Technical University of Munich

TUM School of Engineering and Design

Chair of Transportation Systems Engineering

Master's Thesis

The Factors behind Willingness to Pool and Successful Ride-Pooling:

How COVID-19 Influences That

Merindha Arty Sekardani

Supervised by:
M.Sc Mohamed Abouelela

5 February 2023

Declaration of Authorship

I, Merindha Arty SEKARDANI, confirm that this master's thesis titled "The Factors behind Willingness to Pool and Successful Ride-Pooling: How COVID-19 Influences That" is my own work and I have documented all sources and material used.

Signed:

Date: 05.02.2023

Abstract

The fast-growing popularity of ride-haling has invoked concerns regarding its negative impacts especially pertaining to sustainability. Pooling of hailed rides is dubbed as a potential solution. However, as of today, the rate of ride-pooling requests and matching success are still very low. The influencing factors of willingness to pool have been extensively researched, but the same cannot be said about pooling success. Moreover, there is only a little knowledge on how a pandemic such as COVID-19 could affect the mentioned factors.

This thesis took the trip-level statistical modelling approach to investigate the most influential factors to willingness to pool and materialisation of pooling request using the ride-hailing scene of Chicago as a case study. Taking notes from past studies' limitations, this thesis considered a wide range of potential factors including exogenous ones such as weather and crime rate. This thesis also proposed a methodology which enable working with large-sized trip data without risking loss of information to aggregation and sampling. Two statistical selection methods—Backward Stepwise Elimination and Lasso Regression—were utilised and compared.

At the end of this study, the potential driving factors to pooling decision and success in both non-pandemic and pandemic contexts were identified. The results showed that while trip impedance, temporal attributes, and weather possibly remain influential for both outcomes, the magnitude and direction of effects could change depending on the pandemic context. This thesis also discovered that post-outbreak, pandemic-related variables may pose the biggest impacts on willingness to pool and pooling success. Other findings include the potential effects of additional taxing on certain parts of the city, while built environment, spatiodemographic attributes, and crime rate may pose little to no impact.

Acknowledgements

My sincere gratitude goes to my supervisor M.Sc. Mohamed Abouelela who with tremendous patience gave me continuous support and guidance throughout this whole thesis.

I would also like to thank my family whom I owe everything to and my friends for their care and encouragement.

Lastly, I would like to thank the woman in the mirror for keeping it together and pushing through.

Contents

List of Figures

List of Tables

List of Abbreviations

ACS	American Community Survey
AIC	Akaike Information Criterion
AICc	Corrected Akaike Information Criterion
ANOVA	Analysis of Variance
AUC	Area under the Curve
BIC	Bayesian Information Criterion
BLR	Binomial Logistic Regression
CAGR	Compound Annual Growth Rate
CBD	Central Business District
CO	Carbon Monoxide
COVID-19	Coronavirus Disease 2019
CV	Cross-Validation
Df	Degree of Freedom
ETR	Extra Trees Regression
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GHG	Greenhouse Gas
GLM	Generalised Linear Models
GPS	Global Positioning System
GTT	Ground Transportation Tax
GVIF	Generalised Variance Inflation Factor
GWR	Geographically Weighted Regression
GWTR	Geographically and Temporally Weighted Regression
HC	Hydrocarbons
ICT	Information and Communication Technologies
IDE	Integrated Development Environment
ITS	Interrupted Time Series
LASSO or Lasso	Least Absolute Shrinkage and Selection Operator
log	Logarithmic
ML	Machine Learning
MLE	Maximum Likelihood Estimation
MLR	Multivariate Linear Regression
MSE	Mean Squared Error
NOx	Nitrogen Oxides
OD	Origin-Destination
OLS	Ordinary Least Squares
PCLR	Principal Component Logistic Regression
R	R statistical programming language
RFR	Random Forest Regression
ROC	Receiver Operator Characteristics
SAR	Spatial Autoregressive model

SDE	Spatial Durbin Error model
SE	Spatial Error model
SEM	Structural Equation Modelling
SP	Stated Preference
TN	True Negative
TNC or TNP	Transportation Network Company or Transportation Network Provider
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
US	United States
USD	United States Dollar
VIF	Variance Inflation Factor
VKT or VMT	Vehicle Kilometres Travelled or Vehicle Miles Travelled
WAV	Wheelchair Accessible Vehicle
WTP	Willingness to Pool

Chapter 1

Introduction

This writing begins by introducing the thesis starting from the the existing problems, needs, and research gaps which invoke the undertaking of this thesis. The recent development in the shared mobility and its relevance to the global transportation challenges is discussed below. Furthermore, subsequent research questions are elaborated along with the contributions of this thesis, research framework, and thesis structure.

1.1 Problem Statement

In this era of increasingly pressing climate issues, the transportation sector has been under the spotlight as it is responsible for 16.2% and 27.6% of global greenhouse gas (GHG) emission [1] and energy consumption [2] respectively. Out of all the energy consumed globally for transport, the United States massively takes up 25% with more than 60% of it dedicated for road passenger travels [3]. This issue is exacerbated by the ever-growing motorisation which places the United State's road passenger-kilometres as the second highest in the world with a steady continuous growth [4].

The recent advancements of information and communication technologies (ICT) have accelerated the growth of shared mobility—a concept in which travelers could gain access to a shared vehicle for a short term ([5]). The global shared mobility scene has seen a growth over the years and is expected to keep climbing ([6], [7]). The same also applies in the United States ([8], [9]). Shared mobility encompasses various kinds of vehicles and business models, however, ride-hailing—in which a passenger is driven to their destination with a car—is the most dominant in the United States ([10]). A company which offers this ride-hailing service is often referred to as a Transportation Network Provider (TNP).

The tendency of American cities to form decentralised sprawls [11] has resulted in car dependency [12] which promotes the vast adoption of ride-hailing. However, from the point of view of sustainability, there has been mixed opinions regarding the impacts of ride-hailing on the environment. The debate is centred around whether ride-hailing decreases or rather increases the circulation of cars on the streets, and, thereby, the subsequent consequences. These positive and negative potentials of ride-hailing will be elaborated on later parts of this thesis.

