA). **Figure 5.2** illustrates the spatial distributions of treated and control units for the two analyses.

Table 5.2 Study design of proposed analysis

Characteristics of LSOA	Number of LSOA	Analysis	
		I. LCH only	II. Marginal effect of LCC
LCH only	99	Treated units	Untreated units
LCH and LCC	33	N/A	Treated units
Neither LCH nor LCC	201	Untreated units	N/A

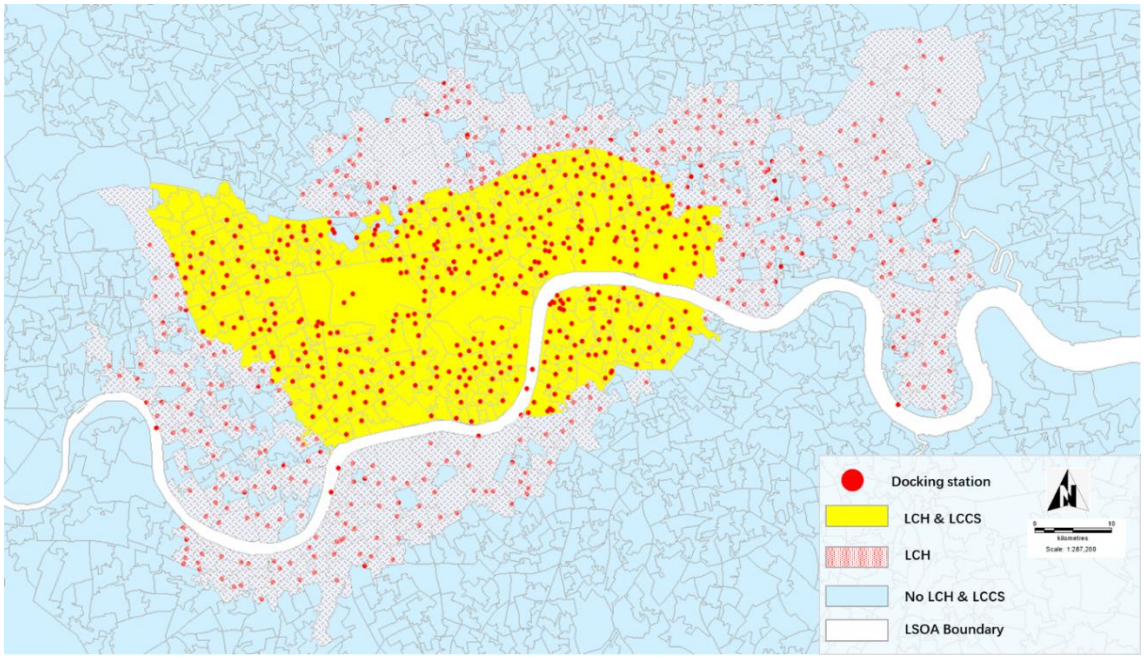


Figure 5.2 Distribution of LSOA by policy interventions

5.3 Estimation results of PSM

Prior to the evaluation of policy intervention on bicycle crash incidence, it is necessary to construct an appropriate 'control' group for every 'treated' unit using PSM approach. Firstly, a balancing test would be conducted to assess the validity of PSM, so that the propensity to receive 'treatment' is independent of the outcome. **Table 5.3** presents the results of balancing test. As shown in **Table 5.3**, the 'treated' and 'control' groups were imbalanced for all covariates at the 5% significance level before matching (U-Unmatched). Favourably, bias in effectiveness evaluation can be eliminated by refining the control groups using the proposed matching algorithm. The 'treated' and 'control' groups are balanced for all covariates after matching (i.e. M - Matched). This justifies the Conditional Independence Assumption (CIA).

Table 5.3 Results of balancing test for treatment and control groups

Covariate	Unmatched (U)/ Matched (M)	Mean		% reduction		t-test	
		Treatment	Control	% bias	bias	t-statistic	p-level
Income	U	49,514	45,063	35.7	97.0	3.90	0.000*
	M	49,514	49,646	-1.1		-0.09	0.928
Population density	U	13.12	13.23	-1.9	-566.3	-0.21	0.836
	M	13.12	13.90	-13.0		-1.20	0.233
Male	U	0.499	0.493	25.4	60.5	2.86	0.004*
	M	0.499	0.501	-10.0		-0.92	0.359
Age above 64	U	0.089	0.089	0.4	-2082	0.04	0.965
	M	0.089	0.086	8.6		0.84	0.401
Age under 16	U	0.162	0.183	-44.4	91.6	-4.78	0.000*
	M	0.162	0.161	3.7		0.35	0.730
Business and office area	U	25.26	19.00	22.2	74.1	2.29	0.023*
	M	25.26	23.64	5.8		0.39	0.695
Road area	U	47.15	45.63	4.0	-102.9	0.42	0.678
	M	47.15	44.07	8.1		0.76	0.447
Green area	U	71.52	88.11	-17.3	80.0	-1.95	0.052
	M	71.52	74.82	-3.5		-0.36	0.717
Class A road	U	4.479	3.771	23.3	38.8	2.62	0.009*
	M	4.479	4.046	14.3		1.41	0.160
Class B road	U	0.489	0.604	-8.0	60.9	-0.83	0.405
	M	0.489	0.534	-3.1		-0.36	0.719
Minor road	U	0.493	1.001	-41.2	75.7	-4.17	0.000*
	M	0.493	0.618	-10.0		-1.17	0.243
Traffic flow	U	18,103	14,559	29.5	65.5	3.21	0.001*
	M	18,103	19,327	-10.2		-0.80	0.426
Bicycle flow	U	880.3	561.3	49.8	92.0	5.57	0.000*
	M	880.3	854.8	4.0		0.33	0.744
	U	0.069	0.024	21.4	74.3	2.53	0.012*

Covariate	Unmatched (U)/ Matched (M)	Mean		% reduction		t-test	
		Treatment	Control	% bias	bias	t-statistic	p-level
Cycle Superhighway	M	0.069	0.058	5.5		0.44	0.661

** Statistical significance at the 5% level*

Additionally, validity of PSM can be assessed graphically based on the propensity score distributions of treated and control groups. Overlap area in the frequency distribution of propensity score indicates 'common support'. Units in the region of common support are referred to as 'on support' and 'off-support' otherwise. As shown in **Figure 5.3**, overlaps of treated and control groups are enough, and all units are 'on support'. Hence, the Common Support Condition (CSC) assumption is justified.

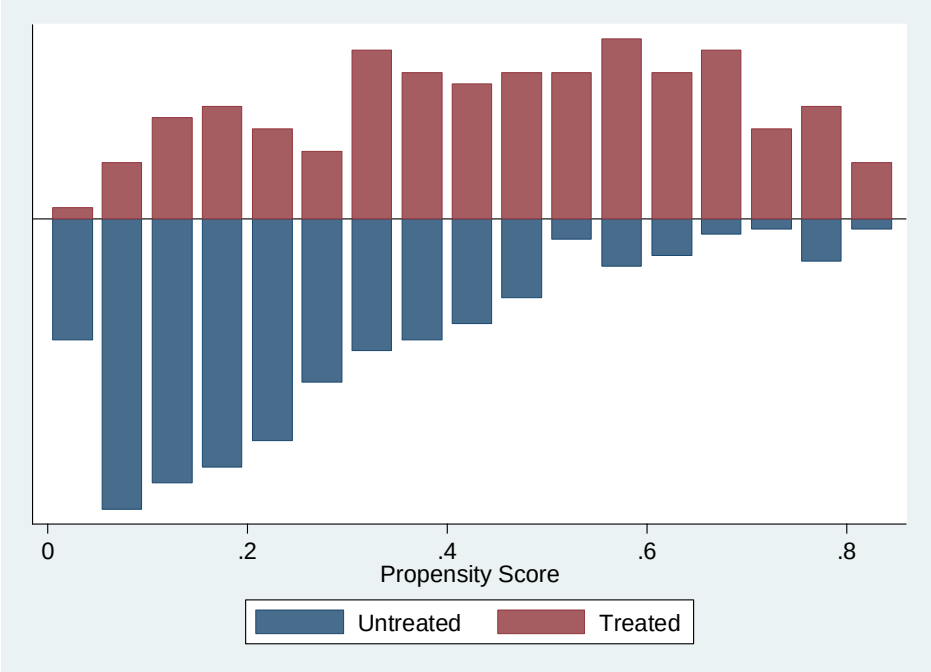


Figure 5.3 Results of overlap test

5.4 Safety effects of LCH and LCC schemes

5.4.1 Safety effect of LCH scheme

Table 5.4 illustrates the estimation results of the effect of LCH on (i) overall bicycle crash; (ii) killed and severely injured (KSI) bicycle crash; and (iii) slightly injured bicycle crash. As shown in **Table 5.4**, overall bicycle crash (37.7%) and slightly injured crash (31.8%) increased significantly when LCH is implemented, both at the 5% level, after controlling the possible confounding factors using PSM. It could be because of the increase in cyclists on the roads. Indeed, 49% of bicyclists in London admitted that they were encouraged to cycle by the LCH (ITV, 2014). To this end, we also evaluated bicycle usage changes in the treated LSOAs. As shown in **Table 5.5**, the increase in bicycle usage (when LCH was present) is remarkable at the 5% level. Such an increase in bicycle usage (37.3%) is comparable to overall and slight bicycle crashes (32-38%, as shown in Table 5.4). This justified that the unfavourable safety effect by LCH could be attributed to the increase in bicyclists on the roads (TfL, 2018). Moreover, the results indicated no significant difference in the occurrence of KSI bicycle crash between treated and control LSOAs. It could be because most bicycle docking stations are in the area where the speed limits are usually lower than 30 mph. Therefore, it is unlikely that the injury risk be elevated (Li and Graham, 2016).

Table 5.4 Effect of LCH on bicycle crash incidence

Outcome	Sample	Treated	Untreated	Difference	Standard error	t-statistic	% change
Overall bicycle crash	Unmatched	3.10	1.74	1.35	0.21	6.32	37.7%*
	Matched	3.10	2.25	0.85	0.28	3.01	
Slight bicycle crash	Unmatched	2.61	1.51	1.10	0.18	5.98	31.8%*
	Matched	2.61	1.98	0.63	0.24	2.62	
KSI bicycle crash	Unmatched	0.48	0.23	0.25	0.06	4.08	IS
	Matched	0.48	0.27	0.22	0.08	1.64	

** Statistical significance at the 5% level; IS denotes insignificant.*

Table 5.5 Results of PSM for bicycle usage (LCH only)

Outcome	Sample	Treatment	Control	Difference	S.E.	t-stat	% change
Bicycle usage	Unmatched	980	561	418	63.0	6.65	37.3%*
	Matched	980	713	266	83.4	3.20	

** Statistically significant at the 5% level*

5.4.2 Safety effect of LCC scheme

Some LSOAs have both LCH and LCC schemes introduced. Since traffic flow patterns and speed could be changed in areas with LCC, it is crucial to estimate the marginal effect of LCC on bicycle crashes. As shown in **Table 5.6**, the marginal effects of LCC on overall bicycle crash (59.1%) and slightly injured bicycle crash (57.8%) are significant, both at the 5% level. However, as shown in **Table 5.7**, the traffic volume in the LSOAs that have both LCC and LCH is 21% lower than that have LCH only. This could be because of the dramatic increase in bicycles in the treated LSOAs (74.9%, as shown in **Table 5.7**) because of the mode shift after the introduction of the congestion charge (Li et al., 2012; Xie and Olszewski, 2011; Tang, 2016). Again, increase in KSI bicycle crash (66%) can be observed, though it is not significant. It could be because of the expansion of the bicycle infrastructure, particularly the Cycle Superhighways in the area (Li et al., 2017).

Table 5.6 Marginal effect of LCC on bicycle crash

Outcome	Sample	Treated	Untreated	Difference	Standard error	t-statistic	% change
Overall bicycle crash	Unmatched	5.92	3.11	2.81	0.58	4.84	59.1%*
	Matched	5.92	3.72	2.20	0.87	2.52	
Slight bicycle crash	Unmatched	5.02	2.61	2.42	0.50	4.83	57.8%*
	Matched	5.02	3.18	1.84	0.74	2.48	
	Unmatched	0.89	0.51	0.38	0.14	2.84	IS

KSI bicycle crash	Matched	0.89	0.54	0.36	0.19	1.85	

** Statistical significance at the 5% level; IS denotes insignificant.*

Table 5.7 Results of PSM for traffic flow and bicycle usage (LCH and LCC)

Outcome	Sample	Treatment	Control	Difference	S.E.	t-stat	% change
AADT	Unmatched	14916	16857	-1941	1508	-1.29	-21.3%*
	Matched	14684	18670	-3985	1862	-2.14	
Bicycle usage	Unmatched	1572	912	669	116	5.74	74.9%*
	Matched	1572	898	673	153	4.38	

** Statistically significant at the 5% level*

5.5 Concluding remarks

Policy interventions, including bicycle infrastructure development and bicycle sharing scheme, have been implemented worldwide to promote bicycle use. In London, a public bicycle hiring scheme (LCH) was introduced in 2010. Despite the public bicycle rental system effectively promoting green transport and improving the physical well-being of the community (Zhang and Mi, 2018; Ding et al., 2020), the safety effects of bicycle sharing were rarely investigated. This study contributes to the literature by estimating the effects of LCH on bicycle crash incidence, with which the possible confounding factors are considered using the PSM approach. Results of this study indicated that both the overall (38%) and slight bicycle crashes (32%) in areas with LCH introduced are remarkably higher than those with no LCH. However, no significant effect on KSI bicycle crash could be revealed. This could be attributed to effective traffic control measures and the development of bicycle infrastructures.

Moreover, this study also contributes to the literature by exploring the marginal effect of the London congestion charging scheme (LCC) on the LCH. Our results suggested that numbers of overall (59.1%) and slight bicycle crash (57.8%) in the areas with both LCC

and LCH introduced are remarkably higher than those with LCH only. It could be because of the possible mode shift (to active transport modes including cycling and walking) because of the congestion charging scheme (Li et al., 2012; Noland et al., 2008). Also, no significant changes could be found in the KSI bicycle crash.

The above findings are indicative to the decision-making of transport planners, particularly striking the balance between environmental benefit, physical health, traffic safety and societal impact when promoting green transport. However, it is noteworthy that the current approach does not consider the differences in crashes between the treated and control groups that might exist before the introductions of LCH and LCC. The extended study is worth exploring the mediation effects by possible factors before and after the interventions. Moreover, possible influences by the weather conditions and seasonal effects on the association are not considered in this study. Indeed, some covariates have different associations with crashes depending on the season and weather conditions. It is worth exploring the interactions by weather conditions on the safety effect of the bicycle sharing scheme when more comprehensive data are available in future studies (Ding et al., 2020).

Findings of this study can be subjected to the effects of unobserved factors that are not considered in the proposed PSM model because of the availability of required information. For instance, seasonal variation and weather conditions can also affect the cyclists' travel behaviour.

Chapter 6 Effect of built environment and population characteristics on bicycle travel

6.1 Introduction

Cycling is increasingly promoted as a sustainable transport mode. However, bicyclists are more vulnerable to fatality and severe injury in road crashes, compared to vehicle occupants. Identifying the contributory factors to crashes and injuries involving bicyclists is necessary. Therefore, effective engineering countermeasures can be developed to enhance the overall safety of bicyclists and promote the bicycle mode.

Many studies have investigated the effects of built environment and bicycle facilities on bicycle crash frequency at the macro-level using cross-sectional models (Siddiqui et al., 2012; Narayanamoorthy et al., 2013; Chen, 2015; Guo et al., 2018b). To evaluate the bicycle crash risk, it is necessary to estimate the exposure (i.e. quantifying the crash potential of bicyclists). For vehicle crashes, annual average traffic flow (AADT) and vehicle kilometre travelled (VKT), based on comprehensive traffic count data, can be used to estimate the exposure (Pei et al., 2012). However, bicycle count data are rarely available. Bicycle crash exposure may be measured using retrospective and prospective approaches based on self-report data. They are, however, subjected to self-selection problems. This study attempts to address the problem of how to accurately measure bicycle crash exposure based on the revealed bicycle trip data of a public bicycle rental system.

Additionally, effects of possible land use, built environment and bicycle infrastructure attributes on bicycle crash incidence are investigated. For example, Cycle Superhighway (‘Superhighway’) was introduced in London in the early 2010s, targeted to provide cyclists with safer, faster and more direct journeys through the city (Li et al., 2018). In this study, we aim to measure the association between possible factors and bicycle crash frequency at the zonal level, using the integrated crash, environment, population profile and traffic data of London in 2012-2013. A random parameter negative binomial model

would be developed to measure the association. Particularly, effect of the presence of Cycle Superhighway on bicycle crash risk will be considered. Moreover, separate bicycle crash prediction models would be developed for different seasons, i.e. from May to October and from November to April, considering the bicyclists' behaviour under different weather conditions.

Reminder of this chapter is organized as follows. Data is described in Section 6.2. Section 6.3 and Section 6.4 presents the analysis results and discussions, respectively. Finally, concluding remarks and study limitations are summarized in Section 6.5.

6.2 Data

6.2.1 Study area

Figure 6.1 illustrates the boundary of the study area under investigation. The study area covers several Inner London Boroughs like the City of London, Islington, Hackney, Tower Hamlets and Westminster, etc. The geographical area was 49.1 km^2, and the total population was 0.76 million in 2017. Similar to other global cities, cycling has recently become increasingly popular in London. In 2017, average daily bicycle trip was 30,170 in the study area. It constituted 2% of overall trips (TfL, 2018). Bicycles in London can generally be classified into three types: (i) privately owned bicycles, (ii) public bicycle rental systems (with docking stations), e.g. Santander Bike (i.e., LCH) and (iii) dockless bicycle sharing systems, e.g. moBike, Ofo and Urbo (Li et al., 2019). The LCH constituted 74% of overall bicycle trips in the study area (TfL, 2018). There are over 750 docking stations of the LCH in the study area. Therefore, bicycle exposure is estimated based on the ridership data of the LCH. Locations of the docking stations of the LCH are shown in **Figure 6.2**. Four Superhighways, e.g. CS2, CS3, CS7 and CS8, were opened in the study area during the period 2010-2013 (Li et al., 2018). There are three major cycle track or lane types: (i) segregated one-way cycle track, (ii) segregated two-way cycle track and (iii) non-segregated (one-way) cycle lane. The network map and typical layouts of Superhighways are illustrated in **Figure 6.1** and **Figure 6.3**, respectively.

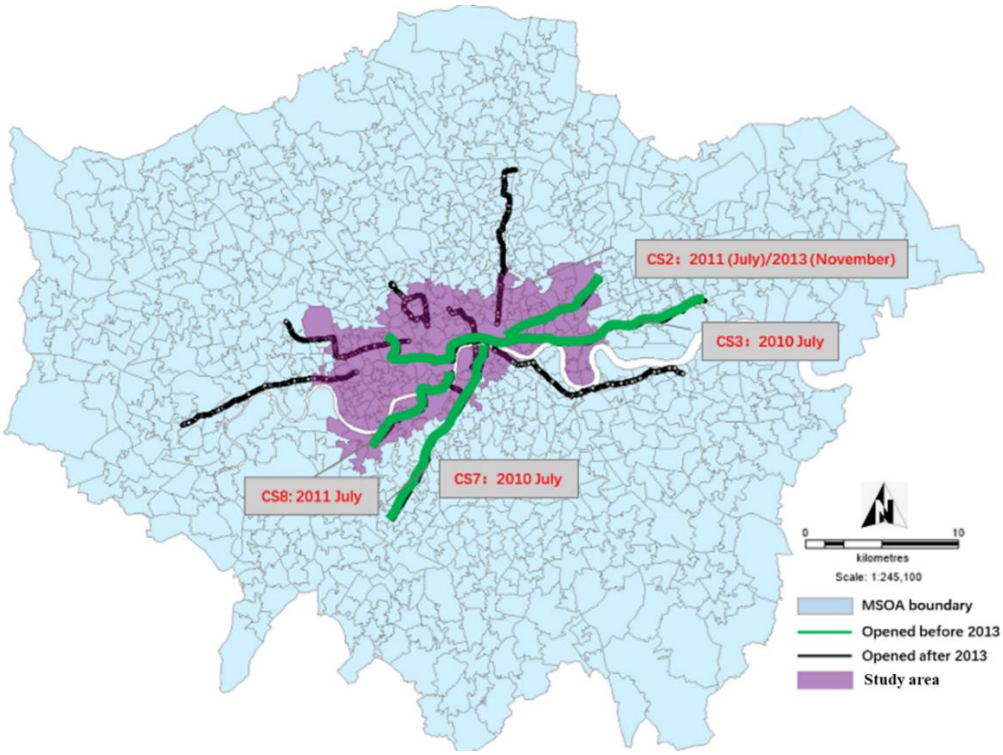


Figure 6.1 Location of the study area

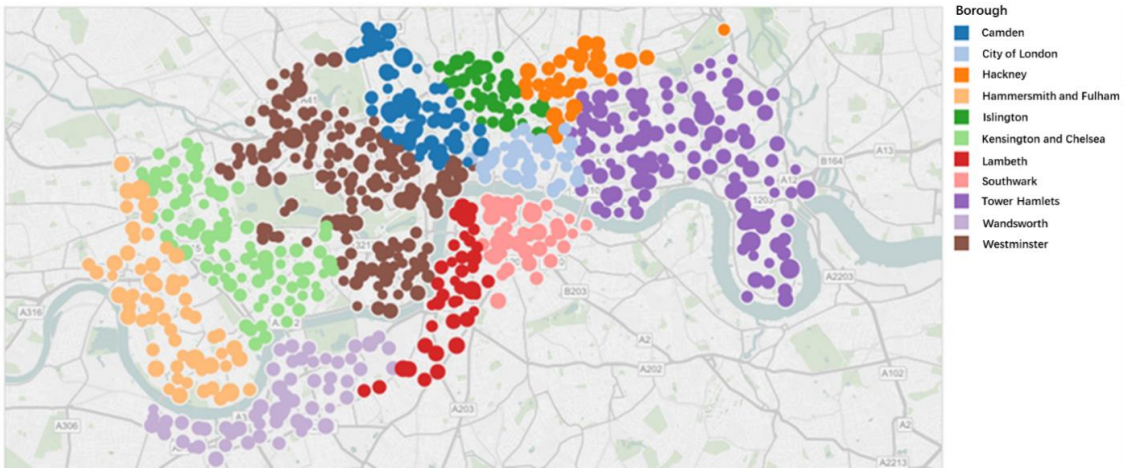


Figure 6.2 Locations of bicycle docking stations in the study area

(*Source: https://kitchen2018blog.blogspot.com/2018/02/boris-bikes-map.html*)

(a) Segregated one-way cycle track

(b) Segregated two-way cycle track

(c) Non-segregated cycle lane

Figure 6.3 Illustrations of Cycle Superhighway

(*Source: https://en.wikipedia.org/wiki/Cycle_Superhighway_3 ; https://www.newcivilengineer.com/archive/new-cycle-superhighway-mooted-21-09-2017/; https://www.geograph.org.uk/photo/2372620*)

6.2.2 Sample

Similar to the zone system adopted in Chapter 4, MSOAs are adopted in this study. Information on bicycle crash incidence, bicycle exposure, land use, road infrastructures, demographic and socioeconomics, and household attributes are matched into the corresponding MSOA using the Geographical Information System (GIS) technique. Specifically, bicycle crash data during the period from 2012 to 2013 is obtained from the Greater London Authority (GLA) collision data extract. Also, information on land use is available in the GLA's dataset. Information on socio-demographics and households in Greater London is available from the Office for National Statistics (ONS) census dataset. In addition, Department for Transport (DfT) dataset provides the transport network data.

In summary, frequencies of bicycle crash of 88 MSOAs in 2012-2013 are modelled. Sample size of the proposed model is 176. There were 2,795 bicycle crashes in the study area in the observation period. To consider the effect of bicycle infrastructure on bicycle

crash incidence, the length of Superhighways in every MSOA is also included in the model. **Table 6.1** summarizes the distribution of the variables considered.

Table 6.1 Summary statistics of the sample

Category	Factor	Attribute	Mean	S.D.	Min.	Max.
Outcome	Frequency of bicycle crash		15.88	18.17	2	144
Road infrastructure	Road density (km per km^2)		6.14	2.59	0.86	13.09
	Cycle superhighway (km)		0.37	0.52	0	1.87
Land use	Proportion for residential		0.20	0.09	0.05	0.42
	Proportion for commercial		0.23	0.10	0.05	0.51
	Proportion for green area		0.20	0.12	0.02	0.60
	Proportion for transport facilities		0.37	0.06	0.24	0.55
Demographics	Population density (per km^2)		19.52	7.23	3.05	35.8
	Gender	Proportion of male	0.51	0.025	0.454	0.61
		Proportion of female	0.49	0.025	0.39	0.55
	Age	Proportion of age above 64	0.10	0.03	0.03	0.24
		Proportion of others	0.90	0.03	0.76	0.96
Socio-economics	Race	Proportion of white	0.53	0.08	0.34	0.69
		Proportion of others	0.47	0.08	0.31	0.66
	Median annual household income (€)		71,369	28,673	37,130	174,960
	Household type	Proportion of couple with children	0.11	0.03	0.05	0.20

		Proportion of others	0.89	0.03	0.80	0.95
Exposure	Total annual bicycle usage time (hour)	Overall	31,141	31,483	2,498	165,577
		May to October only	20,500	21,082	1,006	110,520
		November to April only	10,641	10,802	880	59,813
	Total annual bicycle use frequency	Overall	87,946	81,980	7,049	476,329
		May to October only	54,768	51,755	5,014	289,498
		November to April only	33,178	31,803	2,016	186,831
	AADT		20,365	11,404	4,306	62,889

6.3 Estimation results

First, the multi-collinearity test is conducted to assess the correlations between the independent variables. Results indicate that the variance inflation factor (VIF) are less than five for all independent variables. Therefore, all candidate variables are considered appropriate.

6.3.1 Overall model

In this study, random parameter negative binomial model is applied to measure the association between bicycle crash frequency and possible risk factors, considering the effect of bicycle exposure. **Table 6.2** shows the results of parameter estimation. Three exposure measures considered are population (Model 0), bicycle use time (model 1) and bicycle use frequency (Model 2).

As shown in **Table 6.2**, AIC and BIC of Model 1 are the lowest among the three models. Hence, model using bicycle use time as the exposure measure is considered. Model using population as exposure is underperformed since it does not account for the difference in

travel patterns among individuals (Guo et al., 2018b; Wang et al., 2017; Lee et al., 2015a). On the other hand, studies also indicated that it was appropriate to indicate bicycle safety using relative risk (RR) with respect to travel distance and travel time (Mindell et al., 2012; Vanparijs et al., 2015). Hence, it can be expected model using bicycle use time as the exposure can achieve better goodness of fit.

Results indicate that road density, green area and commercial area, population, age, gender, household income and race contribute to bicycle crash frequency at the 5% significance level. For instance, increases in road density (parameter = 0.04), proportion of green area (1.14), proportion of commercial area (1.60), proportion of elderly (6.49), proportion of male (13.58), median annual household income (0.001), and proportion of white (0.03) are associated with the increase in bicycle crash frequency. Also, the random effects of demographic and socioeconomic characteristics on crash incidence are significant at the 5% level. However, no evidence can be established for the association between bicycle crashes, Cycle Superhighways, household composition and traffic volume. No obvious association between bicycle crashes and traffic volume is revealed could be because of the “safety in number” effect. In other words, the number of bicycle crash does not necessarily increase proportionately with the increase in traffic volume (Bjornskau et al., 2015).

Table 6.2 Results of parameter estimation of overall bicycle crash prediction model

Category	Factor		Model 0		Model 1		Model 2	
			Coefficients	T-stat	Coefficients	T-stat	Coefficients	T-stat
Constant			-21.05**	-6.65	-17.59**	-6.40	-18.50**	-6.79
Road infrastructure	Cycle Superhighway		IS	--	IS	--	IS	--
	Road density		0.07**	3.62	0.04*	2.31	0.05*	2.19
Land use	Proportion of green area		1.25**	3.32	1.14**	3.43	1.56**	3.27
	Proportion of commercial area	Mean	1.50**	3.60	1.60**	3.89	1.87**	3.55
		S.D.	*(9.86**)*					
Demographic	Log (population)	Mean	9.02**	4.39	1.80**	4.19	1.39*	2.56
		S.D.			*(12.08**)*		*(6.76**)*	
	Proportion of age above 64	Mean	5.40**	3.11	6.49**	4.39	6.19**	3.55
		S.D.					*(0.72^)*	
	Proportion of male	Mean	14.57**	8.51	13.58**	8.30	13.74**	6.81
		S.D.	*(0.36**)*		*(0.38**)*			
Socio-economics	Median annual household income		<0.001*	2.36	<0.001*	2.03	<0.001^	1.83
	Proportion of white	Mean	0.03**	4.64	0.03**	5.24	0.03**	3.64
		S.D.			*(<0.001^)*		*(0.002**)*	

	Proportion of couple with children	-0.04**	-2.61	IS	--	IS	--
Exposure	Total annual bicycle use frequency					2.98**	4.63
	Total annual bicycle use time (hour)			2.00**	4.38		
	Log (AADT)	IS	--	IS	--	IS	--
Goodness-of-fit	AIC	1150.43		1136.17		1144.64	
	BIC	1194.99		1186.19		1194.66	

1 ^, * *and* ** *denote statistical significance at the 10%, 5% and 1% levels respectively; IS denotes insignificant.*

6.3.2 Segregated Models

Considering the possible interferences by seasonal effects on the association between bicycle crash incidence and contributory factors, separate bicycle crash prediction models for different seasons: (i) warm season, i.e. May to October and (ii) cold season, i.e. November to April, are developed. **Table 6.3** presents the results of parameter estimation of separate models. Consistent with the overall model results, factors including commercial area, population, elderly, gender and race are significantly correlated to bicycle crash frequency, at the 5% level, in both the warm and cold seasons. Again, the random effects of demographic and socioeconomic characteristics on crash incidence are significant at the 5% level. However, 'Superhighway' is significant only in the cold season (parameter = -0.18), and 'green area' is significant only in the warm season (1.22).

Table 6.3 Results of parameter estimation results of separate model

Category	Factor		Warm Season		Cold Season	
			Coefficients	T-stat	Coefficients	T-stat
Constant			-18.57**	-6.11	-17.59**	-6.40
Road infrastructure	Cycle Superhighway		IS	--	-0.18*	-2.43
	Road density		0.08**	2.91	0.08**	3.37
Land use	Proportion of green area		1.22**	3.01	IS	--
	Proportion of commercial area		1.88**	3.23	1.53**	3.37
Demographic	Log (population)	Mean	2.52**	3.75	1.29*	2.24
		S.D.	*(2.92**)*		*(22.99**)*	
	Proportion of age above 64	Mean	6.75**	3.51	4.84*	2.57
		S.D.	*(0.13*)*			
	Proportion of male	Mean	12.99**	5.57	13.28**	5.67
		S.D.	*(0.17*)*		*(0.17*)*	
Socio-economics	Median annual household income		IS	--	IS	--
	Proportion of white	Mean	0.03**	3.94	0.03**	4.35
		S.D.			*(0.01**)*	
	Proportion of couple with children		IS	--	IS	--
Exposure	Log (AADT)		IS	--	IS	--

	Bicycle use time (hour)	1.13*	1.99	1.81**	3.22
Goodness-of-fit	AIC	1014.47		897.48	
	BIC	1064.50		947.51	

** and ** denote statistical significance at the 5% and 1% levels respectively; IS denotes insignificant.*

6.4 Discussions

6.4.1 Seasonal effect

There is a remarkable variation in the usage of rental bicycles (LCH) across months. **Figure 6.4** illustrates the monthly bicycle rental counts in the study area in 2012 and 2013. As shown in **Figure 6.4**, bicycle usage distribution is similar to the mean daily maximum temperature. In particular, average monthly bicycle usage (ranging from 557,142 to 771,428) in the period from May to October (with mean daily maximum temperature ranging from 61°F to 73°F) was remarkably higher than that (ranging from 282,857 to 454,285) in the period from November to April (with mean daily maximum temperature ranging from 48°F to 59°F). We may consider the commuters who cycle even in the cold season as regular bicyclists, while those who only cycle in the warm season as casual bicyclists. As revealed in the crash statistics in 2012-2013, the total number of bicycle crashes in the warm season (1,680) was remarkably higher than that in the cold season (1,115). A possible reason is that there are more bicyclists in the warm season. Indeed, it is believed that casual bicyclists, who are expected to ride more in the warm season, usually ride for leisure purposes (TfL, 2018). That is why the proportion of green area is positively correlated to the bicycle crash frequency in the warm season only (as revealed in **Table 6.3**). Therefore, it is necessary to implement effective education and promotion measures to enhance the safety awareness and perception of casual bicyclists, particularly children, adolescents and the elderly. On the other hand, presence of Cycle Superhighway is negatively associated with the bicycle crash frequency in the cold season only. This justifies the safety benefit of upgrading bicycle infrastructure, particularly for more skilful regular bicyclists. Nevertheless, it is worth exploring the seasonal trend using the disaggregated model for every month or season when comprehensive data is available in the extended study.

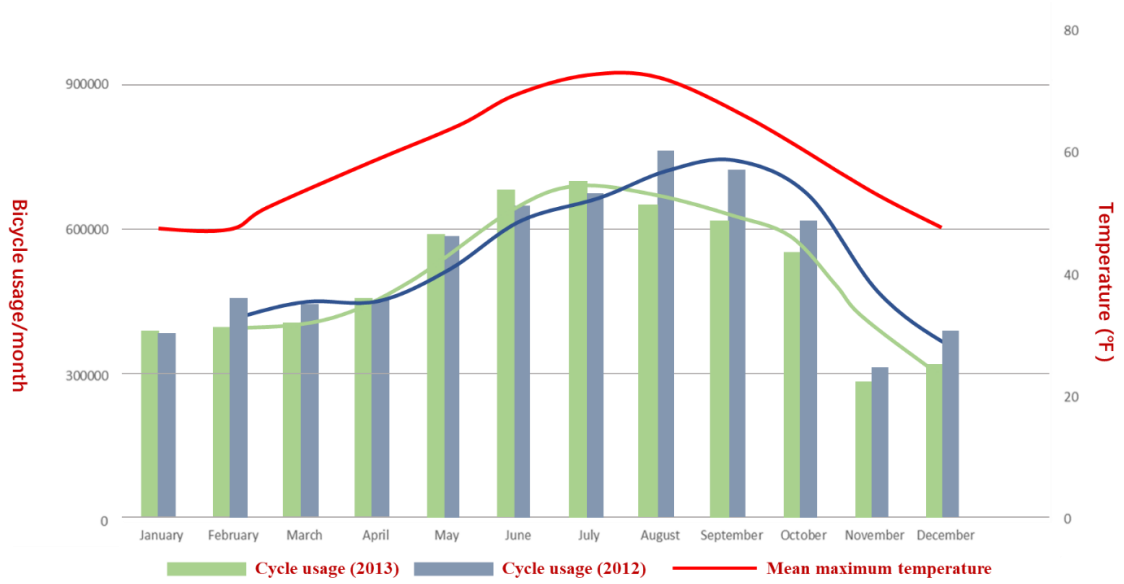


Figure 6.4 Monthly bicycle use frequency and daily maximum temperature in London

6.4.2 Road infrastructure

For the road network characteristic, road density is found to be positively associated with bicycle crash frequency both in overall and separate models. For instance, the number of bicycle crashes would be increased by 28% when road density is increased by 100%. This could be attributed to the increase in potential interactions between bicycles and motor vehicles (Wong et al., 2007; Li et al., 2018). The presence of Cycle Superhighway has a favorable effect on bicycle crash frequency in the cold season. It could be attributed to the increase in driver awareness and safety perception when travelling through the Cycle Superhighway. As illustrated in **Figures 6.3(a), 6.3(b)** and **6.3(c)**, coloured asphalt pavements are applied for the cycle track and cycle lane along the Cycle Superhighway. However, no evidence can be established for significant correlation between bicycle crash frequency and presence of Cycle Superhighway for the overall and warm season models. It could be because most Cycle Superhighways are non-segregated (see **Figure 6.3(c)**). Also, casual bicyclists, who are usually less skilful, are expected to ride more in the warm seasons (Sze et al., 2011). The favourable effect of coloured pavement on drivers' safety awareness could be offset. This finding implies that better designs of Cycle

Superhighway, such as physical separation between the bicycle lane and (motor) traffic lane, would be essential to enhance bicycle safety.

6.4.3 Land use

Results indicate that proportion of commercial area is positively associated with bicycle crash frequency. This could be attributed to the frequent pick-up and drop-off activities on the roadsides in the commercial area. Therefore, potential bicycle crash risk could increase (Wong et al., 2007). Additionally, increase in the proportion of green area is associated with the increase in bicycle crash frequency, particularly in the warm season. It could be attributed to the access to green area for recreational purpose of casual bicyclists in the warm season (Chen, 2015; Guo et al., 2018a). Moreover, commercial (commercial, office and shopping), green area (public park and plantation) and utility (highway) could constitute 80% of the study area (Lubbock, 1963). It is important to enhance the safety level in these areas where pedestrian, bicycle and vehicular traffic flows are high. As shown in **Figure 6.5**, bicycle crashes are widely distributed in the study area. Bicycle facilities, including segregated bicycle tracks, designed crossing, and bicycle signals, could have been introduced at the hot spots of bicycle crashes, especially in the commercial and green areas. Yet, it is worth exploring the effects of weather conditions, e.g. rain, strong wind, fog and snow, etc., on the frequency and severity of bicycle crash, when comprehensive real-time weather data are available (Wen et al., 2019; Zhai et al., 2019a; Xing et al., 2019).

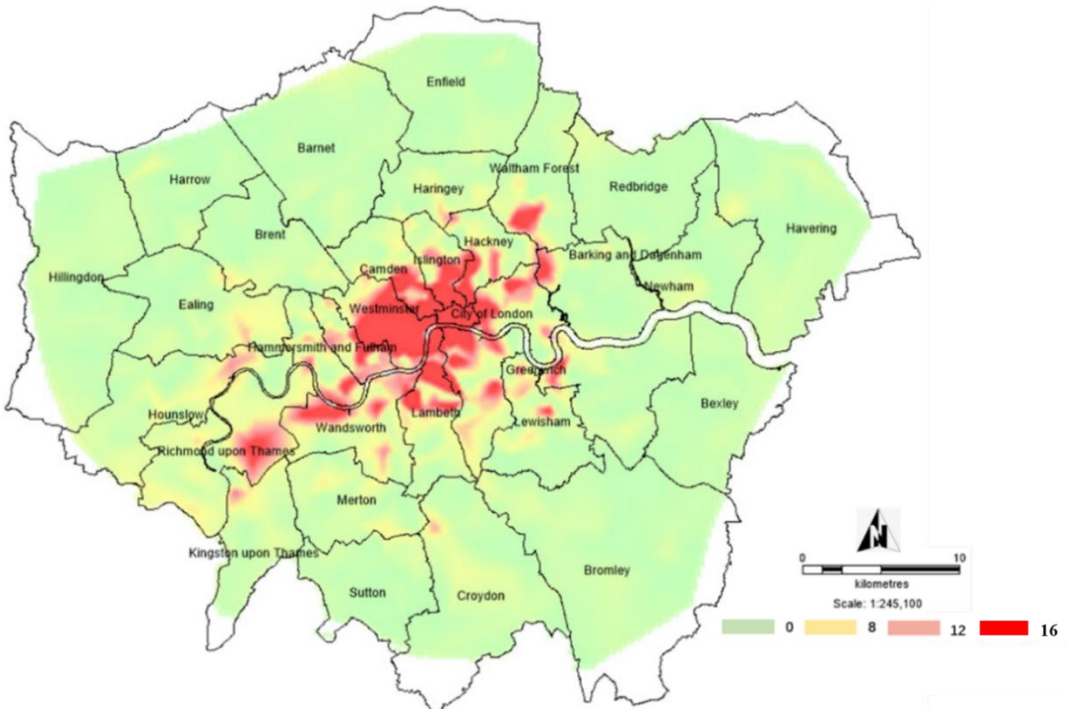


Figure 6.5 Distribution of bicycle crash by MSOA in the analysis period

6.4.4 Demographic and socioeconomics

We also consider the safety effects of population demographic, socioeconomics and household attributes (Mindell et al., 2012; Wei et al., 2013; Ghekiere et al., 2014). Overall, population is positively associated with bicycle crash frequency. However, there is heterogeneity in the population effect based on demographic and socioeconomic characteristics. For instance, increase in the proportion of elderly (age above 64) is associated with the increase in bicycle crash frequency. This could be attributed to the degradation of cognitive performance and impaired mobility of the elderly. Then, the crash likelihood might increase (Palamara and Broughton, 2013). Also, the increase in male proportion is associated with the increase in bicycle crash frequency. This might be because male commuters are generally more aggressive and tend to commit convicted travel behaviour (Guo et al., 2018b). Furthermore, the proportion of white race, which constituted over 75% of the overall population in Greater London, is positively associated with bicycle crashes. Again, there are heterogeneities for the effects of male and race on bicycle crash incidence. This implies the variations in safety perception and behaviours

among male and people of the same race. It can be attributed to the differences in education level, cultural background and family influences, which are not captured in the prediction model, among the people in the same group. For example, higher education people are more risk averse and have a lower tendency to violate traffic rules (Sami et al., 2013; Hung et al., 2011). These findings are indicative to the targeted safety education and promotion strategies that can enhance the safety perception of vulnerable road user groups (TfL, 2018). For the household attribute, results indicate that increase in medium household income by 100% is correlated to the increase in bicycle crash by 19%. Yet, current results only indicate the correlation between bicycle crash frequency and characteristics of residents. It is worth exploring the relationship between population demographic & socioeconomics, bicyclist behaviour and potential crash risk when comprehensive information on bicyclists' safety perception is available in the future survey.

6.5 Concluding remarks

This study examines the relationship between possible risk factors and bicycle crash frequency at the zonal level, using the population census, land use, traffic, bicycle use and crash data of Greater London in 2012-2013. Random parameter negative binomial regression approach is adopted. Crash exposures are estimated based on the frequency and duration of usage of a public bicycle rental system in London.

Results indicate that the model using the duration of bicycle use as the exposure measure is superior to that using the frequency of bicycle use or population. It can be expected as the duration of bicycle use is a better proxy to infer the potential interactions between bicycles and motor vehicles on the roads. Additionally, road density, bicycle facilities, land use, demographic, socioeconomics and household attributes are found to be associated with bicycle crash incidence. It is indicative to the development of infrastructure, traffic management and enforcement strategies that can mitigate the hazards to bicyclists on roads. In particular, the London Cycle Superhighway network, which has favorable effect on bicycle safety, could have been extended. Also, better traffic management and control measures can be implemented to mitigate the risk of

bicyclists in commercial areas, where roadside pick-up & drop-off activities and interactions between bicycles and motor vehicles are frequent. Furthermore, separate bicycle crash prediction models are developed for different seasons. It is believed that the characteristics and travel behaviour of bicyclists are different across different time periods. Casual bicyclists could ride more frequently for recreation purpose in the warm season. That is why proportion of green area is positively associated with bicycle crash frequency in the warm season only. This is indicative to the effective education and promotion strategies that can enhance the safety perception and awareness of bicyclists, especially those of vulnerable groups. Yet, it is worth exploring the contributory factors to the safety perception and, therefore, bicyclists' behaviour and crash risk. Moreover, the bicycle exposure adopted in this study is limited to the usage data of a public bicycle rental system. However, the system only records the origin and destination (i.e. docking station) of a bicycle trip. It is worth exploring to use bicycle travel distance as a proxy of bicycle crash exposure when comprehensive and extensive bicycle counts are available in the future study.

Chapter 7 Effect of built environment and population characteristics on bicycle safety

7.1 Introduction

Bicycle safety has received more and more attention in recent years. Studies have been conducted to identify the possible factors including built environment and bicycle facilities (Guo et al., 2018b; Wei and Lovegrove., 2013; Chen et al.,2016), population and household characteristics (Ghekiere et al., 2014; Vanparijs et al., 2015; Guo et al., 2018a), land use (Chen, 2015) and traffic attributes (Wei and Lovegrove., 2013) that may affect the bicycle safety. To better quantify the potential of bicycle crash involvement and interpret the risk of different entities, it is necessary to measure the crash exposure. In previous studies, bicycle exposures adopted were bicycle flow counts, bicycle trips (Miranda-Moreno et al., 2011), bicycle time travelled (BTT), and bicycle distance travelled (BDT) (Mindell et al., 2012; Poulos et al., 2015) which were measured using retrospective and prospective surveys. Regardless of the sampling framework and survey design, data may be subject to recall and selection biases. In addition, an extensive household travel survey can be expensive and time-consuming. In chapter 6, the transaction records of the London public bicycle rental system were used to estimate the bicycle crash exposure. Although this system covered most bicycle trips in London, exposure measures were limited to bicycle trips and BTT (Ding et al., 2020).

In London, two Cycle Superhighways were introduced in 2010. They provided faster, safer, and more direct routes for bicyclists. The Cycle Superhighways are completely separated from the trafficable roads and footpaths. In addition, segregated crossings are provided at the intersections (Rayaprolu et al., 2020). The minimum width is 4 meters for a bi-directional Cycle Superhighway (European Cyclists' Federation, 2014). Currently, there are six Cycle Superhighways in London (Li et al., 2018). As illustrated in **Figure 7.1**, the total road length in London is 6,139 km. Cycle lanes (known as 'cycleways') are present on 8.1% (i.e. 496 km) of the roads. Overall, the total length of Cycle Superhighways in London is 77 km. Since the bicyclists do not only consider the path

distance, but also the perceived safety and level-of-service when choosing the routes, it is expected that one would prefer the Cycle Superhighways to the traditional cycleway. The roads that have no cycle lane are expected to be the least preferred.

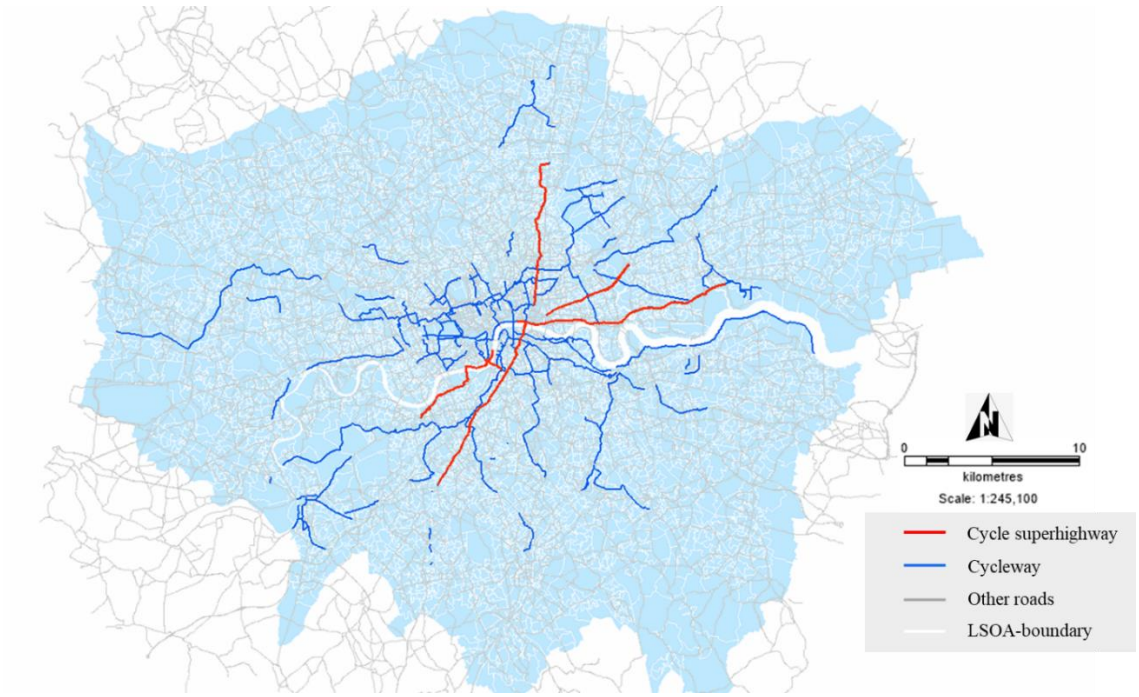


Figure 7.1 Illustration of London road network

In this study, the bicycle routing will be modelled, and the BDT will be estimated based on the origin and destination data of each trip of the London public bicycle rental system. Unlike vehicle drivers, bicyclists generally consider multiple objectives, including travel time and safety, when choosing the route (Ehrgott et al., 2012). Two path analysis models: (a) the simple shortest path model (SPM) that incorporates the effect of path distance only and (b) the weighted shortest path model (WSPM) that incorporates the effects of path distance and perceived safety level, in the route choices are proposed in this study. Then, the negative binomial regression models will be applied to assess the performances of the proposed bicycle path analysis models. Moreover, the associations between bicycle crashes, various exposure measures (bicycle trips, BDT and BTT) and potential influencing factors will be estimated. Findings of this study would indicate the suitability of different bicycle exposure measures. Also, it can improve the understanding on the role of exposure in the bicycle safety analysis.

This chapter is organized as follows. Section 7.2 and Section 7.3 describes the bicycle path analysis and data collection. Analysis results are then given in Section 7.4. Finally, the policy implications are discussed in Section 7.5 and the concluding remarks are given in Section 7.6.

7.2 Bicycle path analysis

In this study, the bicycle transaction records obtained from the London public bicycle rental system (i.e., LCH) are used to estimate the BDT. The dataset records the start time, end time, origin and destination of each bicycle trip. Then, the path of each trip would be determined using the SPM method. Considering the preferences of bicyclists to different bicycle infrastructures, the WSPM is also proposed to model the bicycle path. The model formulations of SPM and WSPM are given as follows.

7.2.1 Simple shortest path model (SPM)

In this model, the shortest path is determined using the Dijkstra's algorithm, assuming that a bicyclist would consider the path distance only in the route choice decision (Deng et at., 2012; Wang, 2012; Sedeño-noda and Colebrook, 2019; Liu and Chen, 2010). The key steps are given as follows.

Step 1: Let V denote the set of vertices of the road network in the algorithm. Denote C_{ij} as the weight that is assigned to the arc connecting V_i and V_j given by

$$C_{ij} = \begin{cases} \infty, \text{if no path between } V_i \text{ and } V_j \\ d_i, \text{otherwise} \end{cases} \tag{7.1}$$

Where d_i denotes the distance of the shortest path originated from the vertex i, and is given by

$$d_i = L_{ij} \tag{7.2}$$

Where L_{ij} is the connection distance between V_i and V_j.

Step 2: Let V_s be the source vertex which is labelled. Estimate the distance between V_s and other unlabelled vertices one by one, then an end vertex V_p will be identified when

$$d_p = min\{d_i | V_p \in V - S\} \tag{7.3}$$

Where d_p is the distance of the shortest path from the source vertex to the end vertex, S is the set of labelled vertices of the shortest path, and (V-S) refers to all unlabelled vertices that are not 'visited' yet.

Step 3: When $V_p = V_t$, then d_p is the distance of the shortest path from V_s to the end point V_t, and the searching process can be stopped. Otherwise, assess another end point by,

$$d_i = min\{d_i, d_k + l_{kj}\}, V_p \in V - S, V_k \in S \tag{7.4}$$

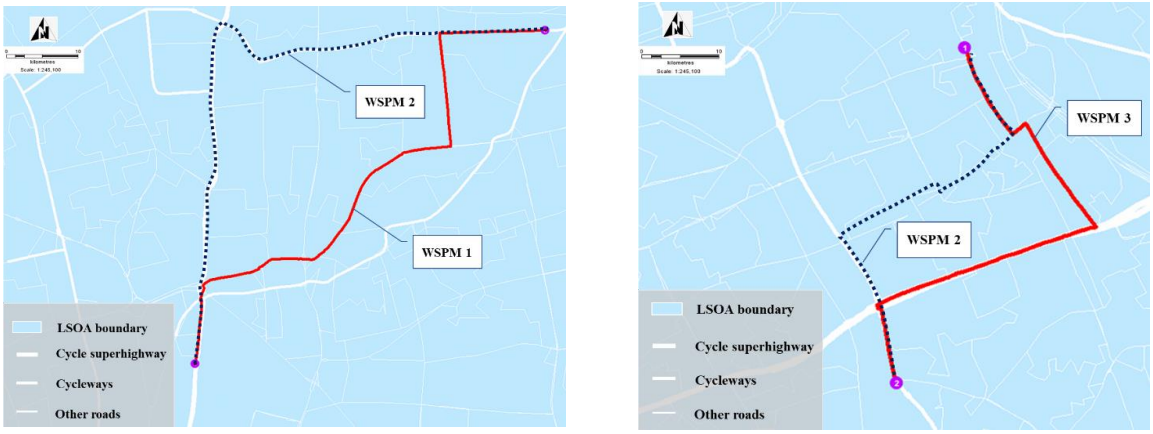
Step 4: Repeat step 2 and step 3 until $V_p = V_t$.

7.2.2 Weighted shortest path model (WSPM)

As mentioned above, not only the path distance, but also the perceived safety and level of service are considered in the bicycle route choice. In this study, it is assumed that Cycle Superhighway and cycleway are preferred by the bicyclists. Therefore, a weighted shortest path method (WSPM) is proposed, with which different weights are assigned to Cycle Superhighways, cycleway and other roads (that have no cycle lanes) respectively in the algorithm. As illustrated in **Table 7.1**, three different scenarios of weight allocation are considered: (i) WSPM1: Cycle Superhighway is preferred, and there is no difference between the cycleway and other roads; (ii) WSPM2: Cycle Superhighway is the most preferred, followed by the cycleway, and other roads are the least preferred; (iii) WSPM3: Similar to WSPM2, just the differences in the weights are magnified. **Figure 7.2** shows an example of the bicycle path choice based on different WSPMs.

Table 7.1 Setting of different weighted shortest path model

Model	Road type				
	(A) Cycle Superhighway		(B) Cycleway		(C) Other roads
WSPM1	W_A	>	W_B	=	W_C
WSPM2	W_A	>	W_B	>	W_C
WSPM3	W_A	>>	W_B	>>	W_C



(a) WSPM1 versus WSPM2

(b) WSPM2 versus WSPM3

Figure 7.2 Bicycle path choices using different WSPM

7.3 Data

The area of interest of this study is the same as that in chapter 6 (see **Figure 6.1**). The observation unit of bicycle crash analysis is the Lower Super Output Area (LSOA) in London. A total of 289 LSOAs are selected. Specifically, bicycle crash data between 2015 and 2016 are obtained from the Greater London Authority (GLA) collision data extract. Land use, population characteristics, traffic flow, and road infrastructure are explanatory variables considered. These can be obtained from the Office for National Statistics (ONS) census dataset and Department for Transport (DfT) dataset.

Furthermore, to examine bicyclist travel behaviour (i.e. trips and time), the transaction records of the London Public Bicycle Rental system - LCH – in the period between 2015

and 2016 are used. The data from multiple sources are mapped to the corresponding Lower Super Output Area (LSOA) using the geographical information system (GIS) approach. **Table 7.2** summarizes the characteristics of the LSOAs.

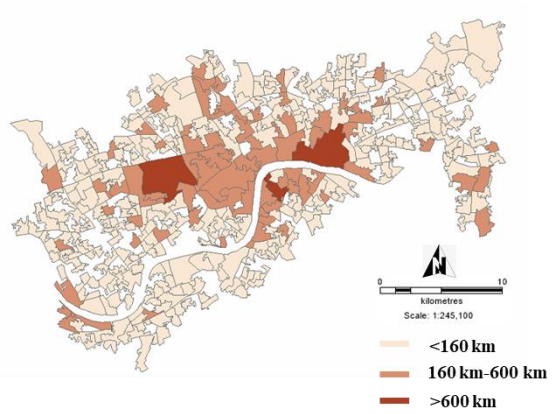
Table 7.2 Summary statistics of the sample

Category	Factor	Attribute	Mean	S. D.	Min.	Max.
Outcome	Frequency of bicycle crash		5.13	5.83	1	38
Land use	Proportion for residential		0.15	0.07	0.02	0.36
	Proportion for commercial		0.25	0.14	0.01	0.56
	Proportion for green area		0.28	0.16	0.03	0.77
	Proportion for transport facilities		0.32	0.16	0.03	0.77
Population characteristics	Population density (per km^2)		14.28	7.36	0.86	39.77
	Population		1,298	464	1,077	3,351
	Gender	Proportion of male	0.52	0.03	0.45	0.65
		Proportion of female	0.48	0.04	0.35	0.55
	Age	Proportion of age above 64	0.11	0.05	0.02	0.3
		Proportion of others	0.89	0.05	0.7	0.98
	IMD (Index of Multiple Deprivation)		24.49	10.46	6.06	53.20
Exposure	Annual BTT (hour)		10,297	13,434	163	14,912
	Annual bicycle trips		28,035	28,748	544	236,240
	VKT		45,849	78,965	51	712,666

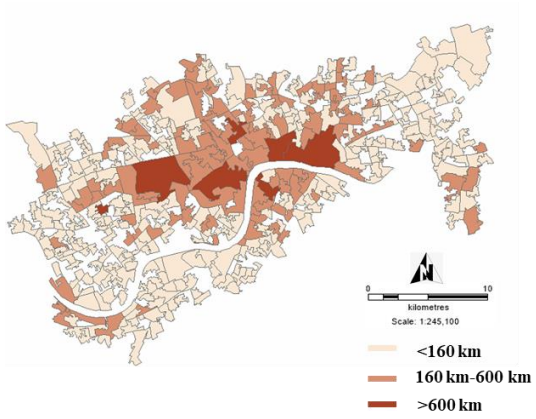
7.4 Estimation results

7.4.1 Estimation of BDTs

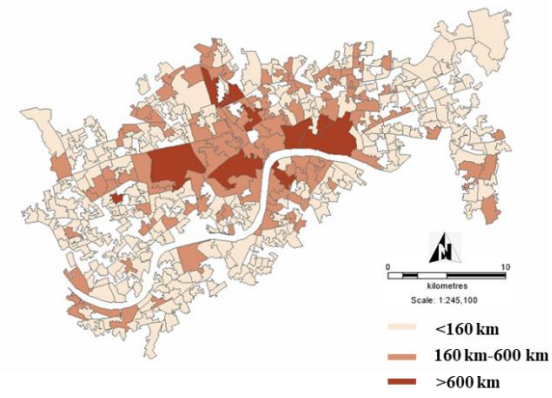
Figure 7.3 and **Table 7.3** illustrate the results of BDT estimations using SPM and WSPMs, respectively. As depicted in **Figure 7.3(a)**, the BDTs seem evenly distributed across the whole study area, when the simple shortest path method is used. As expected, when higher weights are assigned to the cycleway (i.e. WSPM2) and Cycle Superhighway (i.e. WSPM3) in the bicycle path choice analysis, the BDTs would concentrate to the areas that have more cycleway (see **Figure 7.3(b)**) and Cycle Superhighway (see **Figure 7.3(c)**). Among the three WPSMs, as shown in **Table 7.3**, the total estimated BDT is the highest (annual average bicycle distance travelled of 159,600 km per unit) for the WPSM3, followed by the WPSM2 (150,100 km per unit) and then the WPSM1 (146,600 km per unit). This could be attributed to the higher operating speeds of Cycle Superhighways and cycleway. Therefore, the total estimated BDT tends to be higher given the same travel time.

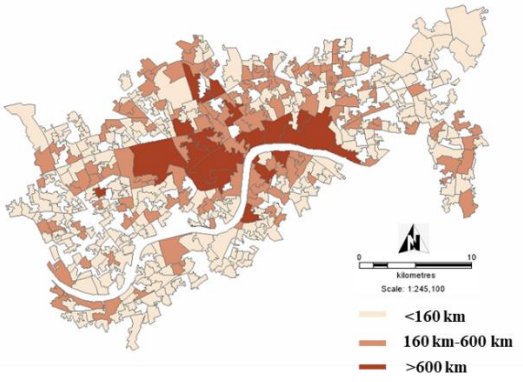


(a) SPM



(b) WSPM1





(c) WSPM2 (d) WSPM3

Figure 7.3 Distributions of BDTs by LSOA

Table 7.3 Estimation results of BDTs by LSOA (10^3 km)

Model	Mean	Standard Deviation	Maximum	Minimum
SPM	133.6	145.3	1,351.4	0.3
WSPM1	146.6	158.6	1,368.6	0.3
WSPM2	150.1	162.3	1,376.7	0.4
WSPM3	159.6	173.4	1,582.3	0.5

7.4.2 Bicycle crash analysis

To eliminate the heteroscedasticity among the variables, variables including population and VKT are logarithmically transformed prior to the parameter estimation (Quddus, 2008). On the other hand, the multi-collinearity test is conducted to assess the correlations between the independent variables. Results indicate that the variance inflation factor (VIF) are less than five for all independent variables. Therefore, all candidate variables are considered appropriate.

(1) BDTs as exposure (SPM versus WSPM)

Since the over-dispersion is prevalent for the data (mean = 5.13 and variance = 33.98), the bicycle crash prediction models, with which the BDTs are used to proxy the bicycle crash exposure, are established using the negative binomial regression model. **Table 7.4** illustrates the model estimation results. As shown in **Table 7.4**, bicycle crash prediction models that incorporate the BDTs estimated by the WSPM are superior to that using the SPM, in accordance with the values of AIC and BIC, regardless of the weights assigned to cycleway and Cycle Superhighway. WSPM2 has the best model fit, with the lowest values of AIC (1268.48) and BIC (1305.46). In addition, differences in AIC and BIC between WSPM2 and SPM are all greater than 10 (Fabozzi et al., 2014). This implies that it is appropriate to assign a higher weight to Cycle Superhighway in bicycle route choice

and safety analysis. Also, the over-dispersion parameter (0.172) of WSPM2 is significant at the 5% level. Therefore, it is appropriate to adopt the NB regression model. The marginal effects of BDTs on the bicycle crash frequency are also estimated (see **Table 7.5**). As shown in **Table 7.5**, bicycle crash frequency is more sensitive to the BDTs that are estimated using the WSPM as compared to that using the SPM. 1% increase in BDT is correlated with 0.47-0.70% increase in bicycle crash frequency when the WSPM is used. On the other hand, 1% increase in BDT is correlated with 0.11% increase in bicycle crash frequency when the SPM is used.

Table 7.4 Results of bicycle crash prediction models using BDTs as exposure

Category	Factor	WSPM1		WSPM2		WSPM3		SPM	
		Coefficient	t-stat	Coefficient	t-stat	Coefficient	t-stat	Coefficient	t-stat
Constant		-11.04**	-6.31	-10.98**	-6.45	-11.27**	-6.45	-10.79**	-6.13
Land use	Proportion of commercial area	2.41**	4.52	2.18**	4.17	2.37**	4.52	2.60**	4.89
	Proportion of green area	1.16**	2.80	1.05**	2.60	1.15**	2.81	1.24**	3.23
Population characteristics	log (population)	1.51**	3.12	1.38**	2.90	1.53**	3.19	1.58**	3.23
	Proportion of age above 64	IS	--	IS	--	IS	--	IS	--
	Proportion of male	5.77**	5.11	5.49**	4.98	5.69**	5.06	5.95**	5.22
	IMD	IS	--	IS	--	IS	--	IS	--
Exposure	log (VKT)	0.63**	7.01	0.59**	6.87	0.62**	7.21	0.64**	7.21
	BDT (km)	0.08*	1.92	0.14**	3.70	0.10**	2.48	0.02	0.53
Over-dispersion parameter	alpha	0.189		0.172		0.186		0.197	
Goodness-of-fit	AIC	1279.07		1268.48		1276.65		1282.45	
	BIC	1315.05		1305.46		1312.64		1318.43	

** and ** denote statistical significance at the 5% and 1% levels respectively; IS denotes insignificant.*

Table 7.5 Marginal effects of BDTs on bicycle crash frequency

Model	Elasticity	p-value
SPM	0.11	0.596
WSPM1	0.47	0.035
WSPM2	0.70	0.000
WSPM3	0.49	0.014

(2) Bicycle trip, BTT and BDT as exposures

Three bicycle crash prediction models incorporating bicycle trips, BTT and BDT respectively, as exposure are also developed using the negative binomial regression model (see **Table 7.6**). As shown in **Table 7.6**, Model 3, which incorporates BDT as the exposure, has the best model fit with the lowest values of AIC and BIC. Again, there are remarkable differences in AIC and BIC between Model 3 and Model 2 (both greater than ten). This indicates that the model using BDT as the exposure is preferred. **Table 7.6** also shows that factors including road density, green area, commercial area, population and gender significantly affect the bicycle crash frequency at the 1% level. Such finding is consistent with that of many previous studies (Ding et al., 2020; Guo et al., 2018a; Chen, 2015; Wei and Lovegrove, 2013). Specifically, increases in the proportion of green area (1.05), proportion of commercial area (2.18), log (population) (1.38), proportion of male (5.49), log (VKT) (0.59) are associated with the increase in bicycle crash frequency. However, effects of IMD and proportion of elderly on bicycle crash frequency are insignificant.

Table 7.6 Results of bicycle crash prediction models with different exposures

Category	Factor	Model 1		Model 2		Model 3	
		coefficient	t-stat	coefficient	t-stat	coefficient	t-stat
Constant		-10.80**	-6.28	-10.78**	-6.19	-10.98**	-6.45
Land use	Proportion of commercial area	2.32**	4.41	2.39**	4.48	2.18**	4.17
	Proportion of green area	1.16**	2.86	1.16**	2.82	1.05**	2.60
Population characteristics	log (population)	1.46**	3.05	1.48**	3.05	1.38**	2.90
	Proportion of age above 64	IS	--	IS	--	IS	--
	Proportion of male	5.55**	4.95	5.61**	4.92	5.49**	4.98
	IMD	IS	--	IS	--	IS	--
Exposure	log (VKT)	0.59**	6.76	0.61**	7.01	0.59**	6.87
	Bicycle trips			0.10**	2.06		
	BTT (hour)	0.13**	3.10				
	BDT (km)					0.14**	3.70
Over-dispersion parameter	alpha	0.182		0.189		0.172	
Goodness-of-fit	AIC	1274.49		1278.56		1268.48	
	BIC	1309.29		1315.52		1305.46	

*** denotes statistical significance at the 1% level; IS denotes insignificant.*

Again, we have estimated the marginal effects of different exposures on bicycle crashes (see **Table 7.7**). As shown in **Table 7.7**, bicycle crash frequency is more sensitive to the BDT (WSPM2), as compared to bicycle trips and BTT. 1% increase in BDT is associated with 0.70% increase in bicycle crash frequency. On the other hand, 1% increase in bicycle trips and BTT is associated with 0.53% and 0.66% increases in bicycle crash frequencies, respectively.