Despite the above dilemma, pooling hailed rides potentially minimise the negative impacts of ride-hailing [13]–[15]. Pooling refers to grouping ride-hailing requests which are similar or have coinciding routes and subsequently carry them out in one trip. However, although the option to pool is readily provided by TNPs in many cities, the data from US cities show that the users' willingness to pool [16], and hence, the average ride-hailing occupancy rate [17], are still very low. This is despite the fact that about 50% of ride-hailing trips are feasible for pooling [14].

1.2 Needs

A survey in the US estimated only around 20% of ride-hailing users would opt for pooling [16], out of which around 15% would be successfully matched and materialised [17]. Considering the major impact the US road transportation scene imposes on the global emission and fuel consumption, it is necessary to increase the pooling rate by understanding the underlying reasons. Therefore, the factors which influence people's willingness to pool as well as the success of the pool request need to be identified. Moreover, the recent COVID-19 pandemic has shown how such occurrence could drastically change people's social perceptions and behaviours [18]. As COVID-19-like pandemics are predicted to occur more frequently in the future [19], it is hence vital to investigate whether such pandemic affects the mentioned pooling influence factors.

In recent years, more governments started to adopt the open government data philosophy in order to promote transparency and participatory governance [20], [21]. For example, as a part of this initiative, the City of Chicago releases the records of each individual TNP trip within the city starting from November 2018. However, as hundreds of thousands of trips are conducted daily in a city, the sheer size of such raw dataset could pose challenges. Since such a dataset is highly valuable for transportation researches, there needs to be a systematic methodology which is proficient in handling, processing, and analysing large trip data.

1.3 Research Gap

To identify the research gaps that exist in the literature, this thesis looked into past studies which explored similar themes. Although many authors have delved into the factors behind willingness to pool, the driving factors of pooling success have been much less researched. Most past studies also made the assessment at an aggregated level rather than at individual trip level, as they may have been hindered by the large size of the trip data. Moreover, the possible effects of exogenous factors such as public holidays, weather, and crime rate have not been considered by many. Additionally, there has been very limited research on ride-pooling in a pandemic context.

1.4 Objectives and Research Questions

This thesis strives to help promoting pooling which has the potential to alleviate the drawbacks of ride-hailing. With higher rate of pooling, the ride-hailing industry could be made more environmentally sensitive and economically efficient for all parties. Thus, this thesis sees the necessity to investigate what contributes to people's willingness to pool and pooling success in order to be able to enhance the pooling rate. Consequently, the objectives of this thesis could be identified as follows:

1. To establish a systematic procedure to handle, analyse, and build statistical models employing large-sized trip data
2. To build models for the willingness to pool and pooling success involving the trip attributes and various exogenous factors before and after the COVID-19 outbreak
3. To identify the most influential factors affecting rider′s willingness to pool and pooling success before and after the COVID-19 outbreak
4. To make a comparison between the influencing factors before and after the COVID-19 outbreak to deduce if such pandemic could shift people′s motivation towards ride-pooling and its success

Ultimately, this thesis strives to answer the following research questions:

- **What are the factors influencing people′s willingness to pool in non-pandemic and pandemic contexts?**
- **What are the factors influencing the success of pooling in non-pandemic and pandemic contexts?**
- **What are the efforts that could be done to encourage ride-hailers to pool and to enhance ride-pooling success?**

1.5 Contributions

The work within this thesis could yield several theoretical, methodological, and practical contributions:

- **Theoretical contributions:**
  1. Review of the impacts and the factors behind ride-hailing
  2. Synthesis of the benefits of ride-pooling
  3. Potential future research
- **Methodological contributions:**
  1. Systematic procedure for handling, processing, and modelling with large data
  2. Comparison and reviews of the Stepwise and Lasso selection techniques
- **Practical contributions:**
  1. Identification of the main factors affecting people′s willingness to pool and pooling success
  2. Identification of the main factors affecting people′s willingness to pool and pooling success in the context of pandemic

The theoretical and methodological contributions of this thesis would be especially beneficial in future researches involving ride-hailing and/or large-sized data. Whereas, the practical contributions are valuable inputs for TNPs and authorities to enhance ride-pooling matching rate, incentivise ride-pooling, as well as to establish strategies in response to COVID-19-like pandemic.

1.6 Research Framework and Report Structure

This thesis report is structured systematically and elaborates the research in seven chapters:

- Chapter 1 identifies the existing concerns pertaining to ride-hailing industry and the needs that shaped the objectives of this thesis
- Chapter 2 delves into the literature to provide more details on the impacts of ride-hailing and ride-pooling along with the previous works and theories that laid the foundation of this thesis' methodologies
- Chapter 3 introduces the study area Chicago and its ride-hailing scene before and after the COVID-19 outbreak
- Chapter 4 elaborates on the systematic methodologies in handling and modelling large-sized trip data
- In Chapter 5 and 6, the results are presented and discussed
- Chapter 7 concludes the research in relation to the research questions and discusses the limitations of this study

Chapter 2

Literature Review

This chapter delves into the literature on the topic of ride-hailing and ride-pooling to provide more context on the thesis's motivation, to outline the theories behind this thesis's methodologies, and to explore previous works on this theme.

2.1 Ride-Hailing

This section aims to provide further understanding on the concept of ride-hailing, as well as elaborate on the impacts. This hence brings forth the discussions on the benefits of pooling.

2.1.1 Ride-Hailing as a Part of Shared Mobility

The shared mobility concept is a part of the bigger sharing economy philosophy, which, in its essence, aims to share underutilised assets in the interests of efficiency and sustainability [22]. Often, this definition coincides with other ideas such as the access economy, community-based economy, collaborative economy, and collaborative consumption [23]. In the realm of shared mobility, these assets correspond to transportation vehicles or any means of passenger or goods mobility.

Its broad nature and rapid development have caused discrepancies of terminologies and classifications within the field of shared mobility. Notwithstanding, Shaheen et al. [24] classified shared mobility into three main groups: sharing of a vehicle or device, sharing of a passenger ride, and sharing of a delivery ride. See Figure (2.1) below for the full classification tree. On top of these modes, some authors would also include conventional mass public transport [25] or other mobility assets such as parking spots [26]–[28] under the umbrella of shared mobility.