Table 7.7 Parameter estimates for the effects of exposures on bicycle crash frequency

Exposure	Elasticity	p-value
BTT	0.66	0.002
Bicycle trips	0.53	0.025
BDT (WSPM2)	0.70	0.000

7.5 Discussions

In previous studies, it is rare that bicycle crash exposure is incorporated into the bicycle crash prediction model, limited to the reliable bicycle count data. Taking the advantage of the availability of bicycle trip data obtained from the public bicycle rental system, we adopt various path analysis approaches to estimate the BDT as bicycle crash exposure.

7.5.1 SPM versus WSPM in estimating BDT

For the estimation of BDT, results indicate that the WSPM is superior to the SPM. Such result is reasonable since the SPM assumes that the bicyclists only consider path distance when making route choice decisions. In contrast, the WSPM assigns different weights to different bicycle facilities. For example, higher weights are assigned to the cycleway and Cycle Superhighway, considering the fact that bicyclists would consider the connectivity, directness, environmental quality and safety when planning the travel routes (Ehrgott et al., 2012; Broach et al., 2012; Hopkinson and Wardman, 1996).

Among the WSPMs, WSPM1 has the worst model performance with the highest values of AIC and BIC. It is because such assignment approach is contradicting with

conventional wisdom that the perceived safety level of the cycleway is higher than that of the roads that have no cycle lane. Indeed, the revealed safety level of the former is 28% higher than that of the latter. Additionally, many studies also indicated that the bicyclists are more willing to ride on the cycleway (Lusk et al., 2011, Broach et al., 2012; Winters and Teschke, 2010). Nevertheless, the bicycle crash frequency models that incorporate the BDT based on WSPM2 (WA>WB>WC) is superior to that based on WSPM3 (WA>>WB>>WC). The latter hypothesizes that preferences toward cycleway and Cycle Superhighways are more substantial. It implies that the bicyclists would give up the safety and level of service by riding on the roads with no cycle lane only if the time saving and/or the reduction in total travel distance was considerable. However, such speculation might be controversial.

Indeed, several studies indicate that there is no noticeable difference in traffic safety among cycleway, Cycle Superhighways and other roads with no cycle lane (Li et al., 2017). It could be because of the heterogeneity in the preference among the bicyclists. For example, even the occasional bicyclists generally prefer the cycleway and Cycle Superhighways, the commuting cyclists may have some other considerations (i.e. route directness and attractiveness) when making the route choice (Ehrgott et al., 2012; Howard and Burns, 2001). Moreover, studies also show that the cycleway is not always considered more desirable than a wider arterial road for experienced bicyclists (Taylor and Mahmassani, 1996; Heinen et al., 2010). Furthermore, factors like gender can also affect safety perception and bicycle route choice (Sener et al., 2009; Stinson and Bhat, 2003). It is, therefore, worth exploring the effects of individual characteristics and trip purpose on the association between route choice and road attributes using the bicyclist survey in future studies.

7.5.2 Bicycle crash exposures

We also assess the use of bicycle trips, BTT, and BDT as exposures in the bicycle crash analysis. Results indicate that bicycle crash frequency model using the BDT as the exposure provides the best model fit. It is because trip distance is more sensitive to the interactions between bicycle and other road users, and therefore potential traffic conflicts,

as compared to trip frequency (Pei et al., 2012). Indeed, there is no noticeable difference in the elasticities between BTT and BDT (see **Table 7.7**).

In this study, factors including land use, population characteristics and traffic conditions that affect the bicycle crash frequency at zonal level are considered. Results show that the proportion of commercial area (2.18) and green area (1.05) are positively associated with bicycle crashes. This can be attributed to the frequent pick-up and drop-off activities at the roadsides in the commercial area (Ding et al., 2020). As for the effect of green area, it is not surprising since considerable portion (31%) of bicyclists in London report that they ride for recreation purposes (TfL., 2015). In addition, log (VKT) (0.59) is positively associated with bicycle crashes. It is consistent with the previous studies (Alkahtani et al., 2018), since the interactions between vehicles and bicycles can increase with the traffic volume. Furthermore, the increase in the proportion of male (5.49) is associated with the increase in bicycle crash frequency. This can be attributed to the difference in safety perception and cycling behaviours among different bicyclist groups (Guo et al., 2018b). Nevertheless, the current study is limited to the average effect of built environment on bicycle safety at the macroscopic level (i.e. LSOA). It is worth exploring the moderating effect of geometric design and road environment on the association between bicycle crashes and BTT and BDT, when detailed crash, traffic and environment data at the microscopic level is available in the future study. On the other hand, it is worth noting that crash occurrence is rare. It is often necessary to accumulate more bicycle crashes over a considerable period when evaluating the safety effect of an intervention. To this end, it is possible to evaluate the bicycle safety level using appropriate surrogate safety measures, e.g. conflicts (Sayed et al., 2013; Kassim et al., 2014; Strauss et al., 2017; Guo et al., 2020).

7.6 Concluding remarks

To assess the bicycle crash risk of different entities and better interpret the relationship between bicycle safety and possible risk factors, it is necessary to have reliable exposure measures such as bicycle count number, bicycle trips, BTT, and BDT. Unlike vehicular crash analysis, extensive bicycle counts are often not available. In chapter 6, detailed

transaction data of the London public bicycle rental system was available to estimate the bicycle crash exposure (i.e., BTT and bicycle trips) at the zonal level, using the data on bicycle trip, origin and destination (Ding et al., 2020). In this study, we revisit the topic of bicycle crash exposure by estimating the BDT of each trip using the shortest path method. Considering the effects of safety perception, attitudes and preferences to different bicycle infrastructures on bicycle route choice, a modified path analysis approach – weighted shortest path method (WSPM) – is proposed.

Results indicate that the bicycle crash frequency model that adopts BDT as the exposure is superior, compared to that using bicycle trips and BTT as the exposures. For instance, safety effects of land use, population characteristics and traffic conditions on bicycle crash frequency are identified. In addition, the bicycle crash frequency models that adopt BDTs estimated using the WSPM apparently have better model fit, compared to that using the SPM. For instance, when the differences between the preferences toward Cycle Superhighway, cycleway and other roads are moderate, the best model fit can be attained. This justify that bicyclists do not only consider path distance, but also other factors such as level of service and perceived safety when choosing the routes (Ehrgott et al., 2012; Broach et al., 2012). Yet, this study does not consider the bicycle crash severity. In the future study, heterogeneity in the bicycle crash risk by collision type and injury severity would be investigated.

Chapter 8 A multivariate Poisson-lognormal model for the correlation in bicycle safety analysis

8.1 Introduction

Although cycling has benefits to environment and physical health, bicyclists are vulnerable road users. Prior studies have identified the environment, traffic and road user factors that affect the risk of bicycle-related crashes (Ding et al., 2020, 2021a; Guo et al., 2018a). However, it is rare that difference in their effects on the risk among different bicycle crash types to be investigated. Indeed, effects of possible factors on bicycle crash frequencies can also vary with collision type (Guo et al., 2018a; Park and Lord, 2007). For example, presence of bicycle infrastructure is more sensitive to bicycle-only crashes, as compared to bicycle-vehicle crashes (De Rome et al., 2014; Teschke et al., 2014; Beck et al., 2016). It is necessary to account for multivariate correlation in the bicycle crash frequency models.

In addition, road network characteristics can affect travel behaviour in terms of trip frequency, path choice, travel time and travel distances, and therefore crash exposure (Zhang et al., 2015; Pei et al., 2016; Quddus, 2008). Hence, it is necessary to consider road network characteristics in the bicycle crash frequency model. Although previous studies have examined the effects of road network characteristics like street connectivity, number of intersections, length of bicycle lanes, and road classes on bicycle crash frequencies (Yasmin and Eluru, 2016; Osama and Sayed, 2017; Kamel and Sayed, 2021). It is rare that the effect of road network accessibility on bicycle crash risk is considered (Marshall and Garrick, 2010; Wei and Lovegrove, 2012; Guo et al., 2018a).

This study aims to examine the effects of possible factors on the frequencies of different bicycle crash types, i.e., bicycle-vehicle and bicycle-bicycle crashes. Crash data from middle layer super output areas (MSOA) of London in 2018 and 2019 are used. Then, multivariate Poisson-lognormal regression approach is applied to measure the association, with which multivariate correlation between bicycle-vehicle and bicycle-

bicycle crashes is accommodated. Furthermore, effects of road network characteristics in terms of connectivity and accessibility are also considered. Findings should shed light on the design of road infrastructure, and implementation of traffic management and control measures that can reduce the bicycle crash risk.

Reminder of this chapter is organized as follows. Section 8.2 describes the method of data collection. Parameter estimation results and discussions are given in Section 8.3 and Section 8.4, respectively. Section 8.5 provide the study recommendation and concluding remarks.

8.2 Data

The area of interest of this study is the same as that in chapter 6 (see **Figure 6.1**). Observation unit is Middle Layer Super Output Area (MSOA) in London. In this study, population socio-demographics, land use, road network, traffic flow, and bicycle crash data of London in 2018 and 2019 are used. In total, there are four types of bicycle crashes: bicycle-vehicle, bicycle-bicycle, bicycle-pedestrian, and single bicycle crashes. However, counts of bicycle-pedestrian and single bicycle crashes are extremely low (Myhrmann et al., 2021; Olesen et al., 2021). Therefore, only bicycle-vehicle and bicycle-bicycle crashes are considered. Overall, 3,743 bicycle-related crashes (3,622 bicycle-vehicle and 121 bicycle-bicycle crashes) are considered.

In this study, effect of road network topology on bicycle crash frequency is also considered. For example, morphological parameters, including connectivity and accessibility of MSOAs, are estimated using space syntax theory (Hillier and Hanson, 1984; Hillier, 1996). Specifically, connectivity refers to the number of direct neighbouring roads that intersect with a given axial road. High connectivity indicates more possibilities for the roads to intersect with each other in the network. On the other hand, accessibility can be estimated by measuring the degree of integration of the network. A poorly integrated point is a location that requires more steps (spaces) to reach from a starting point. Integration is proportional to the reciprocal of mean depth of the network. In space syntax, depth refers to the topological distance between points (nodes).

In general, depth can be represented by global depth (D), local depth (LD), and mean depth (MD) which are given by,

$$D_r = \sum_{r=1,s=1}^{t} d_{rs} \tag{8.1}$$

$$LD_r = \sum_{r=1,s=1,d_{rs}\leq 3}^{t} d_{rs} \tag{8.2}$$

$$MD_r = \frac{D_r}{t-1} \tag{8.3}$$

Where d_{rs} refers to the shortest topological distance between node *r* and *s*, and t refers to the total number of nodes respectively.

Then, accessibility (global integration) of the network can be estimated by,

$$I_r = \frac{t-2}{2(MD_r-1)} \tag{8.4}$$

Spatial distribution of average connectivity and accessibility of the study area are shown in **Figure 8.1** and **Figure 8.2** respectively. In addition, **Table 8.1** summarizes the data used.

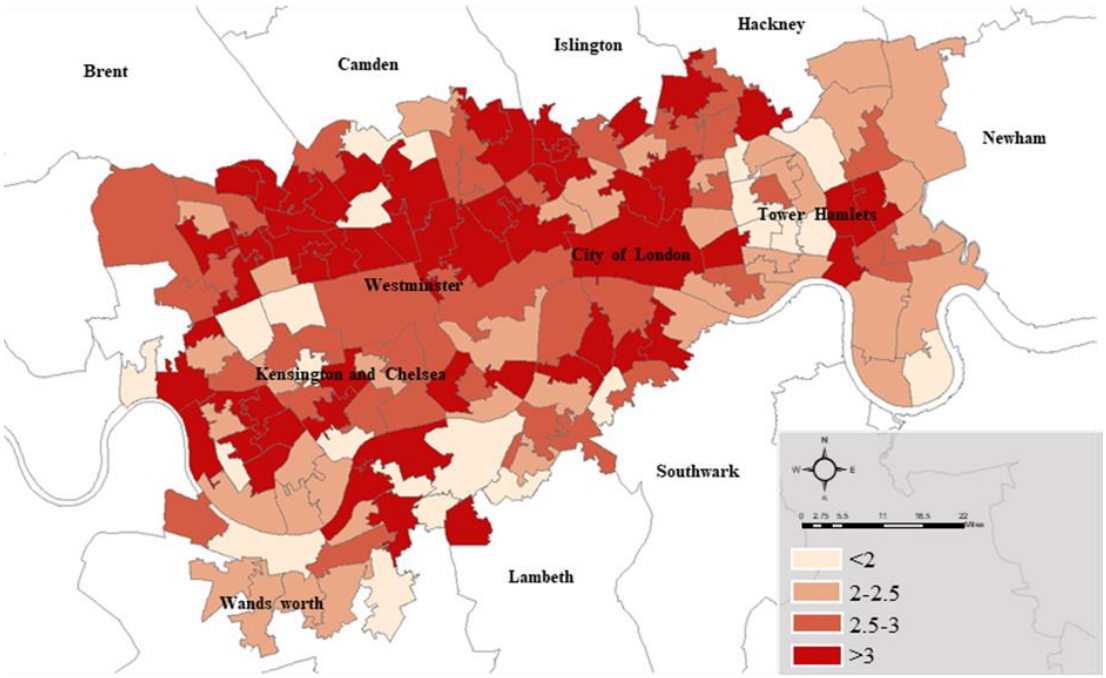


Figure 8.1 Spatial distribution of average connectivity

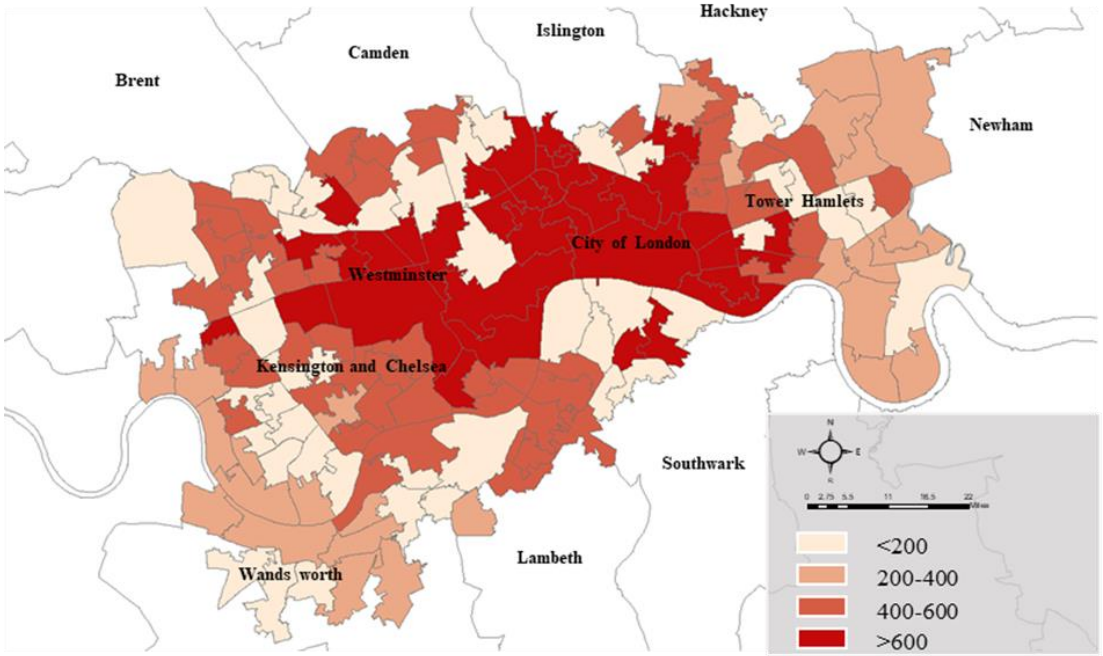


Figure 8.2 Spatial distribution of average accessibility

Table 8.1 Summary statistics of the sample

Category	Factor	Attribute	Mean	S.D.	Min.	Max.
Bicycle-related crash	Bicycle-vehicle crash	Number of bicycle-vehicle crash per year	14.15	14.59	0	127
	Bicycle-bicycle crash	Number of bicycle-bicycle crash per year	0.47	0.93	0	6
Exposure	Population density	Population per square kilometre	64.32	54.52	3.22	280.10
	Traffic flow	Annual average daily traffic	22,023	15,088	1687	96,825
Population socio-demographics	Gender	Proportion of male	0.51	0.02	0.46	0.58
		Proportion of female	0.49	0.02	0.42	0.54
	Age	Proportion of people above age 64	0.10	0.04	0.03	0.24
		Proportion of people below age 16	0.28	0.06	0.15	0.47
	Income	Average annual household income (€)	56,152	8,640	39,800	75,500
Land use	Residential area	Proportion of residential area	0.16	0.07	0.04	0.36

Category	Factor	Attribute	Mean	S.D.	Min.	Max.
	Business and office area	Proportion of business and office area	0.18	0.12	0.01	0.50
	Green area	Proportion of green area	0.34	0.14	0.05	0.74
	Road area	Proportion of road, railway, and footpath area	0.32	0.08	0.15	0.77
Road network characteristics	Road density	Class A road (km per km^2)	3.44	1.94	0.16	9.89
		Class B road (km per km^2)	0.78	0.92	0	5.13
		Minor road (km per km^2)	0.86	0.95	0	4.46
	Connectivity	Average connectivity	2.33	0.47	0.96	3.20
	Accessibility	Average accessibility	621.90	77.09	425.10	812.40
	Bus stop density	Bus stop per square kilometre	0.03	0.02	0	0.14
	Railway station density	Railway station per square kilometre	0.01	0.03	0	0.20
	Intersection density	Intersection per square kilometre	0.39	0.35	0.02	2.54

8.3 Estimation results

In this study, two types of bicycle-related crash, namely bicycle-vehicle crash and bicycle-bicycle crash, will be modelled simultaneously using the proposed multivariate Poisson-lognormal model (i.e., $K = 2$). A multi-collinearity test would be conducted to ensure that all variables considered are independent. For instance, the variance inflation factor (VIF) for all variables should be less than five.

Table 8.2 and **Table 8.3** present the parameter estimation results of multivariate and univariate Poisson-lognormal regression models. As shown in **Table 8.2** and **Table 8.3**, difference in DIC between multivariate Poisson-lognormal model (1910.81) and univariate Poisson-lognormal models (1514.33 + 416.78 = 1931.11) is greater than ten. This justifies that proposed multivariate Poisson-lognormal regression model significantly outperforms the counterpart. In addition, **Table 8.4** present the results of hyper parameter estimation of multivariate Poisson-lognormal regression model. As shown in **Table 8.4**, variance and covariance of errors are all significantly greater than zero. This justifies the existence of over-dispersion. Furthermore, correlation coefficient (ρ_{12}) is significantly greater than zero. This indicates the prevalence of multivariate correlation between the counts of different crash types.

As shown in **Table 8.2**, factors like bicycle usage, traffic volume, household income, residential area, road density, accessibility, and intersection density can affect the bicycle-vehicle crash frequency at the 5% level of significance. On the other hand, factors include bicycle usage, household income, road density, connectivity, accessibility, railway station, and intersection density can affect the bicycle-bicycle crash frequency at the 5% level of significance.

For the crash exposure, bicycle usage is positively associated with both bicycle-vehicle (coefficient = 0.32) and bicycle-bicycle crashes (0.25). Also, traffic volume is negatively associated with bicycle-vehicle crash (-0.22). In contrast, there is no significant effect for traffic volume on bicycle-bicycle crash. For the population characteristics, household income is positively associated with both bicycle-vehicle (0.09) and bicycle-bicycle

crashes (0.38). For the built environment, bicycle-vehicle crash frequency in residential area (-2.42) is lower than that in other areas. However, there is no significant effect for land use on bicycle-bicycle crash.

For the road network characteristics, Class B road density (0.008) and accessibility (0.89) are positively associated with both bicycle-vehicle and bicycle-bicycle crashes. Also, connectivity (1.55) and railway station density (10.95) are positively associated with bicycle-bicycle crash. However, there is no significant effect for connectivity and railway station density on bicycle-vehicle crash. Furthermore, effects of intersection density on bicycle-vehicle crash (0.08) and bicycle-bicycle crash (-0.60) are opposite.

Table 8.2 Results of parameter estimation of multivariate Poisson-lognormal model

Category	Variable	Bicycle-vehicle crash				Bicycle-bicycle crash			
		Mean	SD	95%BCI		Mean	SD	95%BCI	
Intercept	Intercept	IS				IS			
Exposure	ln (Bicycle usage)	0.32	0.03	0.25	0.37	0.25	0.12	0.03	0.44
	ln (Population)	IS				IS			
	ln (AADT)	-0.22	0.17	-0.50	-0.01	IS			
Population socio-demographics	Average annual household income	0.09	0.09	0.02	0.15	0.38	0.69	0.14	0.31
Land use	Residential area	-2.42	0.62	-3.41	-1.37	IS			
Road network characteristics	Class B road density	0.008	0.04	0.01	0.15	0.04	0.13	0.01	0.08
	Connectivity	IS				1.55	0.29	1.09	2.08
	Accessibility	0.89	0.52	0.18	2.04	2.46	1.65	0.18	5.69
	Railway station density	IS				10.95	3.55	5.17	16.82
	Intersection density	0.08	0.14	0.23	0.32	-0.60	0.32	-1.14	-0.09
Goodness-of-fit	$\overline{D}$	1715.69							
	$\widehat{D}$	1520.57							
	P_D	195.12							
	DIC	1910.81							

IS denotes insignificant.

Table 8.3 Results of parameter estimation of Poisson-lognormal model

Category	Variable	Bicycle –vehicle crashes				Bicycle-bicycle crashes			
		Mean	SD	95%BCI		Mean	SD	95%BCI	
Intercept	Intercept	-4.15	1.59	-6.44	-1.51	-11.78	6.38	-22.67	-2.00
Exposure	ln (Bicycle usage)	0.31	0.06	0.22	0.40	0.33	0.13	0.12	0.56
	ln (Population)	0.34	0.17	0.06	0.58	IS			
Land use	Residential area	-2.43	0.60	-3.42	-1.45	IS			
Road network characteristics	Class A road density	0.05	0.02	0.01	0.08	IS			
	Class B road density	0.08	0.04	0.01	0.15	IS			
	Connectivity	IS				1.64	0.29	1.18	2.14
	Accessibility	1.172	0.65	0.02	2.20	3.98	1.8	0.54	6.84
	Railway station density	IS				8.70	3.18	3.45	13.86
	Intersection density	IS				-0.65	0.32	-1.19	-0.15
Goodness-of-fit	$\overline{D}$	1332.7				402.03			
	$\widehat{D}$	1151.1				387.27			
	P_D	181.62				14.76			
	DIC	1514.33				416.78			

IS denotes insignificant.

Table 8.4 Hyper-parameter estimation for multivariate Poisson-lognormal model

Parameter	Mean	SD	95%BCI	
ρ_{12} (ρ_{21})	0.99	0.01	0.98	0.99
${\sigma_{11}}^2$	0.26	0.03	0.21	0.32
${\sigma_{22}}^2$	0.33	0.14	0.12	0.56
${\sigma_{12}}^2({\sigma_{21}}^2)$	0.28	0.07	0.17	0.40

8.4 Discussions

8.4.1 Bicycle crash exposures

As expected, bicycle usage is positively associated with bicycle-vehicle and bicycle-bicycle crashes (Ding et al., 2020, 2021a). In contrast, traffic volume is negatively associated with bicycle-vehicle crash. This might be explained by the compensation theory where drivers adopt more cautious driving behavior to compensate for the increased crash propensity arising from a complex driving environment (Chen et al., 2021). Therefore, risk of possible vehicle-bicycle collision reduces. Indeed, speed limit of 20 miles per hour was imposed in central London (Dumbaugh and Rae, 2009; Guerra et al., 2020). Furthermore, it could be understood that bicycle-bicycle crash should not be sensitive to traffic volume. Hence, there is no significant effect for traffic volume on bicycle-bicycle crash.

8.4.2 Demographic and socioeconomics

Effects of population demographics (i.e., gender and age) and socioeconomics (household income) on bicycle-related crash are investigated. Results indicate that household income is positively associated with both bicycle-vehicle and bicycle-bicycle crashes. Such finding is consistent with that of previous studies (Ding et al., 2020; Guo et al., 2018a). This should be indicative to targeted road safety education and promotional strategies that can increase the safety awareness of bicyclists, and therefore reduce the bicycle crash risk.

8.4.3 Land use

Effects of land use (i.e., residential, commercial and office, industrial, green area, and road area) on bicycle-related crash are examined. Results indicate that proportion of residential area is negatively associated with bicycle-vehicle crash. However, there is no significant effect for residential area on bicycle-bicycle crash. This could be attributed to the implementation of local area traffic management scheme and traffic calming measures in the residential area. This should imply more physical separations between bicycles and vehicles, and reduction in traffic speed. Therefore, risk of bicycle-vehicle crash would be reduced (Zhang et al., 2013).

8.4.4 Road network characteristics

Last but not least, road network characteristics including road density, connectivity, accessibility, transit station, and intersection density are also considered. In particular, Class B road density is positively associated with bicycle-related crashes. In contrast, there is no significant effect for Class A road density on bicycle-related crashes. This could be attributed to the difference in design standards and specifications among different road types. In particular, Class B roads are the minor arterial and collector roads. They usually have lower standard for geometric design like horizontal curves, road width, super elevation, and sight distance. Hence, Class B roads should be more sensitive to bicycle crash, compared to the counterpart. In addition, intersection density is positively associated with bicycle-vehicle crash. This could be because of the higher chance of bicycle-vehicle interactions at the intersections (Wong et al., 2007). In contrast, intersection density is negatively associated with bicycle-bicycle crash. This could be because of the elevated safety awareness of bicyclists when they are approaching the intersections, and thus the risk of bicycle-bicycle conflict would be reduced (Vlakveld et al., 2021). Furthermore, density of railway station is positively associated with bicycle-bicycle crash. This could be attributed to frequent loading and unloading activities near the major public transport hubs (Li et al., 2018, 2019).

For the network topology, average connectivity is positively associated with bicycle-bicycle crash. Also, average accessibility is positively associated with both bicycle-vehicle and bicycle-bicycle crashes. This could be attributed to the bicyclist's route choice behaviour since the well-connected paths are usually more preferred. Therefore, frequent bicycle activities are expected in the well-integrated areas (Quintero et al., 2013). This is particularly true in central London, where both congestion charging scheme and public bicycle rental scheme are imposed (Li et al., 2019; Ding et al., 2021b).

8.5 Concluding remarks

This study aims to investigate the effects of possible factors on bicycle crash frequency, with which possible correlation between different bicycle crash types is accommodated using multivariate Poisson-lognormal approach. Results indicate that proposed multivariate model outperforms conventional univariate model, in term of DIC value. For instance, factors like bicycle usage, household income, road density, and accessibility are positively associated with both bicycle-vehicle and bicycle-bicycle crashes. In contrast, traffic volume and proportion of residential area are negatively associated with bicycle-vehicle crash only. There is no significant effect for traffic volume and land use on bicycle-bicycle crash. Furthermore, connectivity and railway station are positively associated with bicycle-bicycle crash only. There is no significant effect for connectivity and railway station on bicycle-vehicle crash.

Nevertheless, some limitations of this study should be highlighted. For instance, problem of excessive zero observations may exist in bicycle-bicycle crashes. This could result in bias in parameter estimation and poor model fit. Therefore, it is worth investigating for the use of data-driven approaches to resolve the problem of unbalanced crash data (Zhao et al., 2018; Shankar et al., 1997; Ding et al., 2022b). Furthermore, effect of temporal instability on the association could have been considered if the observation period had been extended.

Chapter 9 A deep generative approach for excessive zero observation in safety analysis

9.1 Introduction

Road safety has long been recognized as a major public health and social issue worldwide (Pei et al., 2016; Elamrani et al., 2020). In 2016, there were about 1.35 million road fatalities and over 20 million road injuries round the world. Road crashes are expected to become the fifth leading cause of death by 2030 (WHO, 2018). Crash frequency models are often established to measure the relationship between crash occurrence and possible explanatory factors. Therefore, effective countermeasures can be implemented to mitigate related crash risk and improve overall road safety.

It is well recognized that crashes are rare events. This gives rise to the problem of unbalanced crash and non-crash cases when developing the crash frequency models (Abdel-Aty et al., 2004). For instance, imbalanced data problems may also have existed in the proposed bicycle crash frequency models in the previous chapters. Prior studies indicated that excess zero observations can result in bias in parameter estimation and poor model fit (Miaou, 1994; Shankar et al., 1997). In addition, it can have adverse impact on the identification of crash explanatory factors (Pei et al., 2016; Yu et al., 2020; Cai et al., 2020; Washington et al., 2011).

Although advanced statistical methods and data-driven approaches have developed to model the zero-inflated crash data. These approaches also have deficiencies including sample size, data inconsistency problems, correlations between variables, training stability, robustness and flexibility. In particular, the existed synthetic methods assumed that all variables in the data should be real-valued. They are not capable of handling categorical and nominal data. To this end, a deep generative approach – augmented variational autoencoder – is proposed to generate synthetic crash data for the association measure between crash and possible explanatory factors. This approach is characterized by a factorized generative model and refined objective function. For instance, the

generative model can handle heterogeneous data including real-valued, nominal and ordinal distributions. On the other hand, the refined objective function can control for the random effect by better recognizing both the zero-crash and non-zero crash cases. In this study, comprehensive traffic and crash data of multiple distribution types in Hong Kong between 2014 and 2016 are used. To assess the data generation performance of the proposed augmented variational autoencoder method, a conventional data synthesis technique (synthetic minority oversampling technique-nominal continuous) is also considered. Findings of this study should shed light on both researchers and practitioners for the development of bicycle crash frequency models, with which the problem of excessive zero observations is prevalent when highly disaggregated traffic and crash data by time and space are used.

The reminder of this chapter is organized as follows. Proposed augmented variational autoencoder method is described in Section 9.2. Procedure of data preparation is described in Section 9.3. Then, results and discussions are presented in Section 9.4 and Section 9.5, respectively. Lastly, key findings are summarized in Section 9.6.

9.2 Augmented variational autoencoder

(1) Variational Autoencoder method

Variational Autoencoder method is a deep generative model based on highly-structured homogeneous data generation. It can exploit the correlations between variables and capture the complicated dependencies between samples (Yang et al., 2017; Razavi et al., 2019; Walker et al., 2017). Previous studies suggested that the variational autoencoder method is capable of generation and generalization since the distributions of latent variables are restrained to a specific paradigm. Therefore, a completely new dataset can be generated from the latent space (Boquet et al., 2020).

Let $\mathbf{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ denotes a dataset consisting of N observations, each observation $\boldsymbol{x} \in \mathbb{R}^m$ denotes a m-dimensional vector, with which the value of every element is real number. In this study, a vector represents the values of different traffic and crash attributes

of an entity. To generate a realistic sample, a critical step is to correctly learn the ground-truth distribution of $\boldsymbol{x}$. Let $p(\boldsymbol{x}|\theta)$ denotes the conditional probability of $\boldsymbol{x}$ given θ, then the classical parameter estimation problem can be given by,

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{X}|\theta) \tag{9.1}$$

To estimate the parameters, maximum likelihood estimation and maximum a posterior estimation approaches are commonly used (Lord and Mannering, 2010). However, as $p(\mathbf{X}|\theta)$ is often non-convex, estimation of $\hat{\theta}$ can be hindered. To this end, a latent variable $\boldsymbol{z} \in \mathbb{R}^d$ $(d \ll m)$ proposed by Kingma and Welling (2013) can be applied to resolve the non-convex problem, using variational inference approach. In particular, latent variable $\boldsymbol{z}$ denotes a low-dimensional system, and $\boldsymbol{x}$ can be generated from $\boldsymbol{z}$ in a random process. Therefore, conditional probability of $\boldsymbol{x}$ can be estimated by integrating $p(\boldsymbol{x}|\boldsymbol{z},\theta)$ with respect to a prior distribution of $\boldsymbol{z}$ using the following formulation,

$$p(\boldsymbol{x}|\theta) = \int p(\boldsymbol{x}|\boldsymbol{z},\theta)p(\boldsymbol{z})d\boldsymbol{z} \tag{9.2}$$

However, marginal likelihood of Equation 9.2 is intractable. Computation time for the simulation-based optimization solutions can be considerable. To overcome the problem, as shown in **Figure 9.1**, an ensemble paradigm proposed by Kingma and Welling (2013) can be used to solve the optimization problem. In particular, generative and recognition models are set out to determine the posterior distribution of $p(\boldsymbol{z}|\boldsymbol{x},\theta)$, where (1) generative model $p(\boldsymbol{x}|\boldsymbol{z},\theta)$ can produce the reconstructed sample based on the conditional probability distribution of latent vector $\boldsymbol{z}$ and (2) recognition model $q(\boldsymbol{z}|\boldsymbol{x},\varphi)$ can produce the latent vector $\boldsymbol{z}$ based on the conditional probability distribution of input vector $\boldsymbol{x}$ respectively.

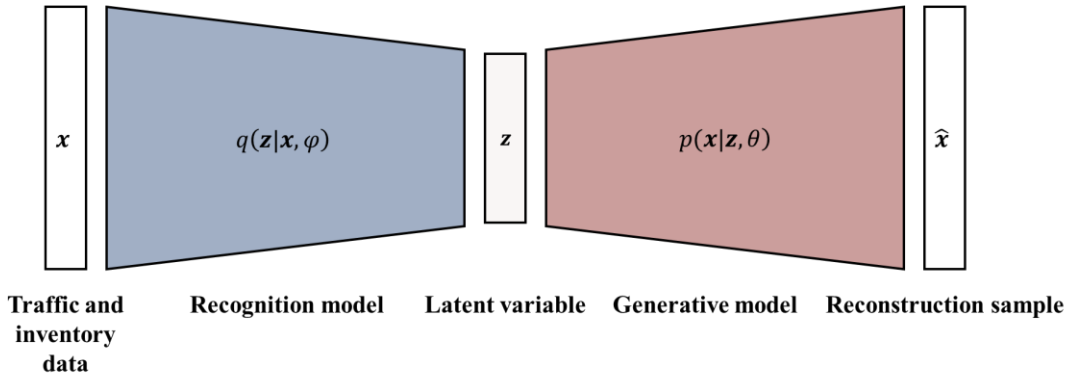


Figure 9.1 Framework of VAE Method

Kullback–Leibler divergence $D_{KL}(\cdot)$ is applied to measure the degree of approximation by the recognition model $q(\boldsymbol{z}|\boldsymbol{x},\varphi)$ to the posterior distribution $p(\boldsymbol{z}|\boldsymbol{x},\theta)$. Kullback–Leibler divergence is specified as,

$$\min_{q(\boldsymbol{z})} D_{KL}(q(\boldsymbol{z})|p(\boldsymbol{z}|\boldsymbol{x})) \tag{9.3}$$

Based on Equation 9.3, a variational lower bound of the likelihood of having a sample can be derived by,

$$\log p(\boldsymbol{x}) \geq -\left\{D_{KL}\left(q(\boldsymbol{z}|\boldsymbol{x})\middle|p(\boldsymbol{z})\right) - \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x})}[\log p(\boldsymbol{x}|\boldsymbol{z})]\right\} \tag{9.4}$$

The estimate given by Equation 9.4 can also be defined as the Evidence Lower Bound. The first term on the right side of Equation is the Kullback–Leibler divergence between the recognition model and posterior distribution, and the second term is the expected value of reconstruction error.

As the ultimate objective is to maximize the marginal likelihood, the objective function defined in Equation 9.3 can be converted into the maximization of the evidence lower bound (i.e., minimization of the magnitude of the evidence lower bound which is negative) with respect to θ and φ using the deep neural networks approach (Kingma and Welling, 2013; Rezende et al., 2014) specified as,

$$\min_{\theta,\varphi} \mathcal{L}(\theta,\varphi) = -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\varphi)}\big(\log p(\boldsymbol{x}|\boldsymbol{z},\theta)\big) + D_{KL}(q(\boldsymbol{z}|\boldsymbol{x},\varphi)||p(\boldsymbol{z})) \tag{9.5}$$

(2) Augmented Variational Autoencoder method

The variational autoencoder method is recognized as a powerful generator for the synthesis of homogeneous data with which all variables are constrained to have the same distribution type (normally multivariate Gaussian distribution) and the parameters of probability density function are optimized using the deep neural networks method (Boquet et al., 2020). However, for a typical crash dataset, variables can be of different types, e.g., continuous, nominal, and ordinal. The above mentioned variational autoencoder method is not capable of processing heterogeneous data. To resolve the problem, the generative model in the variational autoencoder method is modified by factorizing the unified conditional probability density function into several variable-specific probability functions, as shown in **Figure 9.2**. To be specific, let $\mathbf{X} = \{\boldsymbol{x}_i\}_{i=1}^{N}$ denote a heterogeneous crash dataset. Each variable $\boldsymbol{x} \in \mathbb{R}^m$ in the sample dataset can either be continuous, categorical or ordinal in the modified generative model, therefore, probability function $\boldsymbol{h}_m$ of each parameter can be independently characterized using the deep neural networks method with the specification given by,

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_k p(\boldsymbol{x}_k|\boldsymbol{z}) \tag{9.6}$$

Where $p(\boldsymbol{x}_k|\boldsymbol{z})$ is referred as the generation of the k-th variable in crash dataset.

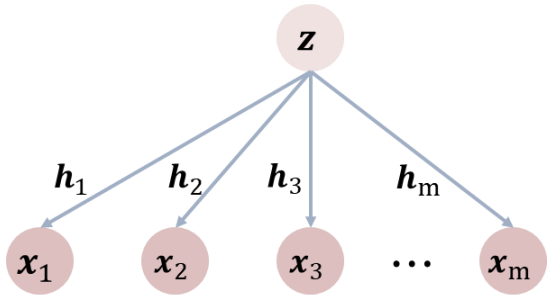


Figure 9.2 The modified generative model

With the factorization structure of the generative model, each variable in the crash dataset can be characterized by an appropriate probability function. Notably, the structure of recognition model in the variational autoencoder method remains unchanged. This ensures that the latent space can accommodate the variations in the interactions between variables.

Generally, objective function of the generative model is determined using non-zero crash cases. However, road crash is a random event. Entities that are identical can have different numbers of crashes. Hence, it is crucial to incorporate the zero crash cases into the crash frequency models. Therefore, the proposed models can better recognize the crash occurrences, given that the data misspecification problem is avoided.

Let $\mathbf{Y} = \{\boldsymbol{y}_i\}_{i=1}^{U}$ denote the zero-crash dataset and $\boldsymbol{y}_i$ refers to the *i*-th observation. In the training stage, the zero-crash cases are fed into the recognition model to derive the latent variables $\boldsymbol{v} \in \mathbb{R}^m$. Previous studies showed that a latent space can be used to detect the anomalous inputs from a normal sample (Xu et al., 2018b; Park et al., 2018). Therefore, the latent variables of the zero-crash dataset are used for the estimation of centroid given by,

$$\overline{\boldsymbol{v}} = \frac{\sum_{i=1}^{U} \boldsymbol{v}_i}{U} \tag{9.7}$$

With the latent space, the generative model is capable of synthesizing crash dataset that has excessive zero-crash observations, given that the difference in the distributions between the recognition model and the latent space of zero-crash cases is immense enough. Therefore, Kullback–Leibler divergence between $q(\boldsymbol{z}|\boldsymbol{x}, \varphi)$ and $p(\overline{\boldsymbol{v}})$ is specified as the distance metric α given by,

$$\alpha = D_{KL}(q(\boldsymbol{z}|\boldsymbol{x}, \varphi)|p(\overline{\boldsymbol{v}})) \tag{9.8}$$

Where α can be regarded as a regularization term or penalty factor that can avoid overfitting of the crash data, with which the original and zero-crash datasets are similar.

Therefore, it can increase the robustness and degree of intelligent of the generative model. Eventually, the optimization problem given by Equation 9.5 is modified as,

$$\min_{\theta,\varphi} \mathcal{L}(\theta,\varphi) = -\mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{x},\varphi)}\big(\log p(\boldsymbol{x}|\boldsymbol{z},\theta)\big) + D_{KL}(q(\boldsymbol{z}|\boldsymbol{x},\varphi)|p(\boldsymbol{z})) - \alpha \quad (9.9)$$

To estimate the model parameters θ and φ, the augmented variational autoencoder model is implemented using the multilayer perceptron approach.

- Generative model

For every variable $\boldsymbol{x}_k$ of the crash dataset, the probability density function is calibrated using the multilayer perceptron approach. The generative processes for a few common data types (e.g., real-valued, count, nominal, and ordinal) are given as follows.

(a) Real-valued data: Gaussian distribution is adopted to characterize the real-valued data specified as,

$$p(\boldsymbol{x}_k) = \mathcal{N}(\boldsymbol{x}_k|\mu_k(\boldsymbol{z}), \sigma_k^2(\boldsymbol{z})) \quad (9.10)$$

Where $\mu_k(\boldsymbol{z})$ and $\sigma_k^2(\boldsymbol{z})$ refer to the mean and variance of Gaussian distribution, that can be generated by the three-layer multilayer perceptron.

(b) Count data: Poisson distribution is adopted to characterize the count data specified as,

$$p(\boldsymbol{x}_k) = \text{Poisson}\big(\boldsymbol{x}_k\big|\lambda_k(\boldsymbol{z})\big) = \frac{\lambda_k(\boldsymbol{z})^{\boldsymbol{x}_k}\exp\left(-\lambda_k(\boldsymbol{z})\right)}{\boldsymbol{x}_k!} \quad (9.11)$$

Where $\lambda_k(\boldsymbol{z})$ refers to the mean of Poisson distribution, with which the mean is determined using the three-layer multilayer perceptron with learnable parameters θ_k specified as,

$$\lambda_k(\boldsymbol{z}) = f_{\theta_k}(\boldsymbol{z}) \quad (9.12)$$

(c) Nominal data**:** Given that there are *Q* possible discrete outcomes, probability of having outcome *r* is determined using the logit function specified as,

$$p(\boldsymbol{x}_k = r) = \frac{\exp\left(-\pi_r(\boldsymbol{z})\right)}{\sum_{q=1}^{Q} \exp\left(-\pi_q(\boldsymbol{z})\right)} \tag{9.13}$$

Which the parameters $\left[\pi_1(\boldsymbol{z}), \pi_2(\boldsymbol{z}), \cdots, \pi_Q(\boldsymbol{z})\right]$ are determined using a three-layer multilayer perceptron.

(d) Ordinal data: Probability of having the outcome *r* is determined using the ordered logit technique specified as,

$$p(\boldsymbol{x}_k = r) = p(\boldsymbol{x}_k \leq r) - p(\boldsymbol{x}_k \leq r - 1) \tag{9.14}$$

And

$$p(\boldsymbol{x}_k \leq r) = \frac{1}{1+\exp\left(-\left(\omega_r(\boldsymbol{z}) - \psi_k(\boldsymbol{z})\right)\right)} \tag{9.15}$$

Where $\omega_r(\boldsymbol{z})$ is the threshold of the observable outcome r, $\psi_k(\boldsymbol{z})$ is the unobserved outcome of $\boldsymbol{x}_k$, and the parameters $[\psi_k(\boldsymbol{z}), \omega_1(\boldsymbol{z}), \omega_2(\boldsymbol{z}), \cdots, \omega_{R-1}(\boldsymbol{z})]$ are determined using a three-layer multilayer perceptron.

- Recognition model

For the crash dataset $\boldsymbol{x}$, assume that the posterior inference is $q(\boldsymbol{z}|\boldsymbol{x}, \varphi) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_z, diag(\boldsymbol{\sigma}_z^2))$. Then, a three-layer multilayer perceptron is deployed to determine the vectors of mean $\boldsymbol{\mu}_z$ and diagonal covariance $\boldsymbol{\sigma}_z^2$ as follows,

$$[\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z^2] = f_\varphi(\boldsymbol{x}) \tag{9.16}$$

Note that latent variable $\boldsymbol{z}$ is extracted from the recognition $q(\boldsymbol{z}|\boldsymbol{x},\varphi)$. However, as the conventional optimizer - stochastic gradient descent – is not capable of estimating the differential of sample operator (Bottou, 2010), a re-parameterization trick is adopted to extract the differentiable samples from $q(\boldsymbol{z}|\boldsymbol{x},\varphi)$ (Kingma and Welling, 2013).

The size of unobserved layers of the above mentioned multilayer perceptron is set at $2m$. As the prior and posterior of latent variable $\boldsymbol{z}$ both follow the Gaussian distribution, the objective function can be refined as follows,

$$\min_{\theta,\varphi} \mathcal{L}(\theta,\varphi) = -\sum_{j=1}^{J}\left(\prod_{k=1}^{m}\log p\left(\boldsymbol{x}_k|\boldsymbol{z}_j\right)\right) - \frac{1}{2}\sum_{i=1}^{m}(1+\log\sigma_i^2-\mu_i^2-\sigma_i^2) - \alpha \quad (9.17)$$

Where J denotes the number of sampling of $\boldsymbol{z}$, and μ_i and σ_i^2 are the i-th components of $\boldsymbol{\mu_z}$ and $\boldsymbol{\sigma_z^2}$.

To synthesize the crash dataset, a vector $\boldsymbol{x}^*$ of m elements with normally distributed noise on the ground-truth crash dataset $\boldsymbol{x}$ is denoted as the input of the generative model. The model output is a reconstructed vector $\widehat{\boldsymbol{x}^*}$ that has the same data structure as $\boldsymbol{x}^*$. To control for the random effect in the recognition of non-zero crash and zero-crash cases, a generative likelihood ρ is defined to indicate the case classification accuracy as,

$$\rho = \sum_{j=1}^{J}\left(\prod_{k=1}^{m}\log p\left(\boldsymbol{x}_k|\boldsymbol{z}_j\right)\right) \quad (9.18)$$

Where $\boldsymbol{x}^*$ denotes the non-zero crash cases when $\rho > 0.5$, and the zero-crash cases when otherwise.

To testify the capability of data generation of the proposed augmented variational autoencoder method, a conventional data generative approach - synthetic minority oversampling technique-nominal continuous - is also considered. Synthetic minority over-sampling technique is a classical over-sampling approach. However, it can process the real-valued data only. Therefore, an alternative approach - synthetic minority oversampling technique-nominal continuous has been proposed. This approach can

generate synthetic data for a combination of real-valued and nominal data (Chawla et al., 2002). For real-valued data, k-nearest neighbours are generated based on the minority class (i.e., non-zero crash observations). Then, a new sample can be created using the k-neighbours. In this study, K is set at 5, in accordance with the prior studies (Cai et al., 2020; Yuan et al., 2019). For nominal data, difference can be determined based on the standard deviations of real-valued data of the minority class.

To assess the capability of the proposed augmented variational autoencoder method for data generation, Jensen-Shannon divergence (JS) can be applied (Lin, 1991; Fuglede and Topose, 2004). JS divergence evaluates the difference in the probability distributions between the synthesized and original data. It can be given by,

$$\mathrm{JSD}(P||Q) = \frac{1}{2}\sum P(X_i)\log\left(\frac{2P(X_i)}{P(X_i)+Q(X_j)}\right) + \frac{1}{2}\sum Q(X_j)\log\left(\frac{2Q(X_j)}{P(X_i)+Q(X_j)}\right) \quad (9.19)$$

$$M = \frac{1}{2}(P+Q) \quad (9.20)$$

Where P and Q denote the probability mass functions of original data X_i and synthesized data X_j, respectively.

9.3 Data

In this study, crash and traffic data in Hong Kong in the period between 2014 and 2016 are used. Traffic count data are available from the Annual Traffic Census (ATC) database. ATC covers 88.5% (i.e., 1860 km) of all trafficable roads in Hong Kong. Of the roads covered, 89 road segments have detailed traffic counts by hour of the day and vehicle type. As shown in **Figure 9.3**, the road segments under investigation are widely distributed in the whole territory. In the ATC system, the road segments are defined in such a way (e.g., between major intersections) that the geometric design and traffic flow characteristics are homogeneous along each segment (Pei et al., 2016).

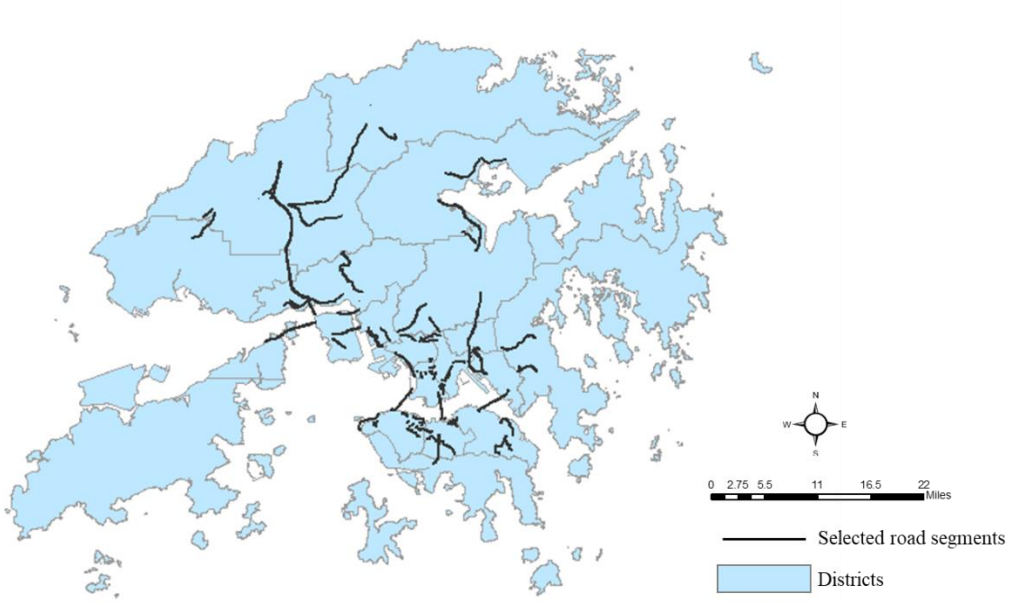


Figure 9.3 Locations of the road segments under investigation

On the other hand, crash data and roadway inventory data are available from the Transport Information System (TIS) and Hong Kong Road Network Database respectively. In the TIS, information on the location, date and time, and crash severity of every crash involving personal injury are available. In Hong Kong, crashes can be categorized into three severity levels, namely fatal crash, severe injury crash, and slight injury crash, in accordance with the degree of injury of the most seriously injured person in a crash. Since the fatal and severe injury crashes are rare, they are combined into one single group, namely fatal and severe injury crashes. In the Hong Kong Road Network Database, information on road class, number of lanes, lane width, horizontal and vertical alignment, intersection control, and speed limit of every road segment are available. In this study, information on crash incidence, traffic flow, and roadway characteristics are mapped to the corresponding road segments under investigation using the geographical information system technique.

In the ATC dataset, annual average hourly traffic flow in the 16-hour period between 7.00 am and 11.00 pm on weekdays are available. This study only considers the crashes occurred in the concerned time periods. Total number of observations is 89 (road segment) x 16 (hour) x 3 (year) = 4,272. Of the 4,272 observations, 2,020 (47.3%) have

zero crash. For fatal and severe injury crashes, 3,858 (90.3%) have zero crash. Problem of excessive zero observations is more prevalent for fatal and severe injury crashes. **Table 9.1** summarizes the data used in this study. As shown in **Table 9.1**, not all variables are real-valued. For example, total and fatal and severe injury crash frequencies are count data, road class is nominal, and speed limit is ordinal.

Table 9.1 Summary statistics of the sample

Variable	Mean	S.D.	Min.	Max.	Data type
Total crash	0.83	1.28	0	22	Count
Fatal and severe injury crash	0.11	0.34	0	3	Count
Road length (km)	2.97	3.47	0.08	19.08	Real-valued
Lane width (m)	3.62	0.52	2.70	6.20	Real-valued
Intersection density (per km)	1.90	3.10	0	13.73	Real-valued
Log (traffic flow)	3.21	0.46	0.67	4.01	Real-valued
Road class (1 = Major road; 0 = otherwise)	0.92	0.27	0	1	Nominal
Presence of bus lane (1 = Yes; 0 = otherwise)	0.16	0.36	0	1	Nominal
Number of lanes	4.85	2.32	2	12	Ordinal
Speed limit (km/h)	64.63	13.09	30	100	Ordinal

9.4 Estimation results

Separate analyses are conducted for (i) total crashes; and (ii) fatal and severe injury crashes. Specifically, the problem of unbalanced crash data is more prevalent for fatal and severe injury crashes. For total crashes, frequency models based on original and synthetic data are compared to justify the suitability of the proposed data generative method. For fatal and severe injury crashes, frequency models based on the balanced data are developed. Hence, explanatory factors that affect the occurrence of fatal and severe injury crashes would be identified.

9.4.1 Temporal stability

Temporal stability of crash frequency models is also assessed. For instance, the likelihood ratio test is conducted to examine the temporal stability across different time periods (i.e., year), using the formulation given as (Washington et al., 2011),

$$\chi^2 = -2\left[LL\left(\beta_{m_1 m_2}\right) - LL\left(\beta_{m_1}\right)\right] \quad (9.21)$$

Where $LL\left(\beta_{m_1 m_2}\right)$ is the log-likelihood at convergence for the converged parameters of the time period m_1 using the data from the time period m_2, and $LL\left(\beta_{m_1}\right)$ is the log-likelihood at convergence for the converged parameters of the time period m_1.

Null hypothesis of the test is that the parameters are constant across different years. **Table 9.2** illustrates the results of likelihood ratio tests. As shown in **Table 9.2**, for total crashes, three out of six chi-square statistics are significant at the 5% level. For fatal and severe injury crashes, three out of six chi-square statistics are significant, again at the 5% level. This implies that temporal instability should be considered when modelling the crash frequency. Hence, separate crash frequency models for different years should be established.

Table 9.2 Results of likelihood ratio test

Total crashes			
Year	2014	2015	2016
2014	N/A	12.25 (5) [0.031]	7.57 (5) [0.181]
2015	10.85 (6) [0.093]	N/A	84.48 (4) [<0.001]
2016	38.95 (6) [<0.001]	5.65 (5) [0.342]	N/A
Fatal and severe injury crashes			
Year	2014	2015	2016

2014	N/A	1.91 (2) [0.385]	8.12 (2) [0.017]
2015	4.71 (3) [0.007]	N/A	5.22 (3) [0.156]
2016	9.86 (2) [0.194]	25.81 (4) [<0.001]	N/A

Note: Degrees of freedom in the parenthesis and significant levels in the brackets

9.4.2 Total crashes

To assess the performance of proposed data generation method, total crash frequency models based on synthetic and original data are established. Specifically, the zero-crash cases remain unaltered, and the non-zero crash cases are synthesized using different deep learning approaches (Scenario 1 – Original data; Scenario 2 – Synthetic data using augmented variational autoencoder method; and Scenario 3 – Synthetic data using synthetic minority oversampling technique-nominal continuous method). As shown in **Table 9.3**, for each dataset, sample size, number of zero-crash case, and number of non-zero crash case remain unchanged in the data generation process. For the performance assessment, both statistical fit and inferences of the explanatory factors are considered (Pei et al., 2016; Yu et al., 2020).

Table 9.3 Number of observations

Dataset		Scenario 1	Scenario 2	Scenario 3
2014	Zero crash case	758	758	758
	Non-zero crash case	666	666 (Synthetic data)	666 (Synthetic data)
	Sample size	1424	1424	1424
2015	Zero crash case	737	737	737

Dataset		Scenario 1	Scenario 2	Scenario 3
	Non-zero crash case	687	687 (Synthetic data)	687 (Synthetic data)
	Sample size	1424	1424	1424
2016	Zero crash case	751	751	751
	Non-zero crash case	673	673 (Synthetic data)	673 (Synthetic data)
	Sample size	1424	1424	1424

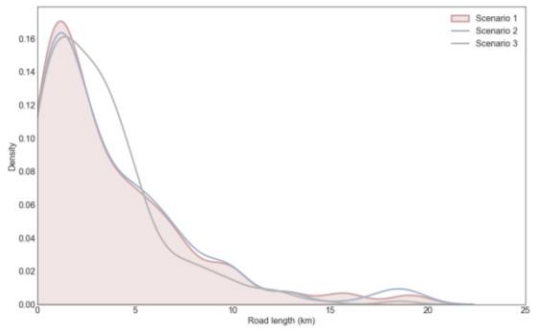
To assess the prediction accuracy of the proposed crash frequency models, the dataset is stratified into two: (i) training (80%), and (ii) test data (20%) (Gooch et al., 2018). As over-dispersion is prevalent, negative binomial regression approach is adopted. Additionally, random parameter approach is used to account for the effect of unobserved heterogeneity.

Table 9.4 summaries the results of prediction accuracy assessment based on root mean square error (RMSE) and mean absolute error (MSE) (Huo et al., 2020). As shown in **Table 9.4**, there is no significant difference in the root mean square error and mean absolute error between the models based on original data (Scenario 1) and synthetic data using the augmented variational autoencoder method (Scenario 2), for all datasets. However, the mean absolute error and root mean square error of the model based on synthetic data using the synthetic minority oversampling technique-nominal continuous method (Scenario 3) are remarkably higher than those using the augmented variational autoencoder method (Scenario 2) in general. This justifies the capability of the proposed deep generative approach. **Figure 9.4 (a)-(h)** illustrates the distributions of original data, synthetic data using the augmented variational autoencoder method, and synthetic data using the synthetic minority oversampling technique-nominal continuous method, for 2014 dataset. As shown in **Figure 9.4(a)-(h)**, deviations of distributions of synthetic data based on the augmented variational autoencoder method from that of original data are

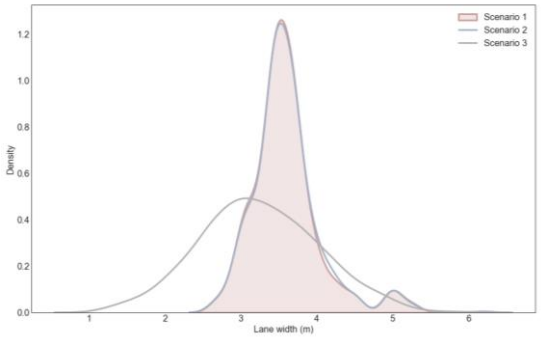
much smaller, compared to that based on the synthetic minority oversampling technique-nominal continuous method, for all variables.

Table 9.4 Prediction accuracy of total crash frequency models

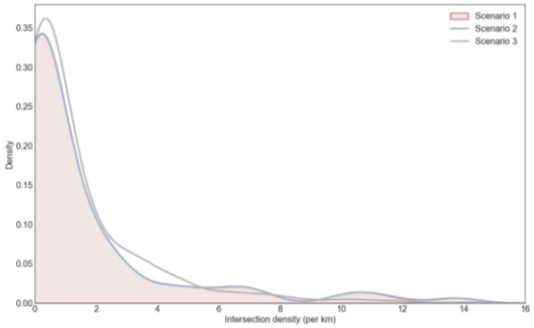
Dataset		Scenario 1	Scenario 2	Scenario 3
2014	MSE	0.923	0.938	1.042
	RMSE	1.187	1.269	1.395
2015	MSE	0.844	0.881	0.997
	RMSE	1.060	1.123	1.264
2016	MSE	0.938	0.959	1.045
	RMSE	1.202	1.276	1.421



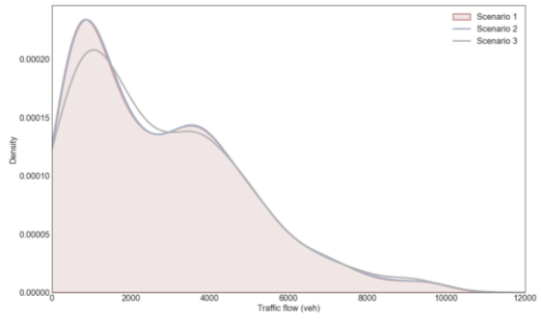
(a) Road length



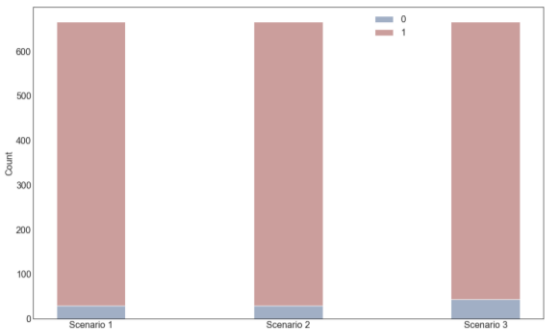
(b) Lane width



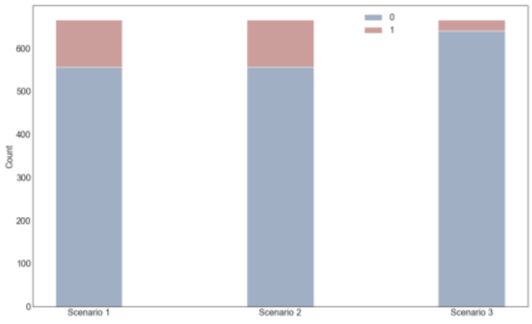
(c) Intersection density



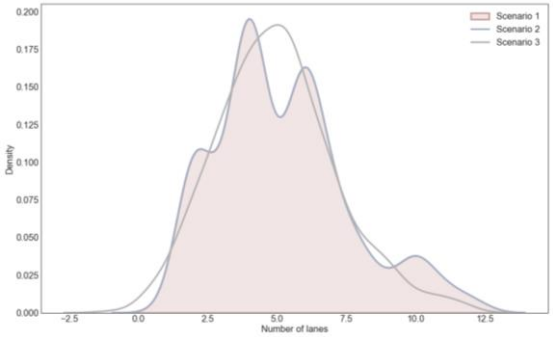
(d) Traffic flow



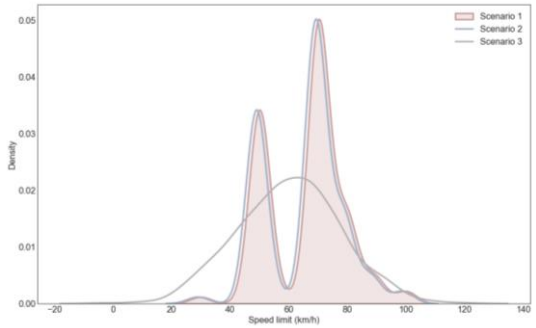
(e) Road class



(f) Presence of bus lane



(g) Number of lane



(h) Speed limit

Figure 9.4 Distributions of synthesized data (2014)

Table 9.5 presents the results of Jensen-Shannon divergence. As shown in **Table 9.5**, values of the Jensen-Shannon divergence of synthetic data based on the augmented variational autoencoder method are comparable to that based on the synthetic minority oversampling technique-nominal continuous method, for the real-valued variables including road length, intersection density, and log (traffic flow). This implies that both data generation approaches are capable of synthesizing simple real-valued data. However, the augmented variational autoencoder method is superior to the synthetic minority oversampling technique-nominal continuous method for the real-valued data that have multi-modal distributions (with remarkably smaller Jensen-Shannon divergence and as indicated in **Figure 9.4**) such as average lane width. As also shown in **Table 9.5**, the augmented variational autoencoder method is superior for the nominal and ordinal variables including road class, presence of bus lane, number of lanes, and speed limit.

Table 9.5 Jensen-Shannon divergence of synthetic data based on data generation approaches

Variable	2014		2015		2016	
	Scenario 2	Scenario 3	Scenario 2	Scenario 3	Scenario 2	Scenario 3
Road length	0.014	0.020	0.015	0.019	0.012	0.022
Lane width	0.012	0.058	0.010	0.064	0.012	0.056
Intersection density	0.008	0.014	0.011	0.016	0.007	0.016
Log (traffic flow)	0.009	0.013	0.007	0.020	0.008	0.017
Road class	0.002	0.006	0.002	0.007	0.003	0.006
Presence of bus lane	0.001	0.005	0.002	0.008	0.001	0.006
Number of lanes	0.005	0.056	0.007	0.061	0.006	0.060
Speed limit	0.021	0.037	0.025	0.049	0.020	0.047

Table 9.6 illustrates the results of parameter estimation for total crash frequency using correlated random parameter negative binomial regression method. As shown in **Table 9.6**, over-dispersions are significant at the 5% level in all models. Results of the total crash frequency models based on original and synthetic data are given as follows.

- Scenario 1

As shown in **Table 9.6**, effects of road length, lane width, and traffic flow are randomly distributed. For instance, road length is positively associated with total crash frequency (marginal effect: 0.05 to 0.07) at the 1% level of significance. Such the effect is randomly distributed in all years. In addition, traffic flow is positively associated with total crash frequency (marginal effect: 0.15 to 0.66) at the 1% level of significance. Such effect is randomly distributed in 2014 and 2016. Furthermore, lane width is negatively associated with total crash frequency (marginal effect: -0.34 to -0.11) at the 1% level of significance in 2015 and 2016. Again, such effect is randomly distributed in 2015.

Table 9.7 illustrates the results of Cholesky matrix for the correlations between random parameters. As shown in **Table 9.7**, there are negative correlations between the random parameters of traffic flow and road length in 2014 (-0.719) and 2016 (-0.350). This implies that the effects of random components of traffic flow and road length are mixed. On the other hand, there is negative correlation between the random parameters of road length and lane width in 2015 (-0.936).

As also shown in **Table 9.6**, effects of intersection density, road class and number of lanes on total crash frequency are fixed. For instance, intersection density is negatively associated with total crash frequency (marginal effect: -0.02) at the 5% level of significance in 2014 only. In addition, total crash risk of major road is significantly higher than that of minor road (marginal effect: 0.39) at the 5% level in 2014 only. Furthermore, number of lanes is positively associated with total crash frequency (marginal effect: 0.06) at the 1% level of significance in 2014 only. Nevertheless, effects of presence of bus lane and speed limit on total crash frequency are not significant in all three years.