Ride-hailing is often synonymous with ride-sourcing, on-demand rides, app-based rides, Transportation Network Companies (TNCs), or Transportation Network Providers (TNPs). More often than not, the term 'ride-sharing' is also used to describe ride-hailing. However, this is technically incorrect as ride-sharing trips should carry more than one passenger [17]. Fundamentally, ride-hailing is described as a mobility service in which a passenger is connected to a community driver through an online platform to be picked up from their current location and driven to their destination using a private car. Grouping of ride-hailing requests into one trip is defined as ride-pooling or ride-splitting. Emphasis is put on its on-demand nature, in which

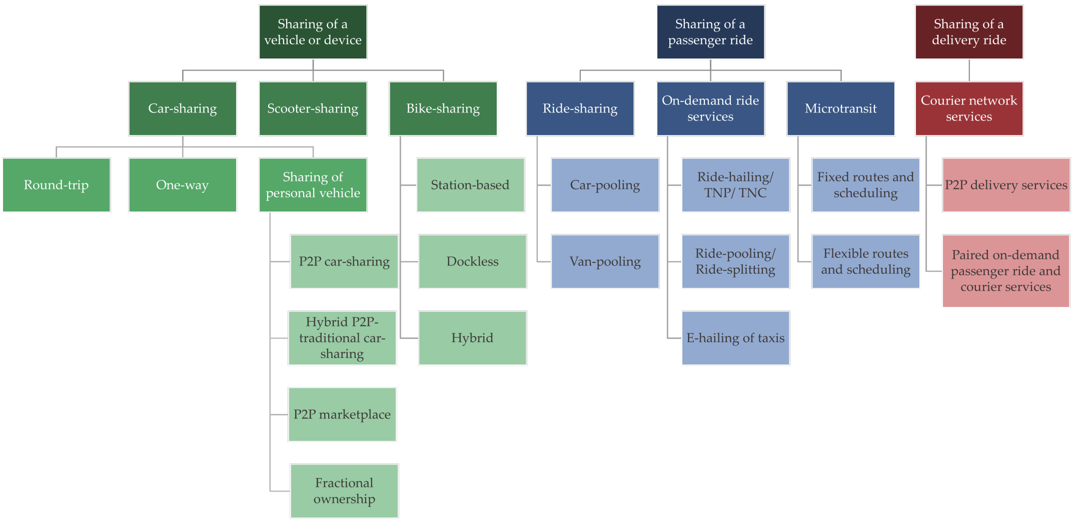


FIGURE 2.1: Classification of shared mobility business models. Adopted from Shaheen et al. [24]

the trip is initiated by the request of the passengers and not of the driver's initiative. This sets apart ride-splitting from car-pooling.

2.1.2 The Growth of Ride-hailing

The concept of sharing a car ride has existed since the World War II era as a response to vehicle shortage [29]. However, the idea was revolutionised by the integration of GPS and online payments technologies in smartphones which birthed today's ride-hailing services. Carma (previously Avego, gocarma.com) was one of the pioneers in the modern ride-hailing [29], however today's global market is dominated by companies such as Didi (web.didiglobal.com), Uber (uber.com), Lyft (lyft.com), and Grab (grab.com) [30], [31]. In the United States, Uber and Lyft are the main players in the market [32], [33].

The ride-hailing market is one of the fastest-growing, with the global market value rising from 1 billion USD to 61 billion USD in a span of a decade [32]. Another source even appraised the ride-hailing market to worth 113 billion USD in 2020 and it is expected to climb with a CAGR of 8.75-20% to reach 220 billion USD in 2025 [32], [33]. This rapid growth is also reflected by the ridership. Uber, one of the biggest TNPs worldwide, reached 111 millions of monthly active users in 2019 after only launching its beta programme in 2011, equating to roughly 1.9 billion trips per calendar quarter [34]. Figure 2.2 illustrates the growth of the number of trips undertaken by Uber over a span of 3 years prior to COVID-19 pandemic.

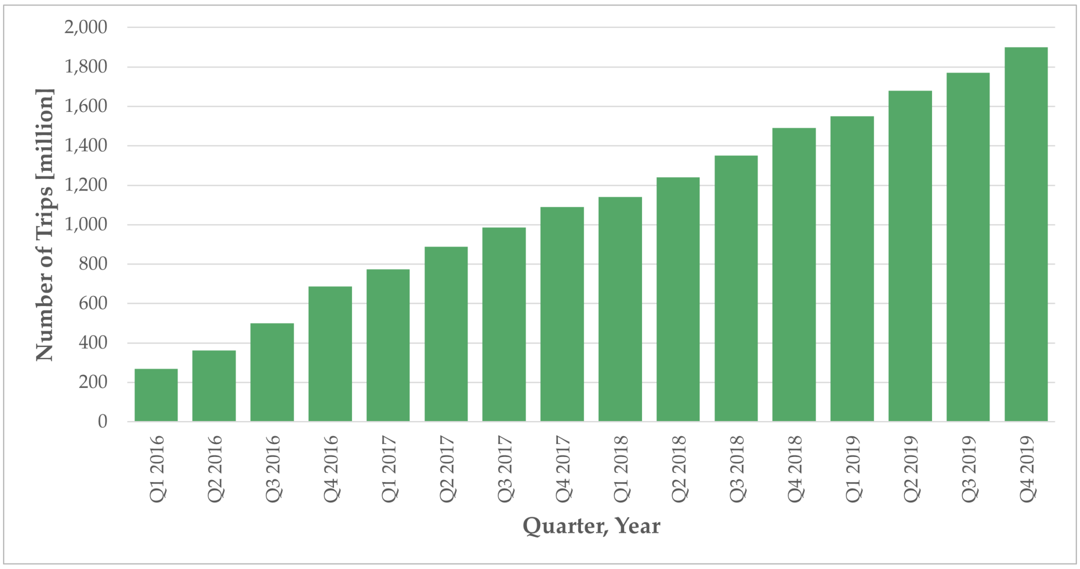


FIGURE 2.2: The quarterly number of trips of Uber period 2016-2019.
Source: Dean [34]

2.1.3 The Factors Behind Ride-hailing Use

Many surveys agree that ride-hailing is mostly popular among the younger, more educated working population [17], [35]–[44]. The skewed age distribution is often regarded as a product of digital divide—the different adeptness in technology—between younger and older generations [43]. The literature is not as unanimous regarding the typical income level. Although some studies associated ride-hailing more with higher-earners [17], [36]–[40], [43], [45], past findings showed that low-earners use ride-hailing more than middle-earners [39], [41], [46]. Qiao & Yeh revealed that this distribution across different income levels could depend on the trip purpose [46]. Gender distributions also vary across studies and may be influenced by overall safety and cultural aspects pertaining to gender roles [42].

The relationship between vehicle ownership and ride-hailing is also complicated as evidences showed that car-less households generate relatively more ride-hailing trips [39], [40], [45], [47], but most users do own or have access to private vehicles [17], [37], [38]. Ride-hailing users also tend to lead a more mobile and/or multi-modal lifestyle as a study found more users own public transit subscription or pass compared to taxi frequenters [43]. Studies also highlighted the significance of attitudinal factors towards ride-hailing use [48], as well as the perceptions of ride-hailing's safety and ease of use [37].

Temporally, the highest number of trips occur in the evening/at night and/or during the weekends [41], [43], [44]. This is in line with the top ride-hailing trip purpose, that is going to or from social/recreational events [35], [39], [42], [44]. Spatially, ride-hailing trips are highly concentrated in urban areas [39], [45]. More specifically, going to or from residential, commercial, and central business district/downtown mixed used locations [41].

Moreover, when people hail a ride rather than driving, it is mainly caused by people avoiding drinking under influence as well as the cost and/or difficulties of parking [36], [39], [47]. Meanwhile, people choose ride-hailing over public transit due to ride-hailing being quicker, more reliable, and public transit not being available at the time of the trip [35], [36], [39]. This naturally brings up a comparison between ride-hailing and conventional taxi. However, ride-hailing was observed and perceived to have shorter and more consistent wait times [35], [44], an attribute that was found to be highly valued by consumers along with the real-time arrival information [49], [50]. Ride-hailing being cheaper, easier to request and pay for, as well as having more transparent fare are also driving factors [38], [44]. The superior popularity of ride-hailing over conventional taxis were further supported as taxi became highly substituted by ride-hailing [37], [46], [51]–[53].

2.1.4 The Impacts of Ride-hailing

With the pervasiveness of ride-hailing, comes questions about its impacts. By providing access to and from less-connected areas, ride-hailing could connect the lower-income groups to more opportunities and activities [46], [54], [55]. It also potentially improves people's quality of life through increasing the mobility of people with physical and cognitive disabilities [47], [56]. Other benefits include the reduction of parking requirement [17], [57] and accommodation of late-night trips which may be unsafe or poorly-served by public transit [35], [58]. However, despite often marketed as a sustainable transport mode, many questions the macro impacts of ride-hailing, especially pertaining to its implications on the environment.

Theoretically, ride-hailing has the ability to alleviate auto-dependency [35], and in turn, reduce private vehicle ownership and circulation which is key for an efficient transportation system [59]. However, as previously stated, the gathered evidences have yielded mixed conclusions. Some studies associated regular ride-hailing usage with owning fewer cars [58], [60]. However, the causality in terms of which begets what is unclear. While a handful of respondents in [17], [36], [51] claimed disposing or foregoing car ownership in response to ride-hailing, the dominant portions reported no change in their attitudes towards owning a car. There are also cases of people acquiring new vehicles to take up a full-time job as a ride-hailing driver [61]–[63], even causing nett increase in vehicle ownership [64].

Concerns arose as some evidences showed that ride-hailing pulls people away from public transit and active modes (walking, cycling, etc.). 14%-37% survey respondents reported they would have taken public transit for their trips had ride-haling not been available [37], [51], [58], [65]. Meanwhile, 10%-24% ride-hailing trips would have been made by walking or cycling [47]. Further investigations discovered that ride-hailing could both be substitutive and complementary [35] to public transit depending on the specific modes [36], trip purpose, and target population [46]. Some surveys have also shown 8% - 22% of induced travel effect—that is trips that would not be conducted had ride-hailing not been available [17], [35], [36], [39].

Another drawback of ride-hailing pertains to deadheading—the kilometers driven without passenger onboard, mostly to reach the pickup location. In US cities, the magnitude estimations range greatly from 43% to 82% of the total distance driven [17], [66], [67], potentially constituting to distance weighted passenger occupancy of only 0.8 [17] or less. Due to deadheading, non-pooled hailed ride was approximated to be 47% more polluting than private vehicle trip [68].

When coupled, the above factors may lead to increased vehicle kilometres/miles travelled (VKT/VMT) of 83.5% [17] to 90% [38]. In the case of the US, research suggested that ride-hailing was responsible for 7.8 millions of daily VMT in the year

2017 [69]. This is supported by evidences of increased traffic delay (congestion) in relation to ride-hailing operations [61], [70].

2.1.5 The Benefits of Pooling

Pooling of ride-hailing potentially alleviates the disadvantages of ride-hailing elaborated above. Although the detours in pooling could increase the rider's travel distance by 20%-35% [71], empirical evidences showed that drivers could have saved 22%-35% of VKT compared to the case should these trips were conducted separately [71], [72]. Even considering trips which substituted public transit or active modes, ride-pooling trips in Hangzhou, China yielded nett decrease of 58,124 VKT daily [73]. In terms of emissions, studies showed pooling alleviating 15%-33% of ride-hailing GHG yields [14], [74]. In another case study from Chengdu, this reduction amounted to 10.601g of CO, 0.691g of NOx, and 1.424g of HC per trip [72]. 7.7%-15% decrease of fuel consumption were also quoted by a few studies [14], [75].

In a study, ride-pooling frequenters were also associated with higher level of private vehicle disposals compared to non-pooled ride-hailing users [44]. This supports Chen et al's [14] claim of 30% reduction of total vehicle count in the streets through pooling which in turn could improve the average velocity of the traffic network, especially during congested situation [76].

In the United States, TNCs and authorities have started to incentivise ride-pooling [13], [77]. Aside from reducing the operational costs [14], passenger demand could be better served even by smaller fleet size [13]. Riders hence also benefit from reduced cost per mile [78], while reduction in congestion lessens the burden of road construction and maintenance for the authorities [77].

These studies [74], [79]–[81] ultimately highlighted the importance of trip-matching algorithm optimisation in order to maximise the above advantages.

2.2 Researches on Willingness to Pool and Pooling Success

In literature, factors associated with ride-pooling adoption (i.e., willingness to pool) have been extensively explored. However, researches on what drives the pooling success have been very limited.