- Scenario 2 and Scenario 3

As shown in **Table 9.6**, goodness-of-fit, in term of Akaike information criterion (AIC) and Bayesian information criterion (BIC), amongst the models based on original data (Scenario 1), synthetic data using the augmented variational autoencoder method (Scenario 2), and synthetic data using the synthetic minority oversampling technique-nominal continuous method (Scenario 3) are comparable. However, effects of explanatory factors as revealed in the crash frequency model based on original data are similar to that based on synthetic data using the augmented variational autoencoder method (Scenario 2) only. For the latter (i.e., Scenario 2), effects of traffic flow, road length and lane width on total crash frequency are randomly distributed. In addition, as also shown in **Table 9.7**, there are negative correlations between the random parameters of traffic flow and road length in 2014 (-0.674) and 2016 (-0.409), and between those of road length and lane width in 2015 (-0.952). Furthermore, consistent with Scenario 1, effects of intersection density, road class and number of lanes on total crash frequency

are fixed. For instance, intersection density is negatively associated with total crash frequency at the 5% level of significance, total crash frequency of major road is significantly higher than that of minor road at the 5% level, and number of lanes is positively associated with total crash frequency at the 1% level of significance in 2014 only. Finally, effects of speed limit and presence of bus lane on total crash frequency are not significant in all years. Above finding justifies that the proposed the augmented variational autoencoder method can generate crash data that have similar inferences, compared with that based on the original data. In contrast, effects of explanatory factors as revealed in the crash frequency models based on synthetic data using the synthetic minority oversampling technique-nominal continuous method (Scenario 3) and original data (Scenario 1) are different. For example, effects of the factors including road class, speed limit and presence of bus lane on total crash frequency as revealed in Scenario 3 are different from that in Scenario 1 and Scenario 2.

Table 9.6 Results of parameter estimation for total crashes

Variable		Scenario 1		Scenario 2		Scenario 3	
		Coefficient	Marginal effect	Coefficient	Marginal effect	Coefficient	Marginal effect
2014							
Constant		-2.11**	N/A	-2.73**	N/A	-1.50**	N/A
Road length	Mean	0.10**	0.07	0.10**	0.07	0.09**	0.06
	S.D.	2.41**	N/A	2.21**	N/A	0.09**	N/A
Lane width		IS	IS	IS	IS	IS	IS
Intersection density	Mean	-0.03*	-0.02	-0.03*	-0.02	-0.04**	-0.03
	S.D.					0.04**	N/A
Log (Traffic flow)	Mean	0.21**	0.15	0.38**	0.25	0.38**	0.27
	S.D.	0.08**	N/A	0.14**	N/A		
Road class		0.57*	0.39	0.48*	0.31	IS	IS
Presence of bus lane		IS	IS	IS	IS	-0.52**	-0.37
Number of lanes		0.09**	0.06	0.08**	0.05	0.05*	0.03
Speed limit		IS	IS	IS	IS	IS	IS

Variable		Scenario 1		Scenario 2		Scenario 3	
		Coefficient	Marginal effect	Coefficient	Marginal effect	Coefficient	Marginal effect
Number of observations		1424		1424		1424	
Mean		0.83		0.83		0.83	
Variance		1.62		1.62		1.51	
Over-dispersion parameter		2.854**		4.363**		7.989**	
Log likelihood at convergence		-1711.15		-1705.60		-1704.74	
AIC		3.11		3.12		3.12	
BIC		50.46		50.47		50.47	
2015							
Constant		-0.73	N/A	-0.92	N/A	-1.68**	N/A
Road length	Mean	0.08**	0.05	0.09**	0.06	0.12**	0.08
	S.D.	2.13**	N/A	2.06**	N/A	2.18**	N/A
Lane width	Mean	-0.51**	-0.34	-0.46**	-0.30	-0.25**	-0.16
	S.D.	0.12**	N/A	0.15**	N/A	0.14**	N/A
Intersection density		IS	IS	IS	IS	IS	IS
Log (Traffic flow)		0.57**	0.38	0.65**	0.43	0.86**	0.54
Road class		IS	IS	IS	IS	IS	IS
Presence of bus lane		IS	IS	IS	IS	-0.63**	-0.39
Number of lanes		IS	IS	IS	IS	IS	IS
Speed limit		IS	IS	IS	IS	-0.01**	-0.01
Number of observations		1424		1424		1424	
Mean		0.83		0.83		0.80	
Variance		1.65		1.65		1.41	

Variable		Scenario 1		Scenario 2		Scenario 3	
		Coefficient	Marginal effect	Coefficient	Marginal effect	Coefficient	Marginal effect
Over-dispersion parameter		3.04**		2.96**		3.617**	
Log likelihood at convergence		-1688.81		-1690.63		-1613.14	
AIC		3.14		3.13		3.23	
BIC		50.49		50.49		50.58	
2016							
Constant		-3.23**	N/A	-3.45**	N/A	-2.94**	N/A
Road length	Mean	0.08**	0.05	0.08**	0.05	0.11**	0.07
	S.D.	2.66**	N/A	2.36**	N/A	1.94**	N/A
Lane width		-0.18**	-0.11	-0.17**	-0.13	IS	IS
Intersection density		IS	IS	IS	IS	IS	IS
Log (Traffic flow)	Mean	1.04**	0.66	1.01**	0.63	1.02**	0.66
	S.D.	0.50**	N/A	0.15**	N/A	0.19**	N/A
Road class		IS	IS	IS	IS	IS	IS
Presence of bus lane		IS	IS	IS	IS	-0.42**	-0.27
Number of lanes		IS	IS	IS	IS	IS	IS
Speed limit		IS	IS	IS	IS	-0.01**	-0.01
Number of observations		1424		1424		1424	
Mean		0.83		0.83		0.84	
Variance		1.61		1.61		1.56	
Over-dispersion parameter		4.87**		5.51**		5.71**	
Log likelihood at convergence		-1697.73		-1688.64		-1719.02	
AIC		3.13		3.14		3.10	

Variable	Scenario 1		Scenario 2		Scenario 3	
	Coefficient	Marginal effect	Coefficient	Marginal effect	Coefficient	Marginal effect
BIC	50.48		50.49		50.45	

** and ** denote statistical significance at the 5% and 1% levels respectively; IS denotes insignificant*

Table 9.7 Cholesky matrix for the correlations between random parameters

2014						
	Scenario 1		Scenario 2		Scenario 3	
	Road length	Log (Traffic flow)	Log (Traffic flow)	Road length	Road length	Intersection density
Road length	2.41 (201.90) [1.000]	-0.06 (-4.06) [-0.719]	2.21 (214.56) [1.000]	-0.08 (-5.89) [-0.674]	0.09 (149.52) [1.000]	-0.03 (-2.65) [-0.756]
Log (Traffic flow)	-0.06 (-4.06) [-0.719]	0.06 (5.55) [1.000]	-0.08 (-5.89) [-0.674]	0.11 (11.03) [1.000]	N/A	N/A
Intersection density	N/A	N/A	N/A	N/A	-0.03 (-2.65) [-0.756]	0.03 (2.34) [1.000]
2015						
	Scenario 1		Scenario 2		Scenario 3	
	Road length	Lane width	Road length	Lane width	Road length	Lane width
Road length	2.13 (202.42) [1.000]	-0.11 (-7.75) [-0.936]	2.06 (194.83) [1.000]	-0.14 (-9.58) [-0.952]	2.18 (155.06) [1.000]	-0.08 (-5.02) [-0.599]
Lane width	-0.11 (-7.75)	0.04 (4.14)	-0.14 (-9.58)	0.04 (4.52)	-0.08 (-5.02)	0.11 (9.98)

	[-0.936]	[1.000]	[-0.952]	[1.000]	[-0.599]	[1.000]
2016						
	Scenario 1		Scenario 2		Scenario 3	
	Road length	Log (Traffic flow)	Road length	Log (Traffic flow)	Road length	Log (Traffic flow)
Road length	2.66 (262.62) [1.000]	-0.48 (-3.10) [-0.350]	2.36 (243.69) [1.000]	-0.06 (-4.30) [-0.409]	1.94 (162.74) [1.000]	-0.16 (-10.54) [-0.826]
Log (Traffic flow)	-0.48 (-3.10) [-0.350]	0.13 (11.96) [1.000]	-0.06 (-4.30) [-0.409]	0.14 (14.11) [1.000]	-0.16 (-10.54) [-0.826]	0.11 (11.53) [1.000]

Note: t-statistics in the parenthesis and correlation coefficients in the brackets

To sum up, crash frequency model based on synthetic (heterogeneous) data using the proposed augmented variational autoencoder method has comparable model fit and inferences, relative to that based on original data. Despite that the statistical fit among the models based on the above data generation approaches (i.e., Scenario 2 and Scenario 3), the proposed augmented variational autoencoder method is more precise for data generation, by incorporating a factorized generative model and a refined loss function. For instance, the factorized generative model is capable of sophisticated data structures like extreme-valued and multi-modal distributions (He and Garcia, 2009; Cai et al., 2020). Even that there may be incremental increase in model complexity, the proposed augmented variational autoencoder method can mitigate the misspecification problem by incorporating excessive zero-crash observations into the refined loss function. Nevertheless, goodness of fit among the models based on synthetic data using augmented variational autoencoder and synthetic minority oversampling technique-nominal continuous methods are comparable. It is not surprising as the model efficiency of the synthetic minority oversampling technique-nominal continuous method is well justified (Cai et al., 2020; Yuan et al., 2019).

9.4.3 Fatal and severe injury crashes

Since fatal and severe injury crashes are extremely rare, fatal and severe injury crash frequency models are often subject to unbalanced crash data (Pei et al., 2016). In this study, ratio of non-zero fatal and severe injury crash to zero-crash cases is 1:9 only. To this end, the proposed augmented variational autoencoder method is adopted to balance the crash data, prior to the estimation of fatal and severe injury crash frequency models. In previous studies, ratio of 1:4 (non-zero crash to zero-crash cases) is commonly adopted for data balancing (Roshandel et al., 2015; Shi and Abdel-Aty, 2015; Yuan et al., 2019) since the marginal improvement in statistical fit is incremental for the increase in the ratio beyond 1:4 (Roshandel et al., 2015; Zheng et al, 2010). Therefore, ratio of 1:4 is also adopted in this study. For instance, the fatal and severe injury crash frequency models based on original and balanced data (Scenario 4 – Original data: Correlated random parameter Poisson regression[1]; Scenario 5 – Original data: Correlated random parameter zero-inflated Poisson regression; and Scenario 6 – Balanced data: Correlated random parameter Poisson regression) are established. As shown in **Table 9.8**, model based on balanced data using the augmented variational autoencoder method (scenario 6) has superior model fit, in term of AIC and BIC. Additionally, as shown in **Table 9.10**, prediction accuracy, in terms of root mean square error and mean absolute error, of the model based on balanced data is better than that based on original data (Scenario 4 and Scenario 5). In addition, model based on balanced data can reveal more significant fatal and severe injury crash explanatory factors. For example, road length, traffic flow, intersection density and presence of bus lane are positively associated with fatal and severe injury crash frequency, and lane width and speed limit are negatively associated with fatal and severe injury crash frequency, respectively. Furthermore, effects of nominal and ordinal variables, i.e., presence of bus lane, number of lanes and speed limit, are less likely to be revealed in the model based on original crash data. Last but not least, correlations between the random parameters are considered. As shown in **Table 9.9**, there are negative correlations between the random parameters of road length and lane width

[1] Results of over-dispersion test indicate that the over-dispersion parameter for fatal and severe injury crash data is not significant at the 5% level. Therefore, correlated random parameter Poisson regression method is applied.

in 2015 (-0.967), and road length and traffic flow in 2016 (-0.886) respectively. Overall, estimation results indicate that crash frequency model based on the proposed augmented variational autoencoder approach is the best among the candidate models, with respect to the statistic fit, predictive performance, and identification of possible explanatory factors.

Table 9.8 Results of parameter estimation for fatal and severe injury crashes

Variable		Scenario 4		Scenario 5		Scenario 6	
		Coefficient	Marginal effect	Coefficient	Marginal effect	Coefficient	Marginal effect
2014							
Constant		-3.96**	N/A	-3.93**	N/A	-2.65**	N/A
Road length	Mean	0.10**	0.01	0.15**	0.002	0.14**	0.03
	S.D.			1.80**	N/A	0.08**	N/A
Lane width		IS	IS	IS	IS	IS	IS
Intersection density		IS	IS	0.13**	0.002	0.05**	0.01
Log (Traffic flow)		IS	IS	0.47**	0.006	0.28**	0.06
Road class		IS	IS	IS	IS	IS	IS
Presence of bus lane	Mean	IS	IS	IS	IS	0.50**	0.11
	S.D.						
Number of lanes		IS	IS	IS	IS	IS	IS
Speed limit		IS	IS	IS	IS	-0.01**	-0.01
Number of observations		1424				1600	
Mean		0.10				0.27	
Variance		0.10				0.27	
Over-dispersion parameter		IS (0.315)				IS (0.0001)	
Ratio of 'zero-crash' to 'non-zero crash' cases		1:10				1:4	

Variable		Scenario 4		Scenario 5		Scenario 6	
		Coefficient	Marginal effect	Coefficient	Marginal effect	Coefficient	Marginal effect
Log likelihood at convergence		-454.70		-486.94		-1034.80	
AIC		5.76		5.62		4.11	
BIC		53.11		52.97		52.51	
2015							
Constant		-0.56	N/A	0.65	N/A	-0.09	N/A
Road length	Mean	0.09**	0.01	0.10**	0.002	0.10**	0.04
	S.D.			2.99**	N/A	1.87**	N/A
Lane width	Mean	-0.85**	-0.10	-0.67**	-0.01	-0.41**	-0.11
	S.D.					0.11**	N/A
Intersection density	Mean	-0.13**	-0.01	-0.15**	-0.002	IS	IS
Log (Traffic flow)		IS	IS	IS	IS	IS	IS
Road class		IS	IS	IS	IS	IS	IS
Presence of bus lane		IS	IS	IS	IS	0.27**	0.07
Number of lanes		IS	IS	IS	IS	0.08**	0.03
Speed limit		IS	IS	IS	IS	-0.02*	-0.01
Number of observations		1424				1647	
Mean		0.12				0.316	
Variance		0.12				0.298	
Over-dispersion parameter		IS (0.259)				IS (0.0002)	
Ratio of 'zero-crash' to 'non-zero crash' cases		1:9				1:4	

Variable		Scenario 4		Scenario 5		Scenario 6	
		Coefficient	Marginal effect	Coefficient	Marginal effect	Coefficient	Marginal effect
Log likelihood at convergence		-507.47		-534.38		-1142.29	
AIC		5.54		5.43		3.92	
BIC		52.89		52.78		52.58	
2016							
Constant		-4.49**	N/A	-4.45*	N/A	-2.48**	N/A
Road length	Mean	0.09**	0.01	0.10**	0.001	0.10**	0.02
	S.D.			2.79**	N/A	2.30**	N/A
Lane width		IS	IS	IS	IS	IS	IS
Intersection density		IS	IS	IS	IS	IS	IS
Log (Traffic flow)	Mean	0.69**	0.07	IS	IS	0.46**	0.11
	S.D.					0.10**	N/A
Road class		IS	IS	IS	IS	IS	IS
Presence of bus lane		IS	IS	IS	IS	0.54**	0.13
Number of lanes		IS	IS	IS	IS	IS	IS
Speed limit		IS	IS	-0.02**	-0.0003	-0.01**	-0.004
Number of observations		1424				1600	
Mean		0.10				0.27	
Variance		0.10				0.27	
Over-dispersion parameter		IS (0.285)				IS (0.0002)	
Ratio of 'zero-crash' to 'non-zero crash' cases		1:10				1:4	
Log likelihood at convergence		-454.34		-462.62		-1023.43	

Variable	Scenario 4		Scenario 5		Scenario 6	
	Coefficient	Marginal effect	Coefficient	Marginal effect	Coefficient	Marginal effect
AIC	5.76		5.72		4.14	
BIC	53.11		53.07		52.53	

** and ** denote statistical significance at the 5% and 1% levels respectively; IS denotes insignificant*

Table 9.9 Cholesky matrix for the correlation between random parameters (Scenario 6)

2015		
	Road length	Lane width
Road length	1.87 (139.20) [1.000]	-0.11 (-5.13) [-0.967]
Lane width	-0.11 (-5.13) [-0.967]	0.03 (2.20) [1.000]
2016		
	Road length	Log (Traffic flow)
Road length	2.30 (176.06) [1.000]	-0.09 (-4.17) [-0.886]
Log (Traffic flow)	-0.09 (-4.17) [-0.886]	0.05 (2.12) [1.000]

Note: t-statistics in the parenthesis and correlation coefficients in the brackets

Table 9.10 Prediction accuracy of the fatal and severe injury crash frequency models

Dataset		Scenario 4	Scenario 5	Scenario 6
2014	MAE	0.54	0.42	0.38
	RMSE	0.95	0.88	0.81

2015	MAE	0.50	0.41	0.39
	RMSE	0.99	0.85	0.75
2016	MAE	0.49	0.44	0.40
	RMSE	0.91	0.87	0.79

9.5 Discussions

9.5.1 Road geometry designs

As shown in the **Table 9.8**, road length is positively associated with fatal and severe injury crash frequency (marginal effect: 0.02 to 0.04) at the 1% level of significance. Such finding is consistent with that of previous studies (Venkataraman et al., 2013; Huang et al., 2016; Guo et al., 2018a). Additionally, effect of road length on the fatal and severe injury crash frequency is randomly distributed (with standard deviation of 0.08 to 2.30). This could be attributed to the variations in geometric design, i.e., vertical and horizontal curvatures, along a road segment. Furthermore, lane width is negatively associated with fatal and severe injury crash frequency in 2015 (marginal effect: -0.11) only. This is because defensive driving maneuvers in emergency are more plausible when the road space increases (Pei et al., 2016; Wong et al., 2007). Again, effect of lane width on the fatal and severe injury crash frequency is randomly distributed in 2015.

9.5.2 Traffic controls

For the effects of traffic control, results indicate that presence of bus lane (marginal effect: 0.07 to 0.13) is positively associated with fatal and severe injury crash frequency. This could be attributed to the increase in possible interactions between buses and other vehicles (Pei et al., 2012). Additionally, intersection density is positively associated with fatal and severe injury crash frequency (marginal effect: 0.01) at the 1% level of significance in 2014 only. This may be attributed to the prevalence of traffic conflicts at the intersections (Wong et al., 2007). Furthermore, speed limit is negatively associated with fatal and severe injury crash frequency in all years. This is because drivers tend to

be more cautious when driving on the roads that have higher speed limits (Behnood and Mannering, 2017; Mannering, 2009; Zhang et al., 2021; Wang et al., 2021).

9.5.3 Traffic conditions

For the effect of traffic condition, traffic flow is positively associated with fatal and severe injury crash frequency in 2014 (marginal effect: 0.06) and 2016 (0.11) only. This could be attributed to the higher likelihood of vehicle interactions under high traffic flow condition. In addition, effect of traffic flow on fatal and severe injury crash frequency is randomly distributed in 2016 (standard deviation: 0.10). Furthermore, number of lanes is positively associated with fatal and severe injury crash frequency in 2015 only (marginal effect: 0.03). This may be attributed to the increase in possible vehicle interactions when number of lanes increase (Pei et al., 2016). Nevertheless, effect of road class on fatal and severe injury crash frequency is not significant.

9.5.4 Correlations between random parameters

Table 9.9 illustrates the correlation estimates between random parameters of the fatal and severe crash frequency model based on balanced data. As shown in **Table 9.9**, there is negative correlation (−0.967) between the random parameters of road length and lane width in 2015. This implies that variations in the effects of road length and lane width change in opposite directions. Random part of road length indicates the possible environmental heterogeneity along a road segment, while that of lane width indicates the possible driver heterogeneity. The heterogeneous effects of road length and lane width offset that of each other. As also shown in **Table 9.9**, there is negative correlation (−0.886) between the random parameters of road length and traffic flow in 2016. Again, this implies that the variations in the effects of road length and traffic flow change in opposite directions. Random part of traffic flow indicates the temporal heterogeneity within an hour. The heterogeneous effects of road length and traffic flow offset that of each other. The counterbalances of heterogeneous effects of road length, traffic flow, and lane width could be attributed to risk compensation of driver (Mannering and Bhat, 2014).

Hence, effect of environmental heterogeneity of the road segment may diminish when temporal and driver heterogeneities are prevalent.

9.6 Concluding remarks

Crash frequency model is often subject to excessive zero observation because of the rare nature of crashes. To address the problem of imbalanced crash data, a deep learning method - Augmented Variational Autoencoder – is proposed to tackle the unbalanced data problem by incorporating a factorized generative model into the objective function and over-sampling of non-zero crash cases, using the crash data in Hong Kong as a case study. Specifically, the crash data is stratified into two by crash severity level, i.e., total crashes and fatal and severe injury crashes.

First of all, data generation performances of the proposed augmented variational autoencoder method, together with another conventional technique – synthetic minority oversampling technique-nominal continuous method, are assessed for the total crashes. Prediction performance and inferences of the total crash frequency models based on original and synthetic data are assessed. Results indicate that prediction accuracy, in terms of RMSE and MAE, of the crash frequency model based on synthetic data using the proposed augmented variational autoencoder method is comparable to that based on original data. Additionally, the proposed method can synthesize heterogeneous data. Distributions of synthetic data based on the proposed method are consistent to that of original data, especially for those that have sophisticated data structure. Furthermore, the crash frequency model based on synthetic data using the proposed method can reveal more significant explanatory factors, compared to that using the conventional synthetic minority oversampling technique-nominal continuous method. Also, correlation between random parameters is considered.

Then, data balancing performance of the proposed augmented variational autoencoder method is assessed for the fatal and severe injury crashes, which are of extremely rare nature. Results indicate that model fit, prediction accuracy, and inferences of the fatal and severe injury crash frequency models based on balanced crash data using the proposed

augmented variational autoencoder method are superior, compared to that based on original data. For instances, road length, traffic flow, intersection density, number of lanes and presence of bus lane are positively associated with the fatal and severe injury crash frequency, while lane width and speed limit are negatively associated with the fatal and severe injury crash frequency, respectively. In addition, there are possible correlations between the random parameters of road length, traffic flow, and lane width.

To sum up, the proposed augmented variational autoencoder method can address the unbalance problem of heterogeneous data. Yet, this study also has limitations. Aggregated crash data are applied in the proposed crash frequency models. It may not be capable of capturing the time-series (i.e., daily and seasonal) variations in the association between crash and possible explanatory factors. In the future study, it is worth exploring the effect of temporal correlation on data generation and association measure when high resolution data are available. Also, it is worth exploring the efficiency of the proposed method on data balancing for the crash data with omitted variables, heteroscedasticity, and endogeneity issues. Furthermore, the proposed augmented variational autoencoder method could be limited to a few contextual (e.g., road geometry) variables. It is worth exploring to consider the environment, traffic, and behavioural variables in the data generating process when comprehensive traffic and safety data are available in the future.

Chapter 10 A deep learning approach for boundary crash problem in safety analysis

10.1 Introduction

The prevalent zonal safety analysis attracts growing interest. A great number of crash frequency models were developed to measure the associations between crash frequency and possible explanatory factors (Guo et al., 2018a; Wei and Lovegrove, 2013; Chen et al 2016; Ding et al., 2020). Countermeasures that target road safety at zonal levels can be implemented to avoid crashes and improve overall road safety. In the zonal safety analysis, crashes are often aggregated as per certain finite spatial units, such as traffic analysis zones, Greater Vancouver neighbourhoods in Canada, wards of London, or census tracts (Lovegrove and Sayed, 2006; Quddus, 2008; Siddiqui and Abdel-Aty, 2012). A considerable proportion of crashes may occur at or near the boundary of geographical units. Such crashes, also known as boundary crashes, can correlate with the explanatory variables of neighbouring geographical units, regardless of the spatial proximity. Therefore, previous studies encountered a fundamental problem of boundary crash allocations in the data preparation.

In preceding studies, mathematical approaches like half-and-half (Sun, 2009; Wei, 2010), collision density ratio (Cui et al., 2015) and iterative method (Zhai et al., 2018) were developed. However, these approaches did not consider the individual crash characteristics (e.g., crash severity), which should, in turn, correlate with the environmental, traffic, and road user characteristics of the corresponding geographical unit.

The main objective of this study is to resolve the boundary crash problem for macro-level crash frequency model by giving due consideration of individual crash characteristics in boundary crash allocation using the proposed approach. For example, association between crash frequency and influencing factors could be modified by covariates like injury severity, and collision mode (Ostrom and Eriksson, 1993; Pei et al., 2016; Ding et

al., 2020; Su et al., 2021). Hence, it is necessary to account for the crash characteristics when allocating the boundary crashes. In addition, crash analysis is often subject to incomplete and missing data. A viable way to handle the missing data is to eliminate the entries with missing values. However, this may systematically bias the parameter estimation (Miaou, 1994; Shankar et al., 1997). To this end, a deep learning approach – crash feature-based allocation method – is developed for the allocation of boundary crashes. Specifically, an integrated augmented masked autoencoder and support vector data description approach is adopted for the recognition of crash pattern, while the incomplete crash entries are masked, for crash feature-based boundary crash allocation of macro-level bicycle crash frequency models (He et al., 2016; Lu et al., 2021).

The reminder of this chapter is organized as follows. Section 10.2 describes the formulation of crash feature-based allocation. Then, data preparation is described in Section 10.3. Then, results and discussions are presented in Section 10.4 and Section 10.5, respectively. Lastly, key findings are summarized in Section 10.6.

10.2 Crash feature-based allocation

(1) Augmented masked autoencoder method

In chapter 9, a modified deep generative approach – augmented variational autoencoder method – was adopted to generate heterogeneous traffic and crash data for crash frequency model. Two connected neural networks: (a) encoder for the transformation of original data into a latent space and (b) decoder for the recovery of data from the coded space, were established. However, capability of augmented variational autoencoder method is subject to the vanilla neural networks (Kipf and Welling, 2016; Lu et al., 2022) and missing data (Johnson and Khoshgoftaar, 2019; Cai et al., 2020). To this end, a novel pattern recognition approach – augmented masked autoencoder method is proposed to complement the missing data (He et al., 2021; Dosovitskiy et al., 2020).

Figure 10.1 depicts the proposed augmented masked autoencoder method. Let $\boldsymbol{x}$ denotes the original data and $\boldsymbol{x}_+$ (assume half of the observations (He et al., 2021; Lu et

al., 2021)) denotes all observed variables including land use, road network, socio-demographics, and crash severity (Siddiqui and Abdel-Aty, 2012; Zhai et al., 2018). Then, encoding of the latent space $\boldsymbol{z}_p$ would be given by,

$$\boldsymbol{z}_p = f_e(\boldsymbol{x}_+, \varphi) \tag{10.1}$$

Where $f_e(\cdot)$ is the neural network, φ is the vector of learnable parameters.

Furthermore, the mask tokens characterized by learnable neurons $\boldsymbol{z}_m$ can be integrated with $\boldsymbol{z}_p$ to constitute a complete hidden vector $\boldsymbol{z}$. These mask tokens enable the Augmented Masked Autoencoder method to learn from the incomplete crash data by substituting the missing properties with learnable neurons, and therefore enlarge the receptive field of hidden layer to absorb more crash information. Hence, output of the decoder can be given by,

$$\tilde{\boldsymbol{x}} = f_d(\boldsymbol{z}, \theta) \tag{10.2}$$

Where θ is the vector of learnable parameters for decoder.

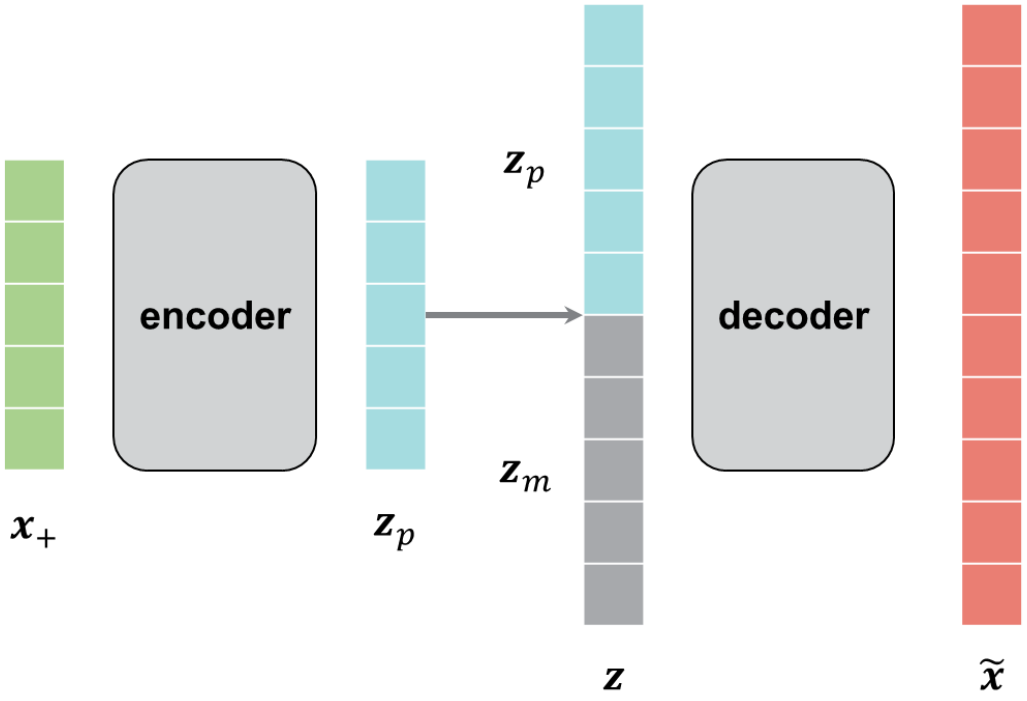


Figure 10.1 Framework of proposed augmented masked autoencoder method

(2) Support vector data description

In this study, support vector data description approach is adopted for the crash feature-based allocation (Tax and Duin, 2004). Let k_i $(i = 1, \dots, N)$ denotes the latent space of geographical unit i and the minimized error function is given by,

$$F(R, \boldsymbol{a}) = R^2 \tag{10.3}$$

Where $\boldsymbol{a}$ refers to the centre of hypersphere and R refers to the radius of hypersphere.

With the constraints,

$$\|k_i - \boldsymbol{a}\| \leq R^2, \forall i \tag{10.4}$$

To relax the constraints of outliers, the error function can be modified as,

$$F(R, \boldsymbol{a}) = R^2 + C \sum_i \xi_i \tag{10.5}$$

Where C is a penalty factor that controls the relative importance of errors.

Furthermore, a Lagrange function that characterizes the objective function for optimal crash allocation is given by,

$$L(R, \boldsymbol{a}, \alpha_i, \gamma_i, \xi_i) = R^2 + C\sum_i \xi_i - \sum_i \alpha_i \{R^2 + \xi_i - (\|k_i\|^2 - 2a \cdot k_i + \|a\|^2)\} - \sum_i \xi_i \gamma_i \quad (10.6)$$

Where α_i and γ_i are Lagrange multipliers.

Finally, d_u, distance from the centre $\boldsymbol{a}$ of a crash embedded latent space $\boldsymbol{z}_u$ for the test data can be estimated using the following expression,

$$d_u = (\boldsymbol{z}_u \cdot \boldsymbol{z}_u) - 2\sum_i \alpha_i (k_i \cdot \boldsymbol{z}_u) + \sum_{i,j} \alpha_i \alpha_j (k_i \cdot k_j) \quad (10.7)$$

A crash would be allocated to geographical unit *i* only if d_u is less than R^2. Such process is iterative until all boundary crashes are assigned.

10.3 Data

10.3.1 Sample

Same study area of chapter 6 (see **Figure 6.1**) is used in this study. Specifically, built environment, road network, population, traffic and bicycle crash data from 289 Lower Layer Super Output areas (LSOAs) in London in the year 2017-2019 would be used. To estimate the bicycle crash exposure, transaction records of London's public cycle hire scheme – LCH– are used (Ding et al., 2020). In summary, Sample size of the proposed model is 867. Above data are mapped into the corresponding Lower Layer Super Output Areas (LSOA) using the geographical information system technique. **Table 10.1** provides the descriptive statistics of the sample.

Table 10.1 Summary statistics of the sample

Category	Variable	Mean	Standard deviation	Min.	Max.
Exposure	Bicycle usage	23,116	20,866	352	122,980
	Annual average daily traffic	17,405	14,104	282	104,745
Socio-demographics	Population	1,989	547	984	4,499
	Proportion of age 65 or above	0.11	0.05	0.02	0.29
	Proportion of male	0.51	0.03	0.42	0.64
	Median household income (€)	26,145	11,752	6,063	57,935
Land use	Proportion of residential area	0.17	0.07	0.03	0.46
	Proportion of commercial area	0.20	0.12	0.01	0.56
	Proportion of green area	0.31	0.15	0.03	0.82
	Proportion of road area	0.32	0.08	0.10	0.77
Road network characteristics	Road density (km per km^2)	6.93	3.96	0.11	22.16
	Cycle path density (km per km^2)	0.06	0.09	0	0.84
	Intersection density (per km^2)	0.11	0.08	0.001	0.72
	Connectivity	2.24	0.83	0.12	4.00
	Global integration	628.60	88.37	421.20	843.32
Crash characteristics	Proportion of fatal bicycle crash	0.002	0.03	0	0.50

	Proportion of severe bicycle crash	0.16	0.26	0	1.00
	Proportion of slight bicycle crash	0.84	0.27	0	1.00
	Proportion of male bicyclist involved	0.73	0.32	0	1.00

Note: Number of observations = 289 Lower Layer Super Output Area x 3 year = 867

10.3.2 Buffer zones and boundary crashes

In this study, optimal buffer zones are set out based on the cumulative distribution of boundary crashes (Siddiqui and Abdel-Aty, 2012). As shown in **Figure 10.2**, starting from 40 meters, the curve slopes tended to be almost flat compared with that between 0 and 40 meters. Therefore, width of buffer zones is set at 40 metres. To this end, of the 4,811 bicycle crashes, 2,909 (60.5%) are considered as boundary crashes and 1,902 (39.5%) are interior crashes respectively. However, it should be noted that boundary buffer could be varied based on the history crash data and study areas. Buffer distance of 40 meters is not a unique threshold for the discrimination between interior or boundary crashes.