Many conducted a stated preference (SP) study through questionnaires and analysed the descriptive statistics to obtain individual-level factors associated with willingness to ride-pool. Examples include Kostorz et al.'s study in Hamburg [82], Mohamed et al. in London [44], and Wang et al. in China [83]. The factors investigated in this kind of approach is generally limited to individual socioeconomic/demographic attributes, temporal variables of the trip, and the trip purpose.

On top of analysing survey statistics, Gehrke et al. [84] also developed trip-level statistical models differentiated by trip purposes to find variables which have significance towards willingness to pool. The methodology consisted of Binomial Logistic Regression (BLR) supported with Backward Stepwise Elimination. This study was carried out in Boston pre-COVID-19 pandemic and suggested to study ride-pooling adoption across multiple time periods and/or contexts.

Likewise, BLR method was also adopted by Wang et al. [85] in a study in China to build a trip-level model for request matching success for ride-pooling between two studied cities. The authors, however, stated that future studies should consider more potential factors, especially pandemic-related variables when including post-outbreak datapoints.

Meanwhile, BLR approach was employed by Taiebat et al. [86] as a mean of exploratory analysis before building Machine Learning (ML) models based on AdaBoosting, Gradient Boosting, and Random Forest methods to predict willingness to pool and pooling success. These methods enable computation of the predictive power and effect direction of each predictor. This study, however, only utilised a small subset of the available ride-hailing trip data from Chicago [86].

Note that the above three studies utilised BLR to model willingness to pool and pool success at a trip level, as the model outcomes are binaries. For aggregated outcomes, linear statistical models such as Multivariate Linear Regression (MLR) with Ordinary Least Squares (OLS) estimations has been used. Examples include Li et al. [87] who spatially aggregated and modelled ride-pool pickup/dropoff counts.

Similar practice was carried out by Dean & Kockelman [88] who aggregated the ride-pooling trips of Chicago into count of trips authorised for pooling and its ratio over all ride-hailing trips per census tract. On top of MLR with OLS, these authors also conducted Spatial Autoregressive (SAR) model and Spatial Error (SE) model to investigate the spatial dependence of the dependent variable. To incorporate temporal variations into the model, Linear Panel Models were developed instead.

Zwick [89] also aggregated the ride-pooling requests in Hamburg per census tract and compared MLR with OLS method with Spatial Durbin Error (SDE) model and Geographically Weighted Regression (GWR) method. Meanwhile, Hou et al. [90] established bins based on the origin-destination (OD) tract pairs, temporal attributes, and whether the OD involves an airport. The ratio of trips authorised for pooling over all ride-hailing trips was then computed for each bin and became the dependent variable in an MLR and an XGBoost models.

Other researches which employed GWR—or its modification, Geographically and Temporally Weighted Regression (GTWR)—include Chen et al [91], Du et al. [92], and Chen et al. [93] who aggregated the willingness-to-pool trips (either as counts or ratios) spatially.

Machine Learning methods were employed by Abkarian et al. [94] (Random Forest Regression (RFR), Extra Trees Regression (ETR), and XGBoost) and Xu et al. [95] (RFR) who both aggregated the willingess-to-pool trips per OD pair. This was done to reveal the possible non-linear pattern, threshold effects, and variable importance of each predictor.

On the other hand, Romeo et al. [96] created hierarchical clusters of census tracts based on their socioeconomic/demographic attributes using the Ward's method. Analysis of Variance (ANOVA) was then executed to observe significant difference in the proportion of willingness-to-pool trips and the proportion of successfully pooled trips between the clusters [96].

Wang et al. [97] took a different approach and used Structural Equation Modelling (SEM) to assess the influence of attitudes in ride-pooling behaviour. Whereas, Abkarian et al. [98] specifically studied the impact of taxing policy on the count and ratio of trips authorised for pooling using the Interrupted Time Series (ITS) technique.

Table 2.1 summarises the mentioned researches. It lists the main methodology, the focus area (willingness to pool ('WTP') or pooling success ('Pool')), the analysis level (trip level or aggregated), and the possible factors considered. Note that the trip impedance includes the trip costs, distance, and duration.

This table shows how researches on pooling success is relatively rare. Many of the studies in Table 2.1 could also expand the list of factors integrated in the analysis, especially with weather-related variables which were found in Gehrke et al. [84] to have significance towards user's decision to pool. Many also aggregated the data, which risked information loss. Moreover, the impacts of COVID-19 on ride-pooling factors are under-researched.

TABLE 2.1: Summary of researches on willingness to pool and pooling success

No.	Author	Focus Area	Methodology	Analysis Level	Factors Considered										Other
					Trip Impedance	Temporal Attribute	Holiday	Socioeconomy / Demography	Land Use / POI	Walka-bility	Transit Access	Crime Level	Weather	Pandemic	
1	Abkarian et al. [94]	WTP	RFR, ETR, XGBoost	Aggregated	•	•		•	•	•	•	•			Area
2	Abkarian et al. [98]	WTP	ITS	Aggregated	•		•								Taxing intervention
3	Chen et al. [91]	WTP	GTWR	Aggregated		•			•						
4	Dean et al. [88]	WTP	MLR, SAR, SEM, Linear Panel	Aggregated	•	•		•			•				Job entropy, Road network density
5	Du et al. [92]	WTP	GWR	Aggregated		•		•	•		•				
6	Gehrke et al. [84]	WTP	SP, BLR	Trip level	•	•		•			•		•		Road network design
7	Hou et al. [90]	WTP	MLR, XGBoost	Aggregated	•	•		•							Airport pickup/ dropoff
8	Kostorz et al. [82]	WTP	SP	-		•		•							Trip purpose
9	Li et al. [87]	WTP	MLR	Aggregated				•	•						
10	Mohamed et al. [44]	WTP	SP	-		•		•							Trip purpose
11	Romeo et al. [96]	WTP, Pool	Cluster analysis	Aggregated	•	•		•			•				
12	Taiebat et al. [86]	WTP, Pool	BLR, AdaBoosting, Gradient Boosting, RFR	Trip level	•	•		•	•		•				Airport, Downtown, or Economically Disconnected Area pickup/dropoff
13	Wang et al. [83]	WTP	SP	-				•							
14	Wang et al. [97]	WTP	SEM	-											Attitudes
15	Wang et al. [85]	Pool	BLR	Trip level	•	•		•	•		•		•		Advanced appointment time
16	Xu et al. [95]	WTP	RFR	Aggregated	•			•	•	•	•	•			
17	Zwick et al. [89]	WTP	MLR, SDE, GWR	Aggregated				•	•		•			•	Hospital beds density

2.3 Statistical Modelling

Maria [99] defined a model as "a representation of the construction and working of some system of interest", which is built "usually based on analogies" and "with a specific goal" according to Chamizo [100]. Statistical modelling hence deals with emulating the generation of observed data through the use of mathematical representations, probability distributions, and various statistical analysis & assumptions.

2.3.1 Binomial (Binary) Logistic Regression

Binomial logistic regression is a statistical modelling technique which falls under the Generalised Linear Models (GLM) family. This logistic regression predicts the probability of an observation to be one of two categories, i.e., the dependent variable is a binary. This method is derived from the sigmoid probability function below:

$$P(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n)}} \tag{2.1}$$

$P(y=1)$ = the probability of an event occurring
x_n = predictor variables
β_0 = intercept
$\beta_{(1..n)}$ = regression coefficient of $x_{(1..n)}$

Considering that the probability of an event not occurring $P(y=0) = 1 - P(y=1)$ and applying log transformation yields the following:

$$\ln \frac{P(y=1)}{1-P(y=1)} = \ln(odds) = logit(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2.2}$$

The probability of an event occurring over the probability of the otherwise is commonly known as *odds*. The above equation enables the assumption that log odds of event y has linear relationships with predictor variables $x_{(1...n)}$. Therefore the coefficients $\beta_{(1...n)}$ could be interpreted as the change in the log odds of y with every unit increase of predictors $x_{(1...n)}$.

To reiterate, binomial logistic regression has the following requirements and assumptions:

- The dependent variable is dichotomous
- There needs to be independence of observations
- All of the categorical variables (including the dependent one) need to have mutually exclusive and exhaustive categories
- Assumes linearity between the continuous predictors and the logit transformations
- Multicollinearity should not exist among the predictor/independent variables
- Homoscedasticity is not required

- Significant outliers, high leverage points, and influential points should not exist among the data
- Due to the use of Maximum Likelihood Estimation (MLE) to estimate the coefficients (see section 2.3.1), the literature suggests minimum data size of 20-50 observations per predictor variable [101], [102].

Coefficient Estimation

The coefficients in the logistic regression is estimated by the probability framework Maximum Likelihood Estimation (MLE). Suppose a model parameter vector β and known observations X, the likelihood $L(\beta|X)$ is the conditional probability of observing X given a specific probability distribution and β values. In other words, $L(\beta|X)$ measures how well the observations support the validity of β values. Therefore, the goal of MLE is to find values of β which maximises $L(\beta|X)$.

In the context of binomial logistic regression, β should be such that:

- for observations with $y = 1$, the product of all probability $P(x)$ should be as near to 1 as possible
- for observations with $y = 0$, the product of all probability $P(x)$ should be as near to 0 as possible, i.e., $1 - P(x)$ should be as close to 1 as possible.

Therefore, across all observations, with x_i being the feature vector of the ith sample, this conditional probability translates to:

$$L(\beta) = \prod_{s}(P(x_i)^{y_i} * (1 - P(x_i))^{1-y_i}) \tag{2.3}$$

As probability function could also be expressed as:

$$P(x_i) = \frac{1}{1 + e^{-\beta x_i}} \tag{2.4}$$

and multiplication of exponentials could be unstable, transformation into the following log-likelihood is widely preferred:

$$l(\beta) = \sum_{i=1}^{n} y_i \beta x_i - \ln(1 + e^{\beta x_i}) \tag{2.5}$$

Some methods to maximise the log-likelihood in Equation 2.5 include:

- Newton-Raphson method [103]
- Bisection method [104]
- Fixed-point interaction [105]

Model Selection and Assessment

The term bias-variance tradeoff was coined by Geman et al. [106] to describe the dilemma in statistical modelling in which a tradeoff needs to be made between the accuracy and the precision of the model's prediction. Figure 2.3 depicts how square of bias error—the inverse of accuracy—decreases with model complexity, whereas the variance–the inverse of precision—behaves the opposite. Therefore, an ideal model is one that minimises the sum of these two errors, i.e., the Mean Squared Error (MSE). By doing so, one avoids underfitting or overfitting the model.

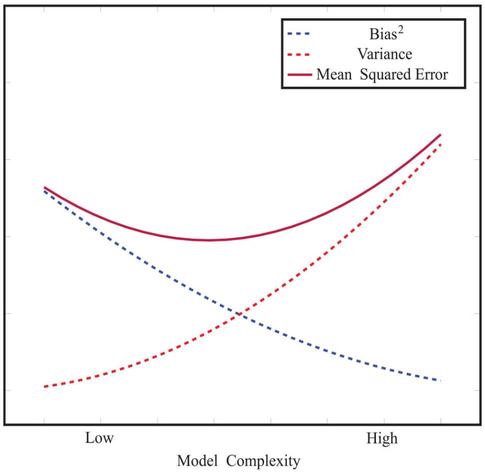


FIGURE 2.3: Illustration of the bias-variance dilemma.
Source: Doroudi [107]

In this sense, the following metrics are commonly referred to in logistic regression model selection and performance assessment:

1. **Akaike Information Criterion (AIC)** [108]
   The term information criterion refers to selection methods which are derived from likelihood functions. The AIC yields relative scores of model quality estimate which could be used to compare the model candidates of the same model class. It is defined as:
   $$AIC = -2\ln L + 2k \tag{2.6}$$
   L = likelihood
   k = number of model predictors

   The likelihood acts as a measure fit, and hence, the minimum AIC is desired. As shown, AIC penalises for any addition of predictor variable by a factor of 2. Therefore, given two model candidates with the same level of fit, AIC would side with the simpler one. However, as sample size grows, AIC tends to expand its choice of models and pick more complex model to reach the most optimum error [109]. For small-sample studies, the corrected version of AIC (AICc) also exists [110].

2. **Bayesian Information Criterion (BIC)** [111]
   It is given as:
   $$BIC = -2\ln L + k\ln n \tag{2.7}$$
   L = the likelihood
   k = the number of model predictors
   n = the number of observations

At large sample size, BIC imposes much larger penalty for an additional predictor compared to AIC.