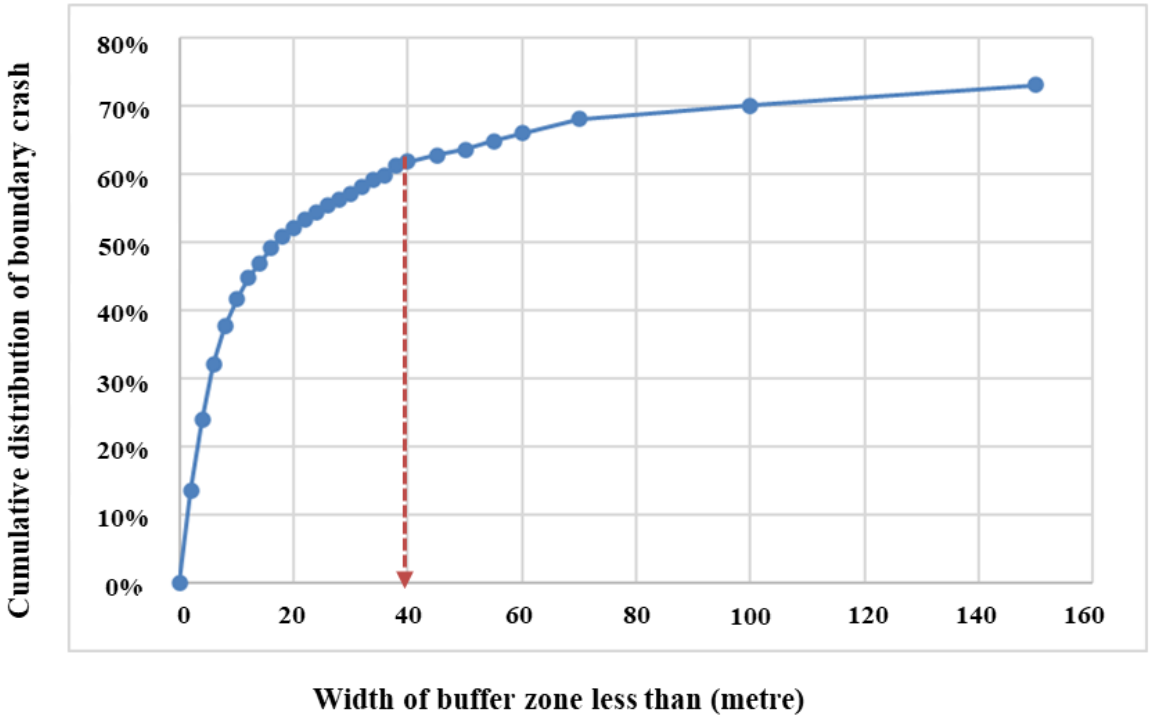


Figure 10.2 Cumulative distribution of crash with respect to the width of buffer zones

In addition, mathematical simulation is adopted to evaluate the performance of proposed crash feature-based allocation method. For example, as shown in **Figure 10.3**, two crashes (no. 8 and no. 9) are classified as boundary crashes and seven are interior crashes when the width of buffer zone is set at 40 metres. Crash no. 8 and no. 9 are treated as the "controls" for the mathematical simulation. When the width of buffer zone increases, some interior crashes would be reconsidered as "boundary crashes" (i.e., crash no. 2, no. 4, and no. 6 as shown in **Figure 10.3**). Then, proposed approach is applied to allocate such boundary crashes, considering the individual crash characteristics. **Table 10.2** summarizes the match percentages that determine how closely the boundary crash allocation and original interior crashes match with each other. As shown in **Table 10.2**, there are negligible changes in the match percentages when the width of buffer zone increases. This should justify the consistency of the proposed crash allocation method.

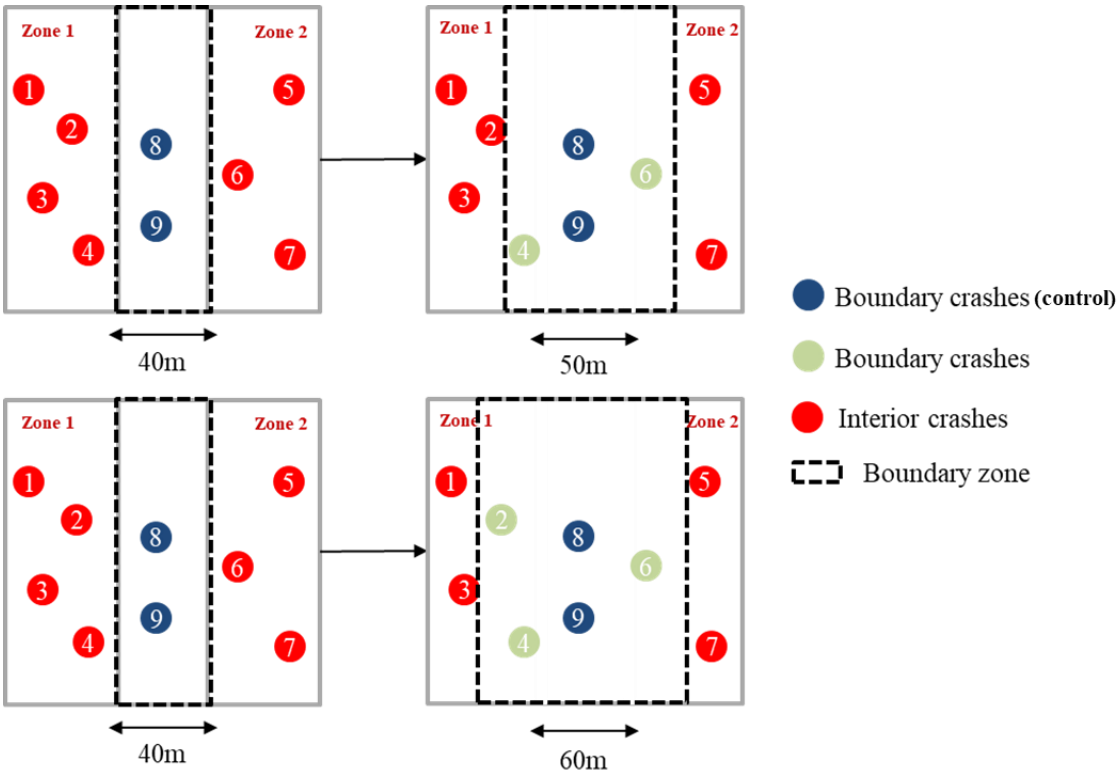


Figure 10.3 Illustration of buffer zone, boundary crashes, and interior crashes

Table 10.2 Match percentages of crash feature-based allocation and interior crashes

Width of buffer zone (change)	Boundary crashes allocated	Match percentage
50 metres (+10 metres)	3,079	94.9%
60 metres (+20 metres)	3,178	94.9%
70 metres (+30 metres)	3,322	95.1%
80 metres (+40 metres)	3,345	95.1%
90 metres (+50 metres)	3,358	95.3%
100 metres (+60 metres)	3,367	94.9%

10.4 Estimation results

In this study, macro-level bicycle crash frequency models are established using correlated random parameters approach, with which boundary crash problem is accounted. For example, proposed crash feature-based, iterative, and half-and-half allocation methods are considered. Performance of the proposed crash feature-based allocation method is

testified based on prediction performance, model fit, and influencing factors identified of the crash frequency models.

10.4.1 Temporal stability

The temporal stability is firstly examined using a likelihood ratio test (See Equation 9.21). As shown in **Table 10.3**, strong temporal instabilities between the estimated models were detected, with significant chi-square statistics at the 5% level. This implies that separate models should be developed to comprehensively understand the effects of factors on the bicycle crash frequency. Therefore, separate bicycle crash frequency models are established for year 2017, 2018 and 2019, respectively.

Table 10.3 Likelihood ratio tests for temporal stability

Year	2017	2018	2019
2017	N/A	52.46 (28) [0.004]	49.21 (26) [0.005]
2018	53.88 (30) [0.005]	N/A	51.73 (28) [0.005]
2019	45.20 (28) [0.025]	55.25 (30) [0.003]	N/A

Note: Degrees of freedom in the parentheses and significant levels in the brackets

10.4.2 Crash frequency model based on different allocation methods

Also, a multi-collinearity test is conducted to assess the correlations between independent variables. Values of variance inflation factor (VIF) are less than five for all variables. **Table 10.4** presents the results of parameter estimation of the bicycle crash frequency models. As shown in **Table 10.4**, prediction performances, in terms of root mean square error (RMSE) and mean absolute error (MSE), of the models based on crash feature-based allocation method are remarkably better than that of the counterparts, even that the improvements in model fit (in terms of AIC and BIC) may not be obvious. Also, more

influencing factors that affect the bicycle crash frequency can be identified. This justifies the use of proposed approach for individual boundary crash allocation when developing macro-level crash frequency models. Based on the results of parameter estimation, bicycle usage, traffic flow, population, male, road density, and global integration are positively associated with bicycle crashes. In contrast, median household income, residential area, and intersection density are negatively associated with bicycle crashes.

Table 10.5 illustrates the estimates of correlations between random parameters. As shown in **Table 10.5**, there are positive correlations between the random parameters of male and residential area in year 2017 (0.926) and 2019 (0.839).

Table 10.4 Results of parameter estimation of bicycle crash frequency models

Variable		Crash feature-based allocation		Iterative allocation		Half and half allocation	
		Coefficient	Marginal effect	Coefficient	Marginal effect	Coefficient	Marginal effect
2017							
Constant		-17.52**	N/A	-20.97**	N/A	-19.03**	N/A
Ln (Bicycle usage)		0.20*	0.29	0.23*	0.22	0.25**	0.22
Ln (Annual average daily traffic)		0.29*	0.44	0.39**	0.38	0.36**	0.32
Ln (Population)		IS	IS	IS	IS	IS	IS
Age 65 or above		IS	IS	IS	IS	IS	IS
Male	Mean	6.58**	9.79	2.87**	7.61	7.59**	6.68
	S.D.	4.08**	N/A	2.87**	N/A		
Median household income		IS	IS	IS	IS	IS	IS
Residential area	Mean	-5.73**	-8.52	-5.17**	-4.96	-6.31**	-5.03
	S.D.	2.59**	N/A	2.38**	N/A	1.12**	N/A
Road area		IS	IS	IS	IS	IS	IS
Road density		0.06*	0.09	0.06*	0.06	0.06*	0.06

Variable		Crash feature-based allocation		Iterative allocation		Half and half allocation	
		Coefficient	Marginal effect	Coefficient	Marginal effect	Coefficient	Marginal effect
Cycle path density		IS	IS	IS	IS	IS	IS
Intersection density		-4.12**	-6.12	-5.48**	-5.26	-5.71**	-5.03
Ln (Global integration)		1.67*	2.48	IS	IS	1.44*	1.26
Over-dispersion parameter		1.45**		1.73**		3.07**	
Log-likelihood at convergence		-1680.79		-1209.77		-1129.22	
AIC		9.15		9.80		9.94	
BIC		53.10		53.76		53.90	
MAE		0.22		0.31		0.36	
RMSE		0.39		0.49		0.58	
2018							
Constant		-9.46	N/A	-12.58	N/A	-11.64*	N/A
Ln (Bicycle usage)		0.43**	0.17	0.43**	0.18	0.30*	0.17
Ln (Annual average daily traffic)	Mean	0.28*	0.11	0.34*	0.14	0.30*	0.17
	S.D.	1.42**	N/A	2.29**	N/A	3.17**	N/A
Ln (Population)		IS	IS	IS	IS	IS	IS
Age 65 or above		IS	IS	IS	IS	IS	IS
Male		8.93**	3.48	9.19**	3.90	9.81**	5.62
Median household income		-0.07**	-0.03	IS	IS	IS	IS
Residential area		-5.65**	-2.20	-6.01**	-2.55	-6.33**	-3.63

Variable		Crash feature-based allocation		Iterative allocation		Half and half allocation	
		Coefficient	Marginal effect	Coefficient	Marginal effect	Coefficient	Marginal effect
Road area		IS	IS	IS	IS	IS	IS
Road density		IS	IS	IS	IS	IS	IS
Cycle path density		IS	IS	IS	IS	IS	IS
Intersection density		-7.11**	-2.77	-8.79**	-3.73	-7.85**	-4.50
Ln (Global integration)		IS	IS	IS	IS	IS	IS
Over-dispersion parameter		1.07**		1.05**		1.43**	
Log-likelihood at convergence		-1140.53		-890.07		-802.57	
AIC		9.92		10.42		10.62	
BIC		53.88		54.37		54.58	
MAE		0.20		0.29		0.35	
RMSE		0.37		0.48		0.57	
2019							
Constant		-20.95**	N/A	-21.10**	N/A	-20.47**	N/A
Ln (Bicycle usage)		0.19*	0.18	0.21**	0.19	0.20**	0.20
Ln (Annual average daily traffic)		IS	IS	IS	IS	IS	IS
Ln (Population)		0.74*	0.72	0.73**	0.64	0.67**	0.65
Age 65 or above		IS	IS	IS	IS	IS	IS
Male	Mean	6.28*	6.07	5.59**	4.96	5.75	5.61
	S.D.	1.78*	N/A			3.30	N/A
Median household income		IS	IS	IS	IS	IS	IS
Residential area	Mean	-6.37**	-6.17	-6.13**	-5.44	-4.70	-4.59
	S.D.	2.66**	N/A	1.01*	N/A		

Variable	Crash feature-based allocation		Iterative allocation		Half and half allocation	
	Coefficient	Marginal effect	Coefficient	Marginal effect	Coefficient	Marginal effect
Road area	IS	IS	IS	IS	IS	IS
Road density	0.07**	0.06	0.05**	0.04	0.05*	0.04
Cycle path density	IS	IS	IS	IS	IS	IS
Intersection density	-4.78**	-4.62	-3.83**	-3.40	-3.94**	-3.85
Ln (Global integration)	1.71**	1.65	1.63**	1.45	1.63**	1.59
Over-dispersion parameter	2.25**		5.66**		3.86**	
Log-likelihood at convergence	-1250.15		-964.77		-886.35	
AIC	9.74		10.26		10.43	
BIC	53.69		54.21		54.38	
MAE	0.21		0.32		0.38	
RMSE	0.37		0.51		0.59	

** and ** denote statistical significance at the 5% and 1% levels respectively; IS denotes insignificant*

Table 10.5 Cholesky matrix for the correlations between random parameters (crash feature-based allocation)

2017						
	Crash feature-based allocation		Iterative allocation		Half and half allocation	
	Male	Residential area	Male	Residential area	Male	Residential area
Male	4.08 (12.70) [1.000]	2.40 (2.42) [0.926]	2.87 (8.98) [1.000]	2.12 (2.14) [0.889]	N/A	N/A

Residential area	2.40 (2.42) [0.926]	0.98 (2.24) [1.000]	2.12 (2.14) [0.889]	1.09 (2.33) [1.000]	N/A	N/A
2019						
	Male	Residential area	Male	Residential area	Male	Residential area
Male	1.78 (5.51) [1.000]	2.24 (2.05) [0.839]	N/A	N/A	N/A	N/A
Residential area	2.24 (2.05) [0.839]	1.44 (3.14) [1.000]	N/A	N/A	N/A	N/A

Note: t-statistics in the parentheses and correlation coefficients in the brackets

10.5 Discussions

10.5.1 Bicycle crash exposures

For the bicycle crash exposure, bicycle crash frequency is positively associated with bicycle usage (all years) and traffic flow (year 2017 and 2018 only). This is consistent with the findings of previous studies. For example, it is effective to estimate the bicycle crash exposure using the bicycle usage data from the public bicycle sharing system (Ding et al., 2020, 2021c). In addition, increase in traffic flow can result in more frequent bicycle-vehicle interactions and conflicts (Wong et al., 2007). However, effect of traffic flow on bicycle crash frequency is random (year 2018 only). This may be because of the heterogeneity of road design and environment (Mannering, 2018; Meng et al., 2021). Furthermore, population is positively associated with bicycle crash frequency (year 2019 only). This is consistent with the findings of previous studies (Siddiqui and Abdel-Aty, 2012; Vanparijs et al., 2015).

10.5.2 Demographic and socioeconomics

For the population socio-demographics, bicycle crash frequency is positively associated with the proportion of male (all years). This could be attributed to the difference in safety perception between male and female. Hence, male bicyclists are more likely involved in a crash (Guo et al., 2018a; Ding et al., 2020). Just, effect of the proportion of male is random (year 2017 and 2019 only). On the other hand, median household income is negatively associated with bicycle crash frequency (year 2018 only). This could be because members of higher income households tend to be risk-averse. Hence, safe cycling behaviour and use of protective devices like helmet are more prevalent (Chen et al., 2020b; Zhu et al., 2021).

10.5.3 Built environments

For the built environment, bicycle crash frequency decreases with residential area (all years). This could be attributed to the implementation of effective local area traffic management and traffic calming measures that can mitigate the vehicle-bicycle conflicts. Also, space allocation for motorized and non-motorized traffic could have been optimized (Su et al., 2021). Just, the effect of residential area is random (year 2017 and 2019 only). This could be because of the heterogeneity of socio-cultural characteristics, risk communication, and public education, which are usually not observed in the population census.

10.5.4 Road network characteristics

For the road network characteristics, bicycle crash frequency increases with road density (year 2017 and 2019 only). This could be attributed to the increase in vehicle-bicycle conflicts, especially when there are limited separations between motorized and non-motorized traffic (Wong et al., 2007; Li et al., 2018; Ding et al., 2020). In addition, bicycle crash frequency increases with the level of global integration (year 2017 and 2019 only). This could be attributed to the increase in vehicle-bicycle conflicts when the degree of

network connection increases (Guo et al., 2018a). However, bicycle crash frequency decreases with intersection density (all years). This might be because of the compensatory strategy of drivers and bicyclists. For example, safety awareness would increase and travelling speed would reduce when one approaches an intersection (Mannering and Bhat, 2014; Chen et al., 2020b). Despite that bicycle lane is recognized to be effective in improving the safety perception of bicyclists. Effect of bicycle path density on bicycle crash is not significant. This could be because of the heterogeneity of geometric design and physical separation between motorized and non-motorized traffic of different bicycle paths in the network. Hence, favourable safety effect of bicycle path could have been offset (Li et al., 2017, 2018; Ding et al., 2021c).

10.5.5 Correlations between random parameters

Last but not least, there is positive correlation between the random parameters of male and residential area. This could be because the heterogeneity of gender effect could be magnified by that of road environment. For example, variations in safety perception and risk-taking behaviour could be more rigorous when the street layout is changed and the local area traffic management is implemented (Zhu et al., 2022).

10.6 Concluding remarks

To enhance overall bicycle safety, macro-level bicycle crash frequency models have been developed to identify influencing factors that affect the bicycle crash risk, and implement optimal urban planning and traffic management measures (Wei and Lovegrove, 2013; Ding et al., 2020, 2021c). However, boundary crash problem is prevalent when the boundaries of geographical units are delineated by roads. Thus, parameter estimation results of crash frequency models can be biased (Zhai et al., 2018; Cui et al., 2015; Siddiqui and Abdel-Aty, 2012). It is necessary to develop an effective method for the allocation of boundary crashes to neighbouring geographical units. In this study, a deep learning approach is developed for boundary crash allocation. For example, crash features are considered using the integrated Augmented Masked Autoencoder and Support Vector Data Description methods.

An illustrative case study using the population, land use, traffic and bicycle crash data from 289 Lower Layer Super Output areas in London is conducted. Cumulative distribution of boundary crash with respect to the width of buffer zone of boundary is estimated. Thus, optimal width of the buffer zone is set at 40 metres. Also, match percentages between the allocated crashes and interior crashes (classified by spatial proximity) are assessed. Consistency of the proposed crash feature-based allocation method is justified. In addition, performances of crash frequency models using correlated random parameters method based on proposed crash feature-based allocation, iterative, and half-and-half methods are compared. Results indicate that prediction performances of the crash frequency models based on crash feature-based allocation method are better than that using iterative and half-and-half methods. Also, more influencing factors that affect the bicycle crash frequency are identified. For example, bicycle and traffic flow, gender, household income, land use, and road network configuration can affect the macro-level bicycle crash frequency. Also, there are significant correlations between the random effects of gender and land use.

Nevertheless, this study also has limitations. First, association between bicycle crash frequency and influencing factors can be modified by time period. There are considerable variations in bicyclist behaviour across different seasons and weather conditions (Ding et al., 2020). Hence, it is worth exploring the effects of temporal variation on boundary crash allocation and parameter estimation when high resolution weather and traffic data, in short time intervals, are available in the future study (Xing et al., 2019). In addition, parameter estimation results may vary with the configuration and scale of geographical units. It is necessary to account for the effect of geographical configuration when allocating the boundary crashes (Zhai et al., 2019b).

Chapter 11 Conclusions and recommendations

11.1 Summary

This dissertation seeks to deepen knowledge of bicycle travel and safety. The study has contributed to the body of literature by strengthening the exploration of a number of fundamental issues. Firstly, the effects of policy interventions on bicycle travel and safety are investigated using a sophisticated inference method known as propensity score matching. Secondly, associations between the built environment, population characteristics, traffic characteristics, and bicycle safety are evaluated to account for the exposure to bicycle crashes. Then, a weighted shortest path approach incorporating the effects of path distance and perceived safety level is proposed for estimating bicycle travel distance. Finally, advanced statistical and deep learning models are developed to address the prevalent issues in (bicycle) safety analysis, such as the correlation between different crash types, excessive zero observations, and boundary crash problems. Overall, the findings can contribute to a greater understanding of the roles of environmental, traffic, and human factors in bicycle travel and safety. Therefore, the optimal urban planning, engineering design, and transportation policy can be implemented to encourage bicycle travel and enhance bicycle safety over time.

Chapter 2 reviews the literature concerning bicycle travel and safety from diverse angles. First, a summary of factors that contribute to bicycle travel is presented. Despite the fact that a number of studies have been conducted to evaluate the effects of policy interventions, rarely has the impact of traffic emission interventions been studied. In fact, the introduction of traffic emission interventions can also motivate the shift to greener modes of transportation, e.g., public transportation, cycling, and walking. Second, the safety effects of built environments, population and socioeconomic factors, and traffic characteristics on bicycle crash frequency are revealed. For the effects of policy intervention on bicycle safety, there was a discernible lack of research. The causal relationships between policy interventions, bicycle travel, and bicycle safety are ambiguous. In addition, accurate bicycle exposure data are essential for quantifying the probability of bicycle crash involvement and interpreting the risk for various entities.

Since bicycle counts are limited, bicycle exposures are generally challenging to measure. Prior studies commonly utilized prospective and retrospective approaches, which may be susceptible to recall and section biases. Lastly, modelling for safety analysis is concluded. Several research gaps are acknowledged. For instance, the differences in potential influencing factors on the risk of various types of bicycle crashes are rarely studied. Possible correlations between different types of bicycle crashes should be considered. In addition, excessive zero observations and boundary crash issues should be highlighted when developing (bicycle) crash frequency models. Overall, the aforementioned research gaps motivate the work performed in Chapters 3 to 10.

Chapter 3 formulates and elaborates the methodologies of the models adopted in this dissertation, including the random-parameter Poisson regression model, the random-parameter negative binomial regression model, the correlated random-parameter models, the multivariate Poisson-lognormal regression, and the propensity score matching approach.

To mitigate the hazardous effects of vehicle emissions, numerous engineering measures and policy strategies have been implemented. After the implementation of the Ultra-Low Emission Zone(ULEZ) in London, CO_2, NO_2 and NO_x emissions were reduced by 6%, 37%, and 35%, respectively. In addition to reducing vehicle emissions in Central London, ULEZ can also relieve traffic congestion (GLA, 2019). Cycling, as a green mode of transportation, not only helps to alleviate traffic congestion and reduce vehicle emissions, but also improves social well-being. As such, it is reckoned that private car users may switch to cycling to avoid steep fees of ULEZ. In Chapter 4, the effects of ULEZ on the utilization of bike-sharing services in London are examined. To account for the confounding effects of other factors, a propensity score matching strategy is adopted. Bike usage data from 699 bicycle docking stations between May and October of 2019 is obtained. The results indicate that bicycle demand significantly increases after the introduction of ULEZ. Specifically, increases in short (less than 15 minutes) and intermediate (between 15 and 30 minutes) bicycle trips are more considerable than increases in long bicycle trips (more than 30 minutes). In addition, the results indicate a remarkable increase in the number of bicycle trips that ended within the ULEZ. The

findings of this study should be indicative of the decision-making of transport planners and engineers, particularly with regard to the policy strategies that can enhance the level of service provided by bike-sharing systems.

In Chapter 5, we attempt to evaluate the effects of policy intervention on bicycle safety, drawing from the data from 333 Lower Super Output Area (LSOA) in London from 2011–2012. Policy interventions, including the London Cycle Hire (LCH) and the London congestion charge (LCC) schemes, are considered. Since cyclists are more susceptible to road injuries, it is hypothesized that the total number of bicycle crashes may increase after the LCH and LCC programs are implemented, as there will be more bicyclists on the roads. Once again, a PSM approach is applied. Consequently, the effects of confounding factors can be eliminated by systematically establishing an "control" group. According to the results, areas with LCH introduced have substantially higher rates of overall (38%) and minor bicycle crashes (32%) compared to areas without LCH. However, there was no discernible effect on the KSI bicycle crash. In addition, the marginal effect of LCC is estimated. The numbers of overall (59.1%) and minor bicycle crashes (57.8%) are significantly higher in areas with both LCC and LCH than in areas with only LCH. Moreover, the KSI bicycle crash evidenced no notable changes.

To assess the risk of bicycle crashes, it is necessary to estimate exposure measures. On the basis of comprehensive traffic count data, annual average traffic flow (AADT) and vehicle kilometre travelled (VKT) can be leveraged to estimate the exposure for vehicle crashes (Pei et al., 2012). However, bicycle count data are rarely available. In Chapter 6, we tackle the conundrum of accurate measurement of bicycle crash exposure based on the bicycle trip data of a public bicycle rental system. A random parameter negative binomial model is developed to measure the association between potential factors and bicycle crash frequency at the zonal level, taking bicycle exposure into account. Moreover, separate bicycle crash frequency models are developed for the warm season and the cold season. Therefore, the behaviour of cyclists in various weather conditions could be considered. The model adopting bicycle use time as the exposure measure, according to the results, is superior to its counterparts (such as the population and bicycle usage frequency) with the lowest AIC and BIC. It is discovered that road density, bicycle

facilities, land use, demographic, socioeconomic, and household attributes are associated with bicycle crash incidence. Notably, the presence of a Cycle Superhighway and the proportion of green area can have seasonal effects on the frequency of bicycle crashes.

In Chapter 6, transaction records from the London public bicycle rental system are utilised to estimate bicycle crash exposure. Although this system covered the majority of bicycle trips in London, exposure measurements are confined to bicycle trips and bicycle travel time (BTT). The bicycle distance travelled (BDT) exposure measure is unavailable. The shortest path method (SPM) and the weighted shortest path method (WSPM) are proposed in Chapter 7 to model the bicycle path selection and estimate the BDT. In particular, the proposed WSPMs consider the effects of path distance and perceived safety level on routing decisions. Initially, bicycle crash frequency models are developed that utilise BDTs as the exposure estimate derived from SPM and WSPM. The results indicate that bicycle crash frequency models that incorporate BDTs using WSPM deliver a superior model fit. In addition, three exposure measures are evaluated, including bicycle trips, BTT, and BDT. The results prove that the bicycle crash frequency model with BDTs as the exposure outperforms those with bicycle trips and BTT as the exposures. The findings of this study should be indicative of the development of bicycle crash frequency models. Furthermore, based on reliable estimates of bicycle exposures, it should facilitate an understanding of the roles of environmental, traffic, and cyclist factors in bicycle crash risk.

Prior studies have identified the environmental, traffic, and road user factors that influence the risk of bicycle-related crashes. However, differences in their effects on risk among various bicycle crash types are rarely investigated. For example, there may be a correlation between the counts of various crash types. In Chapter 8, a multivariate Poisson-lognormal regression method is applied to explore the relationship between bicycle crash frequencies and potential explanatory factors, with the correlation between bicycle-vehicle and bicycle-bicycle crashes in London in 2018 and 2019 being considered. Additionally, the effects of road network characteristics (e.g., connectivity and accessibility) on bicycle crash frequency are considered. In terms of metrics like DIC, the results depict that the proposed multivariate Poisson-lognormal model performs

superiorly to the conventional univariate Poisson-lognormal models. Moreover, variables such as population socioeconomics, land use, and road network characteristics can be identified as influencing factors on bicycle crash risk. For instance, the effects of traffic flow, residential area, network connectivity, and intersection density on crash counts vary between bicycle-vehicle and bicycle-bicycle crashes. The findings of this study are applicable to the implementation of corrective measures that can improve bicycle safety in the long term.

Due to the infrequency of crashes, crash frequency models are recurrently subject to an excess of zero observational data. Chapter 9 proposes a deep generative approach — augmented variational autoencoder — to resolve the predicament of imbalance of crash data. This approach features a factorized generative model and a refined objective function. For example, the generative model is able to process heterogeneous data, such as those with real-valued, nominal, and ordinal distributions. The refined objective function, on the other hand, can control the random effect by better identifying both the zero-crash and non-zero-crash cases. To evaluate the efficacy of the proposed method, comprehensive traffic and crash data of multiple distribution types in Hong Kong from 2014 to 2016 are utilised. Specifically, separate analyses are conducted for total crashes, and fatal and severe injury crashes, respectively. For total crashes, the parameter estimation results of the crash frequency model based on synthetic data using the augmented variational autoencoder method were closer to those based on original data in terms of statistical fit, prediction accuracy, and identified explanatory factors than those based on synthetic data using the synthetic minority oversampling technique-nominal continuous method. Zero-crash observations are prevalent for fatal and severe injury crashes, with a 9-to-1 ratio of zero-crash to non-zero-crash cases. Utilising the proposed augmented variational autoencoder method, crash data is first balanced. Then, fatal and severe injury crash frequency models are estimated by applying correlated random parameter models and original and balanced data, respectively. With the lowest RMSE, the lowest MAE, and the highest number of crash explanatory factors identified, the fatal and severe injury crash frequency model based on balanced data outperforms its counterpart. The ability to identify the correlation between the random parameters is even more significant. The findings of this study should shed light on the development of

bicycle crash frequency models for both researchers and practitioners, as the problem of excessive zero observations is prevalent when highly disaggregated traffic and crash data by time and space is used.

In the conventional safety analysis, traffic and crash data are frequently aggregated at the census tract, street block, and traffic analysis zone levels, which are typically delineated by roads and other physical entities. A significant proportion of crashes may occur at or near the boundaries of geographic units. Such crashes, also defined as boundary crashes, can correlate with the explanatory variables of neighbouring geographical units irrespective of their spatial proximity. The ambiguous allocation of boundary crashes may have an effect on the performance of crash frequency models, resulting in the estimation of erroneous parameters. In Chapter 10, a crash feature-based allocation method based on deep learning is developed for the allocation of boundary crashes. In the crash allocation process, for instance, crash severity and bicyclist characteristics are factored in. To assess the performance of the proposed method, an illustrative case study is conducted using built environment, population, traffic, and bicycle crash data from 289 Lower Super Output Area (LSOA) in London during the period of 2017-2019. High matching percentages of boundary crash allocation are possible, as indicated by the results. Furthermore, in terms of RMSE and MAE, the prediction performances of crash frequency models based on the proposed crash feature-based allocation method are superior to those based on conventional boundary crash allocation methods such as half-and-half and iterative assignment approaches. Last but not least, it is possible to identify additional macroscopic-level factors that affect bicycle crash frequency. The findings should be evident in the spatial safety analysis for various geographical configurations.

11.2 Main findings and contributions

The main findings of this thesis are concluded as follows:

(1) Effects of policy interventions on bicycle travel and safety

Capitalising on an advanced causal inference tool - the propensity score matching method - causal links between policy intervention, bicycle travel, and bicycle safety are evaluated. For instance, favourable effects of the low emission zone (e.g., ULEZ) on bicycle usage are revealed. In addition, the results suggest that bike-sharing and congestion pricing schemes can affect the volume and speed of traffic on the road, causing a rise in the frequency of bicycle crashes, particularly those resulting in slight injuries.

(2) Advance the estimation of exposure in the bicycle safety analysis

Adopting ridership data from a public bicycle rental system, a valid method for measuring bicycle exposures is proposed in this thesis to better quantify the crash potential of bicyclists. The proposed bicycle measure demonstrates its superiority over conventional exposure surrogates (e.g., population). Although the public bicycle rental system covers the vast majority of bicycle trips, exposure measurements are restricted to bicycle trips and bicycle time travelled (BTT). To this end, a weighted shortest path approach (WSPM) is proposed to estimate the bicycle distance travelled (BDT), taking into account the configuration of the cycle lane network and the safety perception of bicyclists. Bicycle crash frequency models that incorporate BDTs using WSPM exhibit a superior model fit than their counterparts, i.e., bicycle trips and BTT.