3. **Pseudo-R^2**
   The physical definition of the log-likelihood-based pseudo-R^2 is debated as some sees it analogously to ordinary least square-R^2 metric which quantifies the proportion of explained variance in linear regression. However, many see pseudo-R^2 more as a measure of goodness-of-fit or association between the predicted and real values [112]–[114]. Among the many proposed Pseudo-R^2, two of the most common are:

   - **McFadden's Pseudo-R^2** (a.k.a. ρ^2) [115]

   $$R^2_{MF} = 1 - \frac{\ln L_c}{\ln L_0} \tag{2.8}$$

   - **Maddala/Cox & Snell's Pseudo-R^2** [116], [117]

   $$R^2_{C\&S} = 1 - e^{(-\frac{2.(\ln L_c - \ln L_0)}{n})} \tag{2.9}$$

4. **Out-of-Sample Accuracy**
   A simple accuracy calculation could be done by determining the rate of correctly classified occurrences over the total prediction. An out-of-sample approach is a more pragmatic and unbiased practice which randomly separate the dataset into train and test sets. 70:30 to 80:20 train-test split proportions are commonly used and mostly adequate [118]–[120].

5. **Confusion Matrix**
   A confusion matrix is useful in the case of imbalance data to show whether the minority class is also well-classified [121]. This matrix maps out the prediction values of a model and the actual values in the format shown in Table 2.2.

TABLE 2.2: Confusion matrix for binary classification

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

From this matrix, several metrics can be derived:

- **Precision**
  The ratio of correctly identified positives over the total predicted positives.

$$TRP = \frac{TP}{TP + FP} \tag{2.10}$$

- **True Positive Rate (TPR)**
  Also known as *sensitivity* or *recall*. A ratio between correctly identified positives over the total observed positives.

$$TRP = \frac{TP}{TP + FN} \tag{2.11}$$

- **True Negative Rate (TNR)**
  Also known as *specificity*. A ratio between correctly identified negatives over the total observed negatives.

$$TNP = \frac{TN}{FP + TN} \tag{2.12}$$

- **False Positive Rate (FPR)**
  The ratio of negatives which are falsely identified as positives over the total observed negatives.

- **Area Under the Curve (AUC) of Receiver Operator Characteristics (ROC)**
  ROC graph plots the TPR against the FPR across all possible cutoff values (i.e., the threshold between classifying an event as positive or negative) as shown in Figure 2.4. The AUC of this graph represents the ability of the model to differentiate the classes. The theoretical best is a vertical graph (i.e., AUC = 1) where a change in the cutoff does not change the sensitivity of the model. Meanwhile, AUC = 0.5 signifies no predictive power and a lower value means that the model performs worse than random chance.

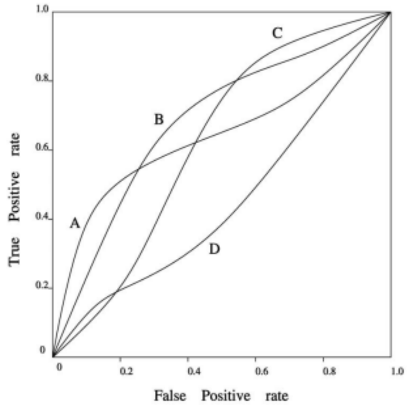


FIGURE 2.4: An example of four ROC curves. Source: Huang & Ling [122]

2.3.2 Collinearity

The term collinearity describes when the independent variables are correlated to each other. A high amount of collinearity may result in the inflation of the regression coefficients' variance which ultimately misleads the identification of relevant predictors [123]. The prevalent practices to detect collinearity include:

1. **Correlation Matrix**
   This matrix simply maps the bivariate correlation between each variable pair. The correlation test depends on the types of the variable pair in question:

   - **Correlation between two continuous variables**
     Both *Pearson* and *Spearman correlation coefficients* measure the bivariate correlation in the scale of -1 to +1 with the sign indicates a negative/positive relationship while the correlation strength is indicated by the absolute number. The main difference lies on the fact that Pearson expect a linear

relationship while Spearman works with a monotonic relationship but not necessarily linear. For normally distributed variables, Pearson and Spearman have similar expected coefficient values [124].

- **Correlation between two categorical variables**
  The *chi-square (χ_2) test of independence* evaluates the χ_2 statistics between groups in a contingency table of a pair of categorical variables. The null hypothesis states that there is significant difference between the two tested groups. Therefore, a significant p-value (i.e., equal or less than alpha value) means that there is in fact no significant difference and the two variables are correlated.

- **Correlation between a continuous and a categorical variable**
  The *point-biserial test* is a Pearson-like correlation test between a continuous and a dichotomous variable. In the case of categorical variables with more than two categories, the variable could be artificially dichotomised through the use of dummy variables.

2. **Variance Inflation Factor (VIF)**
   Essentially, this metric estimates the degree of variance inflation of a regression coefficient due to multicollinearity in the model. This is done by regressing each predictor against other predictors to obtain the R^2 statistics. The VIF value is then calculated as follows:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2.13}$$

Another measure of multicollinearity, Generalised VIF (GVIF), was introduced by Fox & Monette which is suited for when categorical variables are involved [125]. Suppose a regression model of:

$$\boldsymbol{Y} = \beta_0 + \boldsymbol{X_1}\boldsymbol{\beta_1} + \boldsymbol{X_2}\boldsymbol{\beta_2} + \boldsymbol{\epsilon} \tag{2.14}$$

$\boldsymbol{Y}$ = a vector of observations
$\boldsymbol{X_1}$ = a vector which contains related r indicator variables (e.g., each categories of a categorical variable)
$\boldsymbol{X_2}$ = a vector which contains the remaining predictors, excluding the constant
$\boldsymbol{\epsilon}$ = a vector of unobserved error

Thus, GVIF is calculated as follows:

$$GVIF = \frac{det(\boldsymbol{R}_{11})det(\boldsymbol{R}_{22})}{det(\boldsymbol{R})} \tag{2.15}$$

$\boldsymbol{R}_{11}$ = the correlation matrix of $\boldsymbol{X_1}$
$\boldsymbol{R}_{22}$ = the correlation matrix of $\boldsymbol{X_2}$
$\boldsymbol{R}$ = the correlation matrix for all variables, excluding the constant

The measure $(GVIF)^{(1/(2*Df))}$ with Df being the variable's degree of freedom is recommended as these values remain comparable should the predictor variables in question have varying dimensions. The square of $(GVIF)^{(1/(2*Df))}$ is analogous to the regular GVIF/VIF.