(3) Risk factors to different bicycle crash types

A multivariate Poisson-lognormal regression model is constructed to account for the correlation between the frequencies of various types of bicycle crashes. Results indicate that the proposed multivariate Poisson-lognormal model outperforms conventional univariate Poisson-lognormal models. In addition, the differing effects of risk factors on various types of bicycle crashes are identified. For instance, the effects of traffic flow, residential area, network connectivity, and intersection density on crash counts vary between bicycle-vehicle and bicycle-bicycle crashes.

(4) Excessive zero observations in crash frequency model

An advanced deep learning approach, the augmented variational autoencoder method, is proposed to resolve the imbalanced crash data with excessive zero observations. The findings reveal that the proposed method can effectively resolve the challenge of imbalance in heterogeneous data sets. In comparison to other conventional approaches, the proposed method has two advantages: first, it can factorise a unified probability density function to accommodate multiple data types, i.e., interval, nominal, and ordinal, in the data generation process; and second, it can generate an intermediate representation for diverse combinations of zero and non-zero crash cases when formulating the objective function.

(5) Boundary crash problem in crash frequency model

An advanced assignment method, the crash feature-based allocation method, is suggested by considering individual crash characteristics in boundary crash allocation to resolve the boundary crash problem for the macro-level crash frequency model. High matching percentages of boundary crash allocation are achievable, according to the results. In addition, prediction performances of the crash frequency models based on the proposed method, in terms of the RMSE and MAE, are superior to those of the models based on conventional assignment approaches. Furthermore, additional significant risk factors can be identified.

This thesis is able to make contributions to vulnerable road user (i.e. bicyclists) management and educational strategies, traffic controls, bicycle infrastructures, and enforcement strategies that can enhance the safety awareness of bicyclists and reduce their long-term crash risk based on the results of the proposed research questions. Following are some of the potential ramifications of the aforementioned findings. For instance, (i) engineering solutions such as bicycle warning signs and road markings can be put into place in policy intervention areas considering the rise in cycling activity. They not only boost the efficiency of traffic flow, but also enhance bicycle safety overall. (ii) Better design of Cycle Superhighways and cycle lanes, such as physical separation between bicycle lanes and (motor vehicle) traffic lanes, would be required to optimise bicycle safety. (iii) Effective education and promotion strategies that can enhance the

safety perception and awareness of vulnerable groups should be bolstered in light of the growing ageing problem in society. (iv) Accident-prone locations, such as railway stations and intersections, can be the target of enforcement measures and awareness programmes. Consequently, road safety awareness and adherence to traffic regulations can be improved. (v) Effects of risk factors vary across different bicycle crashes. For instance, traffic volume and proportion of residential areas are inversely related to bicycle-vehicle crashes only. Correspondingly, local area traffic management, such as traffic calming, low-speed zones, and cycle priority traffic signals, can be implemented to reduce bicycle-vehicle conflicts and the associated crash risk.

11.3 Limitations

Despite the contributions to the literature described in the above paragraphs, this research should be interpreted in the context of the limitations. Firstly, in the causal inference analysis, the conditional independence assumption (CIA) may not hold if unobserved factors that may influence the outcomes are not included. When more extensive pre-treatment data are available, the difference-in-difference (DID) method based on the propensity score matching approach (PSM) can be applied. Given that differences in their effects on outcomes are time-invariant, the DID matching estimator can reduce the bias arising from differences between treatment and control groups. Moreover, the PSM method utilised in this thesis is incapable of adjusting for spatial correlation effects. This may have an impact on the estimation of policy intervention effects.

Secondly, the bicycle exposure adopted in this thesis is limited to the usage data of the London public bicycle rental system, i.e., LCH. Despite the fact that the bike system constitutes over 70 % of bicycle trips in the study area, parameter estimation results may be subject to bias due to possible differences in the behaviours between different bicyclist groups. Furthermore, we only consider the impact of Cycle Superhighways and cycleways on bicycle route selection. Indeed, perceptions of safety can vary between individuals and travel purposes. When necessary data is available, it is worthwhile to investigate the factors, such as bicycle infrastructure, weather conditions, traffic volume, and bicyclists' perceptions of safety, that may influence bicycle exposure.

There are also limitations in the proposed models. The proposed multivariate model, for instance, only discusses bicycle-vehicle and bicycle-bicycle crashes. When more comprehensive crash data becomes available in the future, risk factors for bicycle-pedestrian and bicycle-only crashes should be further investigated. In the proposed augmented variational autoencoder, each latent variable is modelled independently, and correlations between variables are not considered in the data synthesis procedure. Future research may employ a hierarchical paradigm to address this issue. For the proposed boundary crash allocation method, the mask rate for the proposed augmented masked autoencoder method in this study, which is derived from previous studies, is assumed to be 0.5. It is worthwhile to estimate the optimal mask rate that can optimise individual crash allocation.

11.4 Recommendations for future research

Section 11.1 and 11.2 has outlined the contributions of this thesis regarding the bicycle travel and safety. Nonetheless, the current work can be expanded in the future. The following are the recommendations for future research in three aspects.

11.4.1 Dynamic effects of policy intervention

It is rare that the variation in the effect of policy interventions is examined. The magnitude of an intervention effect can vary over time and space (Gao and Lee, 2019). Admittedly, it takes time for the public to recognise and adjust to a new infrastructure, transportation service, or traffic control and management policy. Likewise, policy intervention may have spatial spillover effects on travel behaviour and related performance attributes. In light of the above, it is critical to study the dynamic effects of policy interventions on bicycle travel and safety when more detailed data regarding time and space are available.

11.4.2 Perception survey

The safety perception of bicyclists may also exert considerable impacts on road safety. When more detailed knowledge of the psychological and physiological information of bicyclists is available, it would be advisable to study the possible impacts of latent characteristics on the likelihood of bicycle crashes in addition to a few demographic and socioeconomic factors. For instance, a stated preference survey can be created to gain knowledge regarding the cyclists' attitudes toward transportation infrastructures, modes of transportation, and traffic features. Long-term crash risk reduction for bicyclists can be greatly facilitated by enforcement and strategies that create a safe or desirable environment for bicyclists.

11.4.3 Multilevel modelling of bicycle crash

The current study focuses on the relationship between bicycle crash frequency, exposure and possible risk factors at the zonal level. Crash frequency models at the microscopic level are required to reveal the effects of road geometries and designs on cycling safety. Both individual and regional characteristics are important predictors of bicycle crash frequency. Evidently, risk factors at different levels may result in counteracting influences (Tseloni et al., 2002; Tin Tin et al., 2013). Therefore, it will be beneficial to conduct multilevel bicycle crash modelling. Additionally, the effect of route uncertainty on the association measure can be incremental. In a future study, it would be worthwhile to explore how the variation in route choice affects the space-time development of bicycle trip distance and, as a result, the exposure of cyclists to the risk of crashes at the microscopic level, e.g. road segment.

11.4.4 Interactions between risk factors

Interaction is present when two or more objects affect one another. In a statistical model, interaction is a term in which the effect of two (or more) variables is not additive. In other words, the effect of factor A plus the effect of factor B is different from the effect of factors A and B combined. It is essential to account for such interaction effects to avoid

poor model performance. In the future study, it is worth assessing the mediated effects of some of the covariates on bicycle crash risk through their effects on bike exposure.

References

Abdel-Aty, M., Siddiqui, C., Huang, H., Wang, X., 2011. Integrating trip and roadway characteristics to manage safety in traffic analysis zones. *Transportation Research Record*, 2213(1), 20-28.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M. F., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record*, 1897(1), 88-95.

Alkahtani, K., Abdel-Aty, M., Lee, J., 2018. A zonal level safety investigation of pedestrian crashes in Riyadh, Saudi Arabia. *International Journal of Sustainable Transportation*, 13 (4), 255-267.

Amoh-Gyimah, R., Saberi, M., Sarvi, M., 2016. Macroscopic modelling of pedestrian and bicycle crashes: A cross-comparison of estimation methods. *Accident Analysis & Prevention*, 93, 147-159.

Aultman-Hall, L., Kaltenecker, M. G., 1999. Toronto bicycle commuter safety rates. *Accident Analysis & Prevention*, 31(6), 675-686.

Bacchieri, G., Barros, A. J., Dos Santos, J. V., Gigante, D. P., 2010. Cycling to work in Brazil: Users profile, risk behaviors, and traffic accident occurrence. *Accident Analysis & Prevention*, 42(4), 1025-1030.

Bakó, B., Berezvai, Z., Isztin, P., Vigh, E. Z., 2020. Does Uber affect bicycle-sharing usage? Evidence from a natural experiment in Budapest. *Transportation Research Part A: Policy and Practice*, 133, 290-302.

Bao, J., He, T., Ruan, S., Li, Y., Zheng, Y., 2017. Planning bike lanes based on sharing-bikes' trajectories. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Canada, 1377-1386.

Barnes, G., Thompson, K., Krizek, K., 2006. A longitudinal analysis of the effect of bicycle facilities on commute mode share. Transportation Research Board 85th Annual Meeting, 22-26 January, Washington DC, United States.

Basso, F., Basso, L. J., Bravo, F., Pezoa, R., 2018. Real-time crash prediction in an urban expressway using disaggregated data. *Transportation Research Part C: Emerging Technologies*, 86, 202-219.

Beck, B., Stevenson, M., et al., 2016. Bicycling crash characteristics: An in-depth crash investigation study. *Accident Analysis & Prevention*, 96, 219-227.

Beck, L. F., Dellinger, A. M., O'neil, M. E., 2007. Motor vehicle crash injury rates by mode of travel, United States: using exposure-based methods to quantify differences. *American Journal of Epidemiology*, 166(2), 212-218.

Behnood, A., Mannering, F., 2017. The effect of passengers on driver-injury severities in single-vehicle crashes: A random parameter heterogeneity-in-means approach. *Analytic Methods in Accident Research*, 14, 41-53.

Bhowmik, M., Rahman, S., Yasmin, N., Eluru., 2019. Alternative model structures for multivariate crash frequency analysis: comparing simulation-based multivariate model with copula-based multivariate model. Transportation Research Board 98th Annual Meeting, 13-17 January, Washington DC, United States.

Blaizot, S., Papon, F., Haddak, M. M., Amoros, E., 2013. Injury incidence rates of cyclists compared to pedestrians, car occupants and powered two-wheeler riders, using a medical registry and mobility data, Rhône County, France. *Accident Analysis & Prevention*, 58, 35-45.

Boquet, G., Morell, A., Serrano, J., Vicario, J. L., 2020. A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection. *Transportation Research Part C: Emerging Technologies*, 115, 102622.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. Proceedings of 19th International Conference on Computational Statistics, Paris, France, August.

Boulange, C., Gunn, L., Giles-Corti, B., et al., 2017. Examining associations between urban design attributes and transport mode choice for walking, cycling, public transport and private motor vehicle trips. *Journal of Transport & Health*, 6, 155-166.

Broach, J., Dill, J., Gliebe, J., 2012. Where do cyclists ride? A route choice model developed with revealed preference GPS data. *Transportation Research Part A: Policy and Practice*, 46(10), 1730-1740.

Brookhart, M. A., Schneeweiss, S., Rothman, K., Glynn, R., Avorn, J., Stürmer, T., 2006. Variable selection for propensity score models. *American Journal of Epidemiology*, 163, 1149-1156.

Brooks, S. P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4), 434-455.

Browne, M., Allen, J., Anderson, S., 2005. Low emission zones: the likely effects on the freight transport sector. *International Journal of Logistics*, 8, 269-281.

Cai, Q., Abdel-Aty, M., Yuan, J., Lee, J., Wu, Y., 2020. Real-time crash prediction on expressways using deep generative models. *Transportation Research Part C: Emerging Technologies*, 117, 102697.

Caliendo, C., De Guglielmo, M. L., Russo, I., 2019. Analysis of crash frequency in motorway tunnels based on a correlated random-parameters approach. *Tunneling and Underground Space Technology*, 85, 243-251.

Caliendo, M., Kopeinig, S., 2005. Some Practical Guidance for the Implementation of Propensity Score Matching. IZA Discussion Paper No. 1588, Bonn, Germany.

Campbell, A. A., Cherry, C. R., Ryerson, M. S., Yang, X., 2016. Factors influencing the choice of shared bicycles and shared electric bikes in Beijing. *Transportation Research Part C: Emerging Technologies*, 67, 399-414.

Casello, J. M., Usyukov, V., 2014. Modeling cyclists' route choice based on GPS data. *Transportation Research Record*, 2430 (1), 155-161.

Cervero, R., Sarmiento, O. L., Jacoby, E., Gomez, L. F., Neiman, A., 2009. Influences of built environments on walking and cycling: lessons from Bogotá. *International Journal of Sustainable Transportation*, 3, 203-226.

Chawla, N. V., Japkowicz, N., Kotcz, A., 2004. Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1), 1-6.

Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. P., 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

Chen, C., Li, K., Teo, S. G., Zou, X., Wang, K., Wang, J., Zeng, Z., 2019. Gated residual recurrent graph neural networks for traffic prediction. Proceedings of the AAAI Conference on Artificial Intelligence, USA, 33, 485-492.

Chen, F., Chen, S., Ma, X., 2018. Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data. *Journal of Safety Research*, 65, 153-159.

Chen, J., Li, Z., Wang, W., Jiang, H., 2016. Evaluating bicycle–vehicle conflicts and delays on urban streets with bike lane and on-street parking. *Transportation Letters*, 10 (1), 1-11.

Chen, L., Chen, C., Srinivasan, R., McKnight, C. E., Ewing, R., Roe, M., 2012. Evaluating the safety effects of bicycle lanes in New York City. *American Journal of Public Health*, 102(6), 1120-1127.

Chen, P. C., Hsieh, H. Y., Su, K. W., Sigalingging, X. K., Chen, Y. R., Leu, J. S., 2020a. Predicting station level demand in a bike-sharing system using recurrent neural networks. *IET Intelligent Transport Systems*, 14, 554-561.

Chen, P., 2015. Built environment factors in explaining the automobile-involved bicycle crash frequencies: A spatial statistic approach. *Safety Science*, 79, 336-343.

Chen, P., Shen, Q., Childress, S., 2017. A GPS data-based analysis of built environment influences on bicyclist route preferences. *International Journal of Sustainable Transportation*, 12, 218-231.

Chen, T., Sze, N. N., Chen, S., Labi, S., Zeng, Q., 2021. Analysing the main and interaction effects of commercial vehicle mix and roadway attributes on crash rates using a Bayesian random-parameter Tobit model. *Accident Analysis & Prevention*, 154, 106089.

Chen, T., Sze, N. N., Saxena, S., Pinjari, A. R., Bhat, C. R., Bai, L., 2020b. Evaluation of penalty and enforcement strategies to combat speeding offences among professional drivers: A Hong Kong stated preference experiment. *Accident Analysis & Prevention*, 135, 105366.

Chipman, M. L., MacGregor, C. G., Smiley, A. M., Lee-Gosselin, M., 1993. The role of exposure in comparisons of crash risk among different drivers and driving environments. *Accident Analysis & Prevention*, 25(2), 207-211.

Chong, S. L., Tyebally, A., Chew, S. Y., Lim, Y. C., Feng, X. Y., Chin, S. T., Lee, L. K., 2017. Road traffic injuries among children and adolescents in Singapore–Who is at greatest risk?. *Accident Analysis & Prevention*, 100, 59-64.

Corcoran, J., Li, T., Rohde, D., Charles-Edwards, E., Mateo-Babiano, D., 2014. Spatio-temporal patterns of a Public Bicycle Sharing Program: the effect of weather and calendar events. *Journal of Transport Geography*, 41, 292-305.

Cottrill, C. D., Thakuriah, P. V., 2010. Evaluating pedestrian crashes in areas with high low-income or minority populations. *Accident Analysis & Prevention*, 42(6), 1718-1728.

Cui, G., Wang, X., Kwon, D., 2015. A framework of boundary collision data aggregation into neighborhoods. *Accident Analysis & Prevention*, 83, 1-17.

Davis, R. M., Pless, B., 2001. BMJ bans "accidents": Accidents are not unpredictable. *British Medical Journal*, 322(7298), 1320-1321.

De Geus, B., Vandenbulcke, G., et al., 2012. A prospective cohort study on minor accidents involving commuter cyclists in Belgium. *Accident Analysis & Prevention*, 45, 683-693.

De Rome, L., Boufous, S., et al., 2014. Bicycle crashes in different riding environments in the Australian capital territory. *Traffic injury prevention*, 15(1), 81-88.

Deliali, A., Campbell, N., Knodler, M., Christofa, E., 2020. Understanding the safety impact of protected intersection design elements: a driving simulation approach. *Transportation Research Record*, 2674(3), 179-188.

Deng, Y., Chen, Y., Zhang, Y., Mahadevan, S., 2012. Fuzzy Dijkstra algorithm for shortest path problem under uncertain environment. *Applied Soft Computing*, 12 (3), 1231-1237.

Ding, H., Guo, Y., Sze, N.N*., 2022a. Effect of the Ultra-Low Emission Zone on the usage of the London Cycle Hire Scheme. *Transportation Letters*, in press.

Ding, H., Lu, Y., Sze, N. N., Chen, T., Guo, Y., Lin, Q., 2022b. A deep generative approach for crash frequency model with heterogeneous imbalanced data. *Analytic Methods in Accident Research*, 34, 100212.

Ding, H., Sze, N. N., 2022. Effects of road network characteristics on bicycle safety: A multivariate Poisson-lognormal model. *Multimodal Transportation*, 1(2), 100020.

Ding, H., Sze, N. N., Li, H., Guo, Y., 2020. Roles of infrastructure and land use in bicycle crash exposure and frequency: a case study using Greater London bike sharing data. *Accident Analysis & Prevention*, 144, 105652.

Ding, H., Sze, N. N., Li, H., Guo, Y., 2021a. Effect of London cycle hire scheme on bicycle safety. *Travel Behaviour and Society*, 22, 227-235.

Ding, H., Sze, N. N., Li, H., Guo, Y., 2021b. Affected area and residual period of London Congestion Charging scheme on road safety. *Transport Policy*, 100, 120-128.

Ding, H., Sze, N. N., Guo, Y., Li, H., 2021c. Role of exposure in bicycle safety analysis: Effect of cycle path choice. *Accident Analysis & Prevention*, 153, 106014.

Dixit, V. V., Pande, A., Abdel-Aty, M., Das, A., Radwan, E., 2011. Quality of traffic flow on urban arterial streets and its relationship with safety. *Accident Analysis & Prevention*, 43(5), 1610-1616.

Dong, N., Huang, H., Xu, P., Ding, Z., Wang, D., 2014. Evaluating spatial-proximity structures in crash prediction models at the level of traffic analysis zones. *Transportation Research Record*, 2432(1), 46-52.

Dong, N., Huang, H., Zheng, L., 2015. Support vector machine in crash prediction at the level of traffic analysis zones: assessing the spatial proximity effects. *Accident Analysis & Prevention*, 82, 192-198.

Dosovitskiy, A., Beyer, L., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010. 11929.

Dumbaugh, E., Rae, R., 2009. Safe urban form: Revisiting the relationship between community design and traffic safety. *Journal of the American Planning Association*, 75(3), 309-329.

Ehrgott, M., Wang, J. Y., Raith, A., Van Houtte, C., 2012. A bi-objective cyclist route choice model. *Transportation Research Part A: Policy and Practice*, 46(4), 652-663.

Elamrani, A., Mousannif, H., Moatassime, H., 2020. A real-time crash prediction fusion framework: An imbalance aware strategy for collision avoidance systems. *Transportation Research Part C: Emerging Technologies*, 118, 102708.

El-Assi, W., Salah Mahmoud, M., Nurul Habib, K., 2017. Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. *Transportation*, 44(3), 589-613.

El-Basyouny, K., Sayed, T., 2009. Collision prediction models using multivariate Poisson-lognormal regression. *Accident Analysis & Prevention*, 41(4), 820-828.

El Esawey, M., Lim, C., & Sayed, T., 2015. Development of a cycling data model: City of Vancouver case study. *Canadian journal of civil engineering*, 42(12), 1000-1010.

Ellison, R., Greaves, S., Hensher, D., 2013. Five years of London's low emission zone: Effects on vehicle fleet composition and air quality. *Transportation Research Part D: Transport and Environment*, 23, 25-33.

Elvik, R., T. Vaa., 2004. The Handbook of Road Safety Measures. Elsevier, London, 2004.

Eren, E., Uz, V. E., 2020. A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable Cities and Society*, 54, 101882.

Ermagun, A., Lindsey, G., Loh, T. H., 2018. Bicycle, pedestrian, and mixed-mode trail traffic: A performance assessment of demand models. *Landscape and Urban Planning*, 177, 92-102.

Esiyok, B., Korkusuz, I., Canturk, G., Alkan, H. A., Karaman, A. G., Hamit Hanci, I., 2005. Road traffic accidents and disability: A cross-section study from Turkey. *Disability and Rehabilitation*, 27(21), 1333-1338.

European Cyclists' Federation (ECF)., 2014. Fast Cycling Routes: Towards Barrier-free Commuting. Available at: https://ecf.com/what-we-do/urban-mobility/fast-cycling-routes

Fabozzi, F. J., Focardi, S. M., Rachev, S. T., Arshanapalli, B. G., 2014. The Basic of Financial Econometrics: Tools, Concepts, and Asset Management Applications. John Wiley & Sons, Inc., United States.

Faghih-Imani, A., Eluru, N., El-Geneidy, A. M., Rabbat, M., Haq, U., 2014. How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal. *Journal of Transport Geography*, 41, 306-314.

Feng, S., Chen, H., Du, C., Li, J., Jing, N., 2018. A hierarchical demand prediction method with station clustering for bike sharing system. Proceedings of the IEEE 3rd International Conference on Data Science in Cyberspace (DSC), China, 829-836.

Fishman, E., Schepers, P., 2016. Global bike share: What the data tells us about road safety. *Journal of Safety Research*, 56, 41-45.

Fishman, E., Schepers, P., 2018. The Safety of Bike Share Systems. ITF Discussion Papers, International Transport Forum, Paris.

Fishman, E., Washington, S., Haworth, N., 2014. Bike share's impact on car use: Evidence from the United States, Great Britain, and Australia. *Transportation Research Part D: Transport and Environment*, 31, 13-20.

Fournier, N., Christofa, E., Knodler Jr, M. A., 2017. A sinusoidal model for seasonal bicycle demand estimation. *Transportation Research Part D: Transport and Environment*, 50, 154-169.

Fuglede, B., Topsoe, F., 2004. Jensen-Shannon Divergence and Hilbert space embedding. In: Proceedings of International Symposium on Information Theory, IEEE, Chicago, United States, June-July.

Gaither, J. R., Gordon, K., et al., 2018. Racial disparities in discontinuation of long-term opioid therapy following illicit drug use among black and white patients. *Drug and Alcohol Dependence*, 192, 371-376.

Gao, X., Lee, G. M., 2019. Moment-based rental prediction for bicycle-sharing transportation systems using a hybrid genetic algorithm and machine learning. *Computers & Industrial Engineering*, 128, 60-69.

García-Palomares, J. C., Gutiérrez, J., Latorre, M., 2012. Optimizing the location of stations in bike-sharing programs: A GIS approach. *Applied Geography*, 35, 235-246.

Gebhart, K., Noland, R. B., 2014. The impact of weather conditions on bike-share trips in Washington, DC. *Transportation*, 41, 1205-1225.

Gehrsitz, M., 2017. The effect of low emission zones on air pollution and infant health. *Journal of Environmental Economics and Management*, 83, 121-144.

Ghekiere, A., Van Cauwenberg, J., et al., 2014. Critical environmental factors for transportation cycling in children: a qualitative study using bike-along interviews. *PloS one*, 9(9), e106696.

Giot, R., Cherrier, R., 2014. Predicting bike-share system usage up to one day ahead. Proceedings of the IEEE symposium on computational intelligence in vehicles and transportation systems, USA, 22-29.

Golob, T. F., Recker, W. W., 2001. Relationships among urban freeway accidents, traffic flow, weather and lighting conditions. *Journal of Transportation Engineering*, 129, 342-353.

González, F., Melo-Riquelme, C., de Grange, L., 2016. A combined destination and route choice model for a bicycle sharing system. *Transportation*, 43(3), 407-423.

Gooch, J. P., Gayah, V. V., Donnell, E. T., 2018. Safety performance functions for horizontal curves and 16 tangents on two lanes, two-way rural roads. *Accident Analysis & Prevention*, 120, 28-37.

Goodfellow, I., Pouget-Abadie, J., et al., 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2672-2680.

Götschi, T., Marko T., Neil, M., Tim, S., Anna, G., James W., 2015. Contrasts in active transport behaviour across four countries: How do they translate into public health benefits? *Preventive Medicine*,74, 42–48.

Greater London Authority, GLA., 2019. Central London Ultra Low Emission Zone-Four Month Report (access on September 2019). https://www.london.gov.uk/sites/default/files/central_london_ulez_4_month_report.pdf.

Green Car Congress., 2020. Survey finds 65% of Londoners changed their usual mode of transport for the ULEZ. https://www.greencarcongress.com/2020/05/20200510-ulez.html.

Green, C. P., Heywood, J. S., Navarro, M., 2016. Traffic accidents and the London congestion charge. *Journal of Public Economics*, 133, 11-22.

Gu, T., Kim, I., Currie, G., 2019. Measuring immediate impacts of a new mass transit system on an existing bike-share system in China. *Transportation Research Part A: Policy and Practice*, 124, 20-39.

Guerra, E., Zhang, H., Hassall, L., Wang, J., Cheyette, A., 2020. Who cycles to work and where? A comparative multilevel analysis of urban commuters in the US and Mexico. *Transportation Research Part D: Transport and Environment*, 87, 102554.

Guo, Y., Li, Z., Liu, P., Wu, Y., 2019. Exploring risk factors with crashes by collision type at freeway diverge areas: accounting for unobserved heterogeneity. *IEEE Access*, 7, 11809-11819.

Guo, Y., Li, Z., Wu, Y., Xu, C., 2018b. Exploring unobserved heterogeneity in bicyclists' red-light running behaviours at different crossing facilities. *Accident Analysis & Prevention*, 115, 118-127.

Guo, Y., Liu, P., Wu, Y., Chen, J., 2020. Evaluating how right-turn treatments affect right-turn-on-red conflicts at signalized intersections. *Journal of Transportation Safety & Security*, 12(3), 419-440.

Guo, Y., Osama, A., Sayed, T., 2018a. A cross-comparison of different techniques for modelling macro-level cyclist crashes. *Accident Analysis & Prevention*, 113, 38-46.

Gutiérrez, M., Hurtubia, R., de Dios Ortúzar, J., 2020. The role of habit and the built environment in the willingness to commute by bicycle. *Travel Behaviour and Society*, 20, 62-73.

Hamann, C., Peek-Asa, C., 2013. On-road bicycle facilities and bicycle crashes in Iowa, 2007–2010. *Accident Analysis & Prevention*, 56, 103-109.

Handy, S. L., 2011. The Davis Bicycle Studies: Why Do I Bicycle But My Neighbour Doesn't?. *Access Magazine*, 1, 16-21.

He, H., Garcia, E. A., 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284.

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., Girshick, R., 2021. Masked Autoencoders Are Scalable Vision Learners. arXiv preprint arXiv:2111.06377.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 27 June – 30 June, Las Vegas, United States, 770-778.

Heesch, K. C., Sahlqvist, S., Garrard, J., 2012. Gender differences in recreational and transport cycling: a cross-sectional mixed-methods comparison of cycling patterns, motivators, and constraints. *International Journal of Behavioural Nutrition and Physical Activity*, 9, 1-12.

Heinen, E., Van Wee, B., Maat, K., 2010. Commuting by bicycle: an overview of the literature. *Transport Reviews*, 30, 59-96.

Heinrich, C., Maffioli, A., Vazquez, G., 2010. A primer for applying propensity-score matching. *Spd Working Papers*, 59, 147-164.

Hess, A. K., Schubert, I., 2019. Functional perceptions, barriers, and demographics concerning e-cargo bike sharing in Switzerland. *Transportation Research Part D: Transport and Environment*, 71, 153-168.

Hijar, M., Vázquez-Vela, E., Arreola-Risa, C., 2003. Pedestrian traffic injuries in Mexico: a country update. *Injury Control and Safety Promotion*, 10(1-2), 37-43.

Hilbe, J.M., 2011. Negative binomial regression. Cambridge University Press.

Hillier, B., 1996. Space is the Machine: A Configurational Theory of Architecture. Cambridge University Press, United Kingdom.

Hillier, B., Hanson, J., 1984. The Social Logic of Space. Cambridge University Press, United Kingdom (1984).

Hoffman, M. R., Lambert, W. E., Peck, E. G., Mayberry, J. C., 2010. Bicycle commuter injury prevention: it is time to focus on the environment. *Journal of Trauma and Acute Care Surgery*, 69(5), 1112-1119.

Hopkinson, P., Wardman, M., 1996. Evaluating the demand for new cycle facilities. *Transport Policy*, 3(4), 241-249.

Howard, C., Burns, E.K., 2001. Cycling to work in phoenix: route choice, travel behaviour and commuter characteristics. *Transportation Research Record*, 1773 (1), 39-46.

Huang, D., Liu, Y., Wang, M., Yang, H., Huang, Q., Li, C., 2020. How to promote users' adoption behaviour of dockless bike-sharing? An empirical study based on extended norms activation theory. *Transportation Letter*, 12, 638-648.

Huang, H., Chin, H. C., Haque, M. M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. *Accident Analysis & Prevention*, 40(1), 45-54.

Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., Abdel-Aty, M., 2016. Macro and micro models for zonal crash prediction with application in hot zones identification. *Journal of Transport Geography*, 54, 248-256.

Hung, K., Huyen, L., 2011. Education influence in traffic safety: A case study in Vietnam. *IATSS Research*, 34, 87-93.

Huo, X., Leng, J., Hou, Q., Zheng, L., Zhao, L., 2020. Assessing the explanatory and predictive performance of a random parameters count model with heterogeneity in means and variances. *Accident Analysis & Prevention*, 147, 105759.

Hyatt, E., Griffin, R., Rue III, L. W., McGwin Jr, G., 2009. The association between price of regular-grade gasoline and injury and mortality rates among occupants involved in motorcycle-and automobile-related motor vehicle collisions. *Accident Analysis & Prevention*, 41(5), 1075-1079.

Islam, Z., Abdel-Aty, M., Cai, Q., Yuan, J., 2021. Crash data augmentation using variational autoencoder. *Accident Analysis & Prevention*, 151, 105950.

ITV, 2014 ITV, 2014. Boris Bikes extended to south-west London. http://www.itv.com/news/london/2013-04-04/boris-bikes-extended-to-south-west-london/.

Ivan, J. N., Deng, Z., Jonsson, T., 2006. Procedure for allocating zonal attributes to link network in GIS environment. Transportation Research Board 85th Annual Meeting, 22-26 January, Washington DC, United States.

Jappinen, S., Toivonen, T., Salonen, M., 2013. Modelling the potential effect of shared bicycles on public transport travel times in Greater Helsinki: An open data approach. *Applied Geography*, 43, 13-24.

Jensen, S. U., 2007. Pedestrian and bicyclist level of service on roadway segments. *Transportation Research Record*, 2031(1), 43-51.

Johnson, F. X., Silveira, S., 2014. Pioneer countries in the transition to alternative transport fuels: comparison of ethanol programmes and policies in Brazil, Malawi and Sweden. *Environmental Innovation and Societal Transitions*, 11, 1-24.

Johnson, J. M., Khoshgoftaar, T. M., 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 1-54.

Johnson, M., Charlton, J., Oxley, J., Newstead, S., 2010. Naturalistic cycling study: identifying risk factors for on-road commuter cyclists. *Association for the Advancement of Automotive Medicine*, 54, 275.

Jones, A. P., Haynes, R., Kennedy, V., Harvey, I. M., Jewell, T., Lea, D., 2008. Geographical variations in mortality and morbidity from road traffic accidents in England and Wales. *Health & Place*, 14(3), 519-535.

Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., Banchs, R., 2010. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6, 455-466.

Kamel, M., Sayed, T., 2021. The impact of bike network indicators on bike kilometers travelled and bike safety: A network theory approach. *Environment and Planning B: Urban Analytics and City Science*, 48 (7), 2055-2072.

Karlaftis, M. G., Vlahogianni, E. I., 2011. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19, 387-399.

Kassim, A., Ismail, K., Hassan, Y., 2014. Automated measuring of cyclist-motor vehicle post encroachment time at signalized intersections. *Canadian Journal of Civil Engineering*, 41(7), 605-614.

Ke, J., Qin, X., Yang, H., Zheng, Z., Zhu, Z., Ye, J., 2021. Predicting origin-destination ride-sourcing demand with a spatio-temporal encoder-decoder residual multi-graph convolutional network. *Transportation Research Part C: Emerging Technologies*, 122, 102858.

Kelly, F. J., Kelly, J., 2009. London air quality: a real world experiment in progress. *Biomarkers*, 14, 5-11.

Kerr, J., Emond, JA., Badland, H., et al., 2016. Perceived neighbourhood environmental attributes associated with walking and cycling for transport among adult residents of 17 cities in 12 countries: The IPEN study. *Environmental Health Perspectives*, 124, 290.

Kim, D., Shin, H., Im, H., Park, J., 2012. Factors influencing travel behaviours in bike-sharing. Transportation Research Board 91st Annual Meeting, 22-26 January, Washington DC, United States.

Kim, T. S., Lee, W. K., Sohn, S. Y., 2019. Graph convolutional network approach applied to predict hourly bike-sharing demands considering spatial, temporal, and global effects. *PloS One*, 14, e0220782.

Kingma, D. P., Welling, M., 2013. Auto-Encoding Variational Bayes. Proceedings of the International Conference on Learning Representations (ICLR), Banff, Canada, April.

Kipf, T. N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Lajunen, T., Özkan, T., Porter, B. E., 2016. Bicycle safety. *Transportation Research Part F: Traffic Psychology and Behaviour*, 41, 179-181.

Larsen, J., El-Geneidy, A., 2011. A travel behaviour analysis of urban cycling facilities in Montréal, Canada. *Transportation Research Part D: Transport and Environment*, 16(2), 172-177.

Lee, A. E., Underwood, S., Handy, S., 2015b. Crashes and other safety-related incidents in the formation of attitudes toward bicycling. *Transportation Research Part F: Traffic Psychology and Behaviour*, 28, 14-24.

Lee, J., Abdel-Aty, M., Cai, Q., 2017. Intersection crash prediction modelling with macro-level data from various geographic units. *Accident Analysis & Prevention*, 102, 213-226.

Lee, J., Abdel-Aty, M., Choi, K., Huang, H., 2015a. Multi-level hot zone identification for pedestrian safety. *Accident Analysis & Prevention*, 76, 64-73.

Lee, J., Abdel-Aty, M., Jian, X., 2014. Development of zone system for macro-level traffic safety analysis. *Journal of Transport Geography*, 38, 13-21.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. *Accident analysis & prevention*, 34(2), 149-161.

Leuven, E., Sianesi, B., 2003. PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Version: 4.0.4 10 nov2010. Available: http://ideas.repec.org/c/boc/bocode/s432?001.html. Accessed 2014 Feb 09.

Li, H., Ding, H., Ren, G., Xu, C., 2018. Effects of the London Cycle Superhighways on the usage of the London Cycle Hire. *Transportation Research Part A: Policy and Practice*, 111, 304-315.

Li, H., Graham, D.J., 2016. Quantifying the causal effects of 20 mph zones on road casualties in London via doubly robust estimation. *Accident Analysis & Prevention*, 126, 65-74.

Li, H., Graham, D. J., Liu, P., 2017. Safety effects of the London cycle superhighways on cycle collisions. *Accident Analysis & Prevention*, 99, 90-101.

Li, H., Graham, D. J., Majumdar, A., 2012. The effects of congestion charging on road traffic casualties: A causal analysis using difference-in-difference estimation. *Accident Analysis & Prevention*, 49, 366-377.

Li, H., Zhang, Y., Ding, H., Ren, G., 2019. Effects of dockless bike-sharing systems on the usage of the London Cycle Hire. *Transportation Research Part A: Policy and Practice*, 130, 398-411.

Li, P., Abdel-Aty, M., Yuan, J., 2020. Real-time crash risk prediction on arterials based on LSTM-CNN. *Accident Analysis & Prevention*, 135, 105371.

Li, Y., Zheng, Y., Zhang, H., Chen, L., 2015. Traffic prediction in a bike-sharing system. Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Washington, 1-10.

Lin, D., Zhang, Y., Zhu, R., Meng, L., 2019. The analysis of catchment areas of metro stations using trajectory data generated by dockless shared bikes. *Sustainable Cities and Society*, 49, 101598.

Lin, J., 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37 (1), 145-151.

Lin, L., He, Z., Peeta, S., 2018. Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. *Transportation Research Part C: Emerging Technologies*, 97, 258-276.

Liu, J., Sun, L., Chen, W., Xiong, H., 2016. Rebalancing bike sharing systems: A multi-source data smart optimization. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, 1005-1014.

Liu, X., Chen, Y., 2010. Application of Dijkstra algorithm in logistics distribution lines. Proceedings of the 3rd International Symposium on Computer Science and Computational Technology (ISCSCT), Jiaozuo, China, page 48-50.

Ljubenkov, D., Kon, F., Ratti, C., 2020. Optimizing Bike Sharing System Flows Using Graph Mining, Convolutional and Recurrent Neural Networks. Proceedings of the IEEE European Technology and Engineering Management Summit (E-TEMS), Germany, 1-6.

Lord, D., Manar, A., Vizioli, A., 2005. Modelling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. *Accident Analysis & Prevention*, 37(1), 185-199.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5), 291-305.

Lovegrove, G., Sayed, T., 2006. Macro-level collision prediction models for evaluating neighbourhood traffic safety. *Canadian Journal of Civil Engineering*, 33, 609-621.

Lu, Y., Ding, H., Ji, S., Sze, N.N., He, Z., 2021. Dual attentive graph neural network for metro passenger flow prediction. *Neural Computing and Applications*, 33(20), 13417-13431.

Lu, Y., Wang, W., Hu, X., Xu, P., Zhou, S., Cai, M., 2022. Vehicle Trajectory Prediction in Connected Environments via Heterogeneous Context-Aware Graph Convolutional Networks. *IEEE Transactions on Intelligent Transportation Systems*, in press.

Lubbock, E., 1963. Parliamentary Debates (Hansard). House of Commons.

Lusk, A. C., Furth, P. G., Morency, P., Miranda-Moreno, L. F., Willett, W. C., Dennerlein, J. T., 2011. Risk of injury for bicycling on cycle tracks versus in the street. *Injury Prevention*, 17(2), 131-135.

Ma, J., Kockelman, K. M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis & Prevention*, 40(3), 964-975.

Malyshkina, N. V., Mannering, F. L., 2010. Zero-state Markov switching count-data models: An empirical assessment. *Accident Analysis & Prevention*, 42(1), 122-130.

Mandic, S., Flaherty, C., et al., 2018. Effects of cycle skills training on cycling-related knowledge, confidence and behaviour in adolescent girls. *Journal of Transport & Health*, 9, 253-263.

Mannering, F., 2009. An empirical analysis of driver perceptions of the relationship between speed limits and safety. *Transportation Research Part F: Traffic Psychology and Behaviour*, 12, 99-106.

Mannering, F., 2018. Temporal instability and the analysis of highway accident data. *Analytic Methods in Accident Research*, 17, 1-13.

Mannering, F. L., Bhat, C. R., 2014. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, 1, 1-22.

Mannering, F., Bhat, C. R., Shankar, V., Abdel-Aty, M., 2020. Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic Methods in Accident Research*, 25, 100113.

Mannering, F. L., Shankar, V., Bhat, C., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research*, 11, 1–16.

Margaryan, S., 2021. Low emission zones and population health. *Journal of Health Economics*, 76, 102402.

Marshall, W. E., Garrick, N. W., 2010. Street network types and road safety: A study of 24 California cities. *Urban Design International*, 15(3), 133-147.

Martin, J. L., 2002. Relationship between crash rate and hourly traffic flow on interurban motorways. *Accident Analysis & Prevention*, 34(5), 619-629.

McNeil, N., Broach, J., Dill, J., 2018. Breaking barriers to bike share: Lessons on bike share equity. *Institute of Transportation Engineers*, 88, 31-35.

Meng, F., Sze, N. N., Song, C., Chen, T., Zeng, Y., 2021. Temporal instability of truck volume composition on non-truck-involved crash severity using uncorrelated and correlated grouped random parameters binary logit models with space-time variations. *Analytic Methods in Accident Research*, 31, 100168.

Menghini, G., Carrasco, N., Schüssler, N., Axhausen, K. W., 2010. Route choice of cyclists in Zurich. *Transportation Research Part A: Policy and Practice*, 44(9), 754-765.

Miaou, S. P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention*, 26(4), 471-482.

Midgley, P., 2011. Bicycle-sharing schemes: enhancing sustainable mobility in urban areas. *United Nations, Department of Economic and Social Affairs*, 8, 1-12.

Mindell, J. S., Leslie, D., Wardlaw, M., 2012. Exposure-based,'like-for-like'assessment of road safety by travel mode using routine health data. *PloS one*, 7(12), e50606.

Miranda-Moreno, L. F., Strauss, J., Morency, P., 2011. Disaggregate exposure measures and injury frequency models of cyclist safety at signalized intersections. *Transportation Research Record*, 2236(1), 74-82.

Myhrmann, M. S., Janstrup, K. H., Møller, M., Mabit, S. E., 2021. Factors influencing the injury severity of single-bicycle crashes. *Accident Analysis & Prevention*, 149, 105875.

Narayanamoorthy, S., Paleti, R., Bhat, C. R., 2013. On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. *Transportation Research Part B: Methodological*, 55, 245-264.

National Highway Traffic Safety Administration, NHTSA, 2019. Traffic Safety Facts 2017: A compilation of motor vehicle crash data from the fatality analysis reporting system and the General Estimates System. Washington DC: U. S Department of Transportation.

Nikitas, A., Michalakopoulos, N., Wallgren, P., 2014. Bike-sharing: Is safety an issue adversely affecting its potential for being embraced by urban societies? Proceedings

of the 3rd International Cycling Safety Conference, 18–19 November, Gothenburg, Sweden.

Noland, R. B., Quddus, M. A., 2004. Analysis of pedestrian and bicycle casualties with regional panel data. *Transportation Research Record*, 1897(1), 28-33.

Noland, R. B., Quddus, M. A., Ochieng, W. Y., 2008. The effect of the London congestion charge on road casualties: an intervention analysis. *Transportation*, 35(1), 73-91.

Olesen, A. V., Madsen, T. K. O., Hels, T., Hosseinpour, M., Lahrmann, H. S., 2021. Single-bicycle crashes: An in-depth analysis of self-reported crashes and estimation of attributable hospital cost. *Accident Analysis & Prevention*, 161, 106353.

Osama, A., Sayed, T., 2017. Evaluating the Impact of Connectivity, Continuity, and Topography of Sidewalk Network on Pedestrian Safety. *Accident Analysis & Prevention*, 107, 117-125.

Palamara, P., Broughton, M., 2013. An investigation of pedestrian crashes at traffic intersections in the Perth Central Business. *Journal of Public Health*, 93, 1456–1463.

Pan, Y., Zheng, R. C., Zhang, J., Yao, X., 2019. Predicting bike sharing demand using recurrent neural networks. *Procedia Computer Science*,147, 562-566.

Park, D., Hoshi, Y., Kemp, C. C., 2018. A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3 (3), 1544-1551.

Park, E. S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modelling crash frequency by severity. *Transportation Research Record*, 2019(1), 1-6.

Park, H. C., Park, B-J., Park, P. Y., 2022. A multiple membership multilevel negative binomial model for intersection crash analysis. *Analytic Methods in Accident Research*, in press.

Park, H. C., Yang, S., Park, P. Y., Kim, D. K., 2020. Multiple membership multilevel model to estimate intersection crashes. *Accident Analysis & Prevention*, 144, 105589.

Pei, X., Sze, N. N., Wong, S. C., Yao, D., 2016. Bootstrap resampling approach to disaggregate analysis of road crashes in Hong Kong. *Accident Analysis & Prevention*, 95, 512-520.

Pei, X., Wong, S. C., Sze, N. N., 2012. The roles of exposure and speed in road safety analysis. *Accident analysis & prevention*, 48, 464-471.

Peters, J., Burguillo, M., Arranz, J., 2021. Low emission zones: Effects on alternative-fuel vehicle uptake and fleet CO_2 emissions. *Transportation Research Part D: Transport and Environment*, 95, 102882.

Poulos, R. G., Hatfield, J., Rissel, C., Flack, L. K., Murphy, S., Grzebieta, R., McIntosh, A. S., 2015. An exposure based study of crash and injury rates in a cohort of transport and recreational cyclists in New South Wales, Australia. *Accident Analysis & Prevention*, 78, 29-38.

Prati, G., Fraboni, F., De Angelis, M., Pietrantoni, L., Johnson, D., Shires, J., 2019. Gender differences in cycling patterns and attitudes towards cycling in a sample of European regular cyclists. *Journal of Transport Geography*, 78, 1-7.

Pucher, J., Buehler, R., 2006. Why Canadians cycle more than Americans: a comparative analysis of bicycling trends and policies. *Transport Policy*, 13(3), 265-279.

Pucher, J., Ralph, B, Dafna, M., Adrian, B., 2011. Walking and cycling in the United States, 2001-2009: Evidence from the national household travel surveys. *American Journal of Public Health*, 101, 310–317.

Pulugurtha, S. S., Thakur, V., 2015. Evaluating the effectiveness of on-street bicycle lane and assessing risk to bicyclists in Charlotte, North Carolina. *Accident Analysis & Prevention*, 76, 34-41.

Quddus, M. A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. *Accident Analysis & Prevention*, 40(4), 1486-1497.

Quintero, L., Sayed, T., Wahba, M. M., 2013. Safety models incorporating graph theory based transit indicators. *Accident Analysis & Prevention*, 50, 635-644.

Rayaprolu, H., Llorca, C., Moeckel, R., 2020. Impact of bicycle highways on commuter mode choice: A scenario analysis. *Environment and Planning B: Urban Analytics and City Science*, 47 (4), 662-677.

Razavi, A., van den Oord, A., Vinyals, O., 2019. Generating diverse high-fidelity images with VQ-VAE-2. Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, Canada, December.

Reynolds, C. C., Harris, M. A., Teschke, K., Cripton, P. A., Winters, M., 2009. The impact of transportation infrastructure on bicycling injuries and crashes: a review of the literature. *Environmental Health*, 8(1), 1-19.

Rezende, D. J., Mohamed, S., Wierstra, D., 2014. Stochastic backpropagation and approximate inference in deep generative models. Proceedings of the 31st International Conference on Machine Learning, Beijing, China, June.

Rixey, R. A., 2013. Station-level forecasting of bike sharing ridership: Station network effects in three US systems. *Transportation Research Record*, 2387, 46-55.

Rodgers, G. B., 1995. Bicyclist deaths and fatality risk patterns. *Accident Analysis & Prevention*, 27(2), 215-223.

Romanillos, G., Moya-Gómez, B., Zaltz-Austwick, M., Lamíquiz-Daudén, P. J., 2018. The pulse of the cycling city: visualizing Madrid bike share system GPS routes and cycling flow. *Journal of Maps*, 14, 34-43.

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.

Roshandel, S., Zheng, Z., Washington, S., 2015. Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. *Accident Analysis & Prevention*, 79, 198-211.

Rudloff, C., Lackner, B., 2014. Modelling demand for bike sharing systems: neighbouring stations as source for demand and reason for structural breaks. *Transportation Research Record*, 2430, 1-11.

Ruffieux, S., Spycher, N., Mugellini, E., Abou Khaled, O., 2017. Real-time usage forecasting for bike-sharing systems: A study on random forest and convolutional neural network applicability. Proceedings of the Intelligent Systems Conference, London, 622-631.

Ruiz-Padillo, A., Pasqual, F. M., Uriarte, A. M. L., Cybis, H. B. B., 2018. Application of multi-criteria decision analysis methods for assessing walkability: A case study in Porto Alegre, Brazil. *Transportation Research Part D: Transport and Environment*, 63, 855-871.

Saeed, T. U., Hall, T., Baroud, H., Volovski, M. J., 2019. Analysing road crash frequencies with uncorrelated and correlated random-parameters count models: An empirical assessment of multilane highways. *Analytic Methods in Accident Research*, 23, 100101.

Sami, A., Najafi, A., Yamini, N., Moafian, G., et al., 2013. Educational level and age as contributing factors to road traffic accidents. *Chinese Journal of Traumatology*, 16, 281-285.

Santos, G., Shaffer, B., 2004. Preliminary results of the London congestion charging scheme. *Public Works Management & Policy*, 9, 164-181.

Sathishkumar, V. E., Park, J., Cho, Y., 2020. Using data mining techniques for bike sharing demand prediction in metropolitan city. *Computer Communications*, 153, 353-366.

Sayed, T., Zaki, M. H., Autey, J., 2013. Automated safety diagnosis of vehicle–bicycle interactions using computer vision analysis. *Safety Science*, 59, 163-172.

Schlögl, M., Stütz, R., Laaha, G., Melcher, M., 2019. A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset. *Accident Analysis & Prevention*, 127, 134-149.

Schneider, S., Brümmer, V., Abel, T., Askew, C. D., Strüder, H. K., 2009. Changes in brain cortical activity measured by EEG are related to individual exercise preferences. *Physiology & Behaviour*, 98, 447-452.

Sedeño-noda, A., Colebrook, M., 2019. A bi-objective Dijkstra algorithm. *European Journal of Operational Research*, 276 (1), 106-118.

Sener, I. N., Eluru, N., Bhat, C. R., 2009. An analysis of bicycle route choice preferences in Texas, US. *Transportation*, 36, 511-539.

Shankar, V., Milton, J., Mannering, F., 1997. Modelling accident frequencies as zero-altered probability processes: an empirical inquiry. *Accident Analysis & Prevention*, 29(6), 829-837.

Shankar, V. N., Ulfarsson, G. F., Pendyala, R. M., Nebergall, M. B., 2003. Modeling crashes involving pedestrians and motorized traffic. *Safety Science*, 41(7), 627-640.

Shi, Q., Abdel-Aty, M., 2015. Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies*, 58, 380-394.

Siddiqui, C., Abdel-Aty, M., 2012. Nature of modelling boundary pedestrian crashes at zones. *Transportation Research Record*, 2299, 31-40.

Siddiqui, C., Abdel-Aty, M., Choi, K., 2012. Macroscopic spatial analysis of pedestrian and bicycle crashes. *Accident Analysis & Prevention*, 45, 382-391.

Smith, H. L., 1997. matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology*, 27(1), 325-353.

Sohrabi, S., Paleti, R., Balan, L., Cetin, M., 2020. Real-time prediction of public bike sharing system demand using generalized extreme value count model. *Transportation Research Part A: Policy and Practice*, 133, 325-336.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Van der Linde, A., 2014. The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B*, 76(3), 485-493.

Stinson, M. A., Bhat, C. R., 2003. Commuter bicyclist route choice: Analysis using a stated preference survey. *Transportation Research Record*, 1828(1), 107-115.

Strauss, J., Zangenehpour, S., Miranda-Moreno, L., Saunier, N., 2017. Cyclist deceleration rate as surrogate safety measure in Montreal using smartphone GPS data. *Accident Analysis & Prevention*, 99, 287-296.

Su, J., Sze, N. N., Bai, L., 2021. A joint probability model for pedestrian crashes at macroscopic level: Roles of environment, traffic, and population characteristics. *Accident Analysis & Prevention*, 150, 105898.

Sun, J., 2009. Sustainable road safety: Development, transference and application of community-based macro-level collision prediction models. Master of Applied Science Dissertation, University of British Columbia, May 2009, Vancouver, Canada.

Sun, Y., 2018. Sharing and riding: How the dockless bike sharing scheme in China shapes the city. *Urban Science*, 2(3), 68.

Sze, N. N., Su, J., Bai, L., 2019. Exposure to pedestrian crash based on household survey data: Effect of trip purpose. *Accident Analysis & Prevention*, 128, 17-24.

Talbot, R., Reed, S., Barnes, J., Thomas, P. D., Christie, N., 2014. Pedal cyclist fatalities in London: analysis of police collision files (2007-2011).

Tang, C.K., 2016. Traffic externalities and housing prices: evidence from the London congestion charge. SERC Discussion Papers 0205, Spatial Economics Research Centre, LSE.

Tax, D.M., Duin, R.P., 2004. Support vector data description. *Machine Learning*, 54(1), 45-66.

Taylor, D., Mahmassani, H., 1996. Analysis of stated preferences for intermodal bicycle-transit interfaces. *Transportation Research Record*, 1556 (1), 86-95.

Teschke, K., Frendo, T., et al., 2014. Bicycling crash circumstances vary by route type: a cross-sectional analysis. *BMC Public Health*, 14(1), 1-10.

Thornley, S. J., Woodward, A., Langley, J. D., Ameratunga, S. N., Rodgers, A., 2008. Conspicuity and bicycle crashes: preliminary findings of the Taupo Bicycle Study. *Injury Prevention*, 14(1), 11-18.

Tin Tin, S., Woodward, A., Ameratunga, S., 2010. Injuries to pedal cyclists on New Zealand roads, 1988-2007. *BMC Public Health*, 10(1), 1-10.

Tin Tin, S., Woodward, A., Ameratunga, S., 2013. The role of multilevel factors in geographic differences in bicycle crash risk: a prospective cohort study. *Environmental Health*, 12(1), 1-10.

Train, K., 2009. Discrete choice methods with simulation. Cambridge university press.

Transport for London, TfL, 2015. Barclays Cycle Hire customer satisfaction and usage survey: Wave 9 (Quarter 3 2014/15).

Transport for London, TfL, 2018. Travel in London, report. http://content.tfl.gov.uk/travel-in-london-report-11.pdf.

Transport for London, TfL., 2005. Central London congestion charging scheme: impact monitoring. Available: http://www.tfl.gov.uk.

Transport for London, TfL., 2019. Ultra-Low Emission Zone (access on December 11 2020). https://tfl.gov.uk/modes/driving/ultra-low-emission-zone.

Trapp, G. S., Giles-Corti, B., Christian, H. E., Bulsara, M., Timperio, A. F., McCormack, G. R., Villaneuva, K. P., 2011. On your bike! a cross-sectional study of the individual, social and environmental correlates of cycling to school. *International Journal of Behavioural Nutrition and Physical Activity*, 8, 1-10.

Tseloni, A., Osborn, D. R., Trickett, A., Pease, K., 2002. Modelling property crime using the British Crime Survey. What have we learnt?. *British Journal of Criminology*, 42(1), 109-128.

Tunaru, R., 2002. Hierarchical Bayesian models for multiple count data. *Austrian Journal of Statistics*, 31(2&3), 221-229.

Turner, S., Francis, T., Roozenburg, A., Transport, N. L., 2006. Predicting accident rates for cyclists and pedestrians (p. 72). Wellington: Land Transport New Zealand.

Vandenbulcke, G., Thomas, I., de Geus, B., Degraeuwe, B., Torfs, R., Meeusen, R., Panis, L. I., 2009. Mapping bicycle use and the risk of accidents for commuters who cycle to work in Belgium. *Transport Policy*, 16(2), 77-87.

Vandenbulcke, G., Thomas, I., Panis, L. I., 2014. Predicting cycling accident risk in Brussels: a spatial case–control approach. *Accident Analysis & Prevention*, 62, 341-357.

Vanparijs, J., Panis, L. I., Meeusen, R., De Geus, B., 2015. Exposure measurement in bicycle safety analysis: A review of the literature. *Accident Analysis & Prevention*, 84, 9-19.

VE, S., Cho, Y., 2020. A rule-based model for Seoul Bike sharing demand prediction using weather data. *European Journal of Remote Sensing*, 53, 166-183.

Venkataraman, N. S., Ulfarsson, G. F., Shankar, V., Oh, J., Park, M., 2011. Model of relationship between interstate crash occurrence and geometrics: exploratory insights from random parameter negative binomial approach. *Transportation Research Record*, 2236(1), 41-48.

Venkataraman, N., Ulfarsson, G. F., Shankar, V. N., 2013. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. *Accident Analysis & Prevention*, 59, 309-318.

Vlakveld, W., Mons, C., Kamphuis, K., Stelling, A., Twisk, D., 2021. Traffic conflicts involving speed-pedelecs (fast electric bicycles): A naturalistic riding study. *Accident Analysis & Prevention*, 158, 106201.

Walker, K., Marino, A., Gupta, M., Hebert., 2017. The pose knows: Video forecasting by generating pose futures. Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, October.

Wang, B., Kim, I., 2018. Short-term prediction for bike-sharing service using machine learning. *Transportation Research Procedia*, 34, 171-178.

Wang, C., Xie, Y., Huang, H., Liu, P., 2021. A review of surrogate safety measures and their applications in connected and automated vehicles safety modelling. *Accident Analysis & Prevention*, 157, 106157.

Wang, C., Xu, C., Xia, J., Qian, Z., 2017. Modelling faults among e-bike-related fatal crashes in China. *Traffic Injury Prevention*, 18(2), 175-181.

Wang, J., Huang, H., 2016. Road network safety evaluation using Bayesian hierarchical joint model. *Accident Analysis & Prevention*, 90, 152-158.

Wang, L., Abdel-Aty, M., Lee, J., Shi, Q., 2019a. Analysis of real-time crash risk for expressway ramps using traffic, geometric, trip generation, and socio-demographic predictors. *Accident Analysis & Prevention*, 122, 378-384.

Wang, L., Abdel-Aty, M., Ma, W., Hu, J., Zhong, H., 2019b. Quasi-vehicle-trajectory-based real-time safety analysis for expressways. *Transportation Research Part C: Emerging Technologies*, 103, 30-38.

Wang, S., 2012. The improved Dijkstra's shortest path algorithm and its application. *Procedia Engineering*, 29, 1186-1190.

Wang, X., Jin, Y., Abdel-Aty, M., Tremont, P., Chen, X., 2012. Macro-level model development for safety assessment of road network structures. *Transportation Research Record*, 2280, 100-109.

Wang, X., Lindsey, G., Schoner, J. E., Harrison, A., 2016. Modelling bike share station activity: Effects of nearby businesses and jobs on trips to and from stations. *Journal of Urban Planning and Development*, 142, 04015001.

Washington, S., Karlaftis, M., Mannering, F., Anastasopoulos, P., 2011. Statistical and Econometric Methods for Transportation Data Analysis (2nd Edition), Chapman and Hall/CRC, New York (2011).

Wei, F., 2010. Boundary effects in developing macro-level CPMs: A case study of city of Ottawa. University of British Columbia, Vancouver, Canada.

Wei, F., Lovegrove, G., 2013. An empirical tool to evaluate the safety of cyclists: Community based, macro-level collision prediction models using negative binomial regression. *Accident Analysis & Prevention*, 61, 129-137.

Wen, H., Zhang, X. Zeng, Q., Sze, N.N., 2019. Bayesian spatial-temporal model for the main and interaction effects of roadway and weather characteristics on freeway crash incidence. *Accident Analysis & Prevention*, 132, 105249.

Winters, M., Teschke, K., 2010. Route preferences among adults in the near market for bicycling: findings of the cycling in cities study. *American Journal of Health Promotion*, 25(1), 40-47.

Wolff, H., 2014. Keep your clunker in the suburb: Low-emission zones and adoption of green vehicles. *The Economic Journal*, 124, 481-512.

Wong, S. C., Sze, N. N., Li, Y. C., 2007. Contributory factors to traffic crashes at signalized intersections in Hong Kong. *Accident Analysis & Prevention*, 39(6), 1107-1113.

Wood, J. S., Donnell, E. T., Porter, R. J., 2015. Comparison of safety effect estimates obtained from empirical Bayes before–after study, propensity scores-potential outcomes framework, and regression model with cross-sectional data. *Accident Analysis & Prevention*, 75, 144-154.

Wood, J., Donnell, E., 2017. Causal inference framework for generalizable safety effect estimates. *Accident Analysis & Prevention*, 104, 74–87.

World Health Organization, WHO, 2018. Global status report on road safety 2018: Summary. World Health Organization.

Xie, L., Olszewski, P., 2011. Modelling the effects of road pricing on traffic using ERP traffic data. *Transportation Research Part A: Policy and Practice*, 45, 512–522.

Xing, F., Huang, H., Zhan, Z., Zhai, X., Ou, C., Sze, N. N., Hon, K. K., 2019. Hourly associations between weather factors and traffic crashes: non-linear and lag effects. *Analytic Methods in Accident Research*, 24, 100109.

Xu, C., Ji, J., Liu, P., 2018a. The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. *Transportation Research Part C: Emerging Technologies*, 95, 47-60.

Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., Liu, Y., Zhao, Y., Pei, D., Feng, Y., Chen, J., Wang, Z., Qiao, H., 2018b. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. Proceedings of the 2018 World Wide Web Conference, Lyon, France, April.

Yamamoto, T., Hashiji, J., Shankar, V. N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. *Accident Analysis & Prevention*, 40(4), 1320-1329.

Yamamoto, T., Shankar, V. N., 2004. Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects. *Accident Analysis & Prevention*, 36(5), 869-876.

Yang, C., Mesbah, M., 2013. Route choice behaviour of cyclists by stated preference and revealed preference. Proceedings of the 2013 Australasian Transport Research Forum.

Yang, K., Wang, X., Yu, R., 2018. A Bayesian dynamic updating approach for urban expressway real-time crash risk evaluation. *Transportation Research Part C: Emerging Technologies*, 96, 192-207.

Yang, Y., Heppenstall, A., Turner, A., Comber, A., 2020. Using graph structural information about flows to enhance short-term demand prediction in bike-sharing systems. *Computers, Environment and Urban Systems*, 83, 101521.

Yang, Z., Hu, J., Shu, Y., Cheng, P., Chen, J., Moscibroda, T., 2016. Mobility modelling and prediction in bike-sharing systems. Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services, Singapore, 165-178.

Yang, Z., Hu, Z., Salakhutdinov, R., Berg-Kirkpatrick, T., 2017. Improved variational autoencoders for text modelling using dilated convolutions. Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, August.

Yao, S., Loo, B. P., 2016. Safety in numbers for cyclists beyond national-level and city-level data: a study on the non-linearity of risk within the city of Hong Kong. *Injury Prevention*, 22(6), 379-385.

Yasmin, S., Eluru, N., 2016. Latent segmentation based count models: analysis of bicycle safety in Montreal and Toronto. *Accident Analysis & Prevention*, 95, 157-171.

Ye, X., Pendyala, R. M., Washington, S. P., Konduri, K., Oh, J., 2009. A simultaneous equations model of crash frequency by collision type for rural intersections. *Safety Science*, 47(3), 443-452.

Yu, R., Wang, Y., Zou, Z., Wang, L., 2020. Convolutional neural networks with refined loss functions for the real-time crash risk analysis. *Transportation Research Part C: Emerging Technologies*, 119, 102740.

Yuan, J., Abdel-Aty, M., Gong, Y., Cai, Q., 2019. Real-time crash risk prediction using long short-term memory recurrent neural network. *Transportation Research Record*, 2673(4), 314-326.

Zacharias, J., 2005. Non-motorized transportation in four Shanghai districts. *International Planning Studies*, 10(3-4), 323-340.

Zeng, Q., Wen, H., Huang, H., Pei, X., Wong, S. C., 2017. A multivariate random-parameters Tobit model for analysing highway crash rates by injury severity. *Accident Analysis & Prevention*, 99, 184-191.

Zhai, X., Huang, H., Gao, M., Dong, N., Sze, N.N., 2018. Boundary crash data assignment in zonal safety analysis: An iterative approach based on data augmentation and Bayesian spatial model. *Accident Analysis & Prevention*, 121, 231-237.

Zhai, X., Huang, H., Sze, N. N., Song, Z., Hon, K. K., 2019a. Diagnostic analysis of the effects of weather condition on pedestrian crash severity. *Accident Analysis & Prevention*, 122, 318-324.

Zhai, X., Huang, H., Xu, P., Sze, N.N., 2019b. The influence of zonal configurations on macro-level crash modelling. *Transportmetrica A*, 15, 417-434.

Zhan, X., Aziz, H. A., Ukkusuri, S. V., 2015. An efficient parallel sampling technique for Multivariate Poisson-Lognormal model: Analysis with two crash count datasets. *Analytic Methods in Accident Research*, 8, 45-60.

Zhang, Y., Bigham, J., Li, Z., Ragland, D., Chen, X., 2013. Associations between Road Network Structure and Pedestrian-Bicyclist Accidents. 92nd Transportation Research Board Annual Meeting Compendium of Papers. No. 13-4316.

Zhang, Y., Li, H., Sze, N.N., Ren, G., 2021. Propensity score methods for road safety evaluation: Practical suggestions from a simulation study. *Accident Analysis & Prevention*, 158, 106200.

Zhang, Y., Mi, Z., 2018. Environmental benefits of bike sharing: a big data-based analysis. *Applied Energy*, 220, 296–301.

Zhang, Y., Thomas, T., Brussel, M., Van Maarseveen, M., 2017. Exploring the impact of built environment factors on the use of public bikes at bike stations: case study in Zhong shan, China. *Journal of Transport Geography*, 58, 59-70.

Zhao, M., Liu, C., Li, W., Sharma, A., 2018. Multivariate Poisson-lognormal model for analysis of crashes on urban signalized intersections approach. *Journal of Transportation Safety & Security*, 10(3), 251-265.

Zheng, Z., Ahn, S., Monsere, C.M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis & Prevention*, 42 (2), 626-636.

Zhong, S., Bushell, M., 2017. Impact of the built environment on the vehicle emission effects of road pricing policies: A simulation case study. *Transportation Research Part A: Policy and Practice*, 103, 235–249.

Zhou, X., 2015. Understanding spatiotemporal patterns of biking behaviour by analysing massive bike sharing data in Chicago. *PloS One*, 10, e0137922.

Zhu, D., Sze, N.N., Feng, X., Yang, Z., 2022. A two-stage safety evaluation model for the red light running behaviour of pedestrians using the game theory. *Safety Science*, 147, 105600.

Zhu, D., Sze, N.N., Feng, Z., 2021. The trade-off between safety and time in the red light running behaviours of pedestrians: A random regret minimization approach. *Accident Analysis & Prevention*, 158, 106214.