2.3.3 Variable Selection

Variable selection is a process of selecting only a subset of the predictor variables to be included in a model. Cassotti & Grisoni [126] identified the purposes of variable selection to be:

- Improving the interpretability through simpler models
- Disregarding insignificant effects, hence reducing noise
- Improving the model's predictive ability
- Speeding up the model's processing time

The following selection algorithms are frequently used in research:

1. **Stepwise Regression**
   Stepwise regression is an iterative process which tests different combinations of predictors until there is no significant improvement in the set performance criterion (usually AIC or BIC). There are different procedures of stepwise regression:

   - **Forward Stepwise Selection**
     This process begins with a null model (a model of only an intercept) and adds one variable at each iteration. The variable which produces the most significant improvement compared to the previous iteration is carried forward to the next iteration. The process stops when an addition of variable no longer significantly improves the model.

   - **Backward Stepwise Selection** (a.k.a Backward Elimination)
     This process begins with the full model M_p (a model containing all of the predictors p in consideration) and removes the least significant variable at each iteration to produce $M_{p-1}, M_{p-2}....M_0$. The AIC/BIC values of all $M_0....M_p$ are compared and one model is selected.

   - **Bidirectional Stepwise Selection**
     This combines both forward and backward selection. Starting with a null model, at each iteration, after adding a significant variable, another variable that is no longer significant is removed.

   Examples of stepwise methods usage for logistic regression: [127]–[129]

2. **Lasso Regression**
   Least Absolute Shrinkage and Selection Operator (Lasso or LASSO) is a *regularisation* method—it penalises a model for having too many variables and removes the less contributive variables by shrinking their regression coefficients to zero. When applied to a logistic regression, Lasso adds a penalty term to the likelihood function as such:

   $$L + \lambda \sum |\beta_i| \tag{2.16}$$

   L = the likelihood function in Equation 2.3
   β_i = the regression coefficient vector
   λ = the shrinkage parameter

   A rise in λ increases the bias while decreases the variance of the model (see the subsection 2.3.1 on bias-variance tradeoff). The optimum λ is typically found through *cross-validation* (CV) minimising the total MSE [130], [131]. However, the *one standard error* λ—the value which yields higher regularisation yet still lies within one standard error of the minimum cross-validated MSE—is also often preferred [132]. Minimising the penalised likelihood in Equation 2.16 shrinks the coefficients which contribute to increasing L (i.e., the ones that contribute most to the error) to zero. One of the advantages of using Lasso penalisation is its ability to deal with multicollinearity [133]. However, the significance of its coefficient estimates could not be simply computed [134].

Examples of penalised logistic regression using Lasso technique: [133], [135], [136]

Chapter 3

Study Area

Chicago, Illinois is the third-most populous city in the United States with the population of 2,696,561 as per 2021 census [137]. Having a sprawling urban form, the city has a wide suburban area [11]. This, coupled with Chicago's mainly radial transportation network (see Figure 3.1), means that there's a large part of Chicago where traveling with public transit is not always convenient. About 47.6% of Chicago residents commute by driving, while only 29% take the public transit or walk [137]. The city's rising transportation emission despite the decreasing population density is a further evidence of Chicago's car dependency [138].

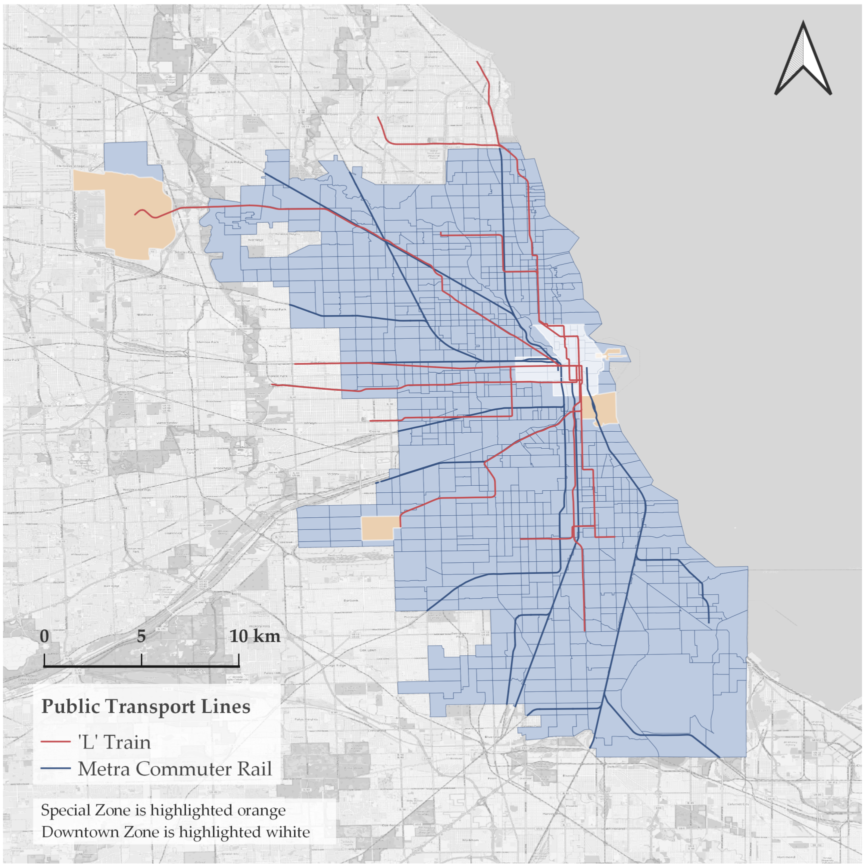

0
5
10 km
Public Transport Lines
'L' Train
Metra Commuter Rail
Special Zone is highlighted orange
Downtown Zone is highlighted wihite

FIGURE 3.1: Map of Chicago's census tracts, Downtown Zone, and public transport network

This is also reflected in Chicago's ride-hailing scene which saw an average of 9.2 millions ride-hailing trips monthly right before the COVID-19 outbreak [139]. Despite a drop in number at the start of the pandemic, it has been making a recovery slowly since (see Figure 3.2).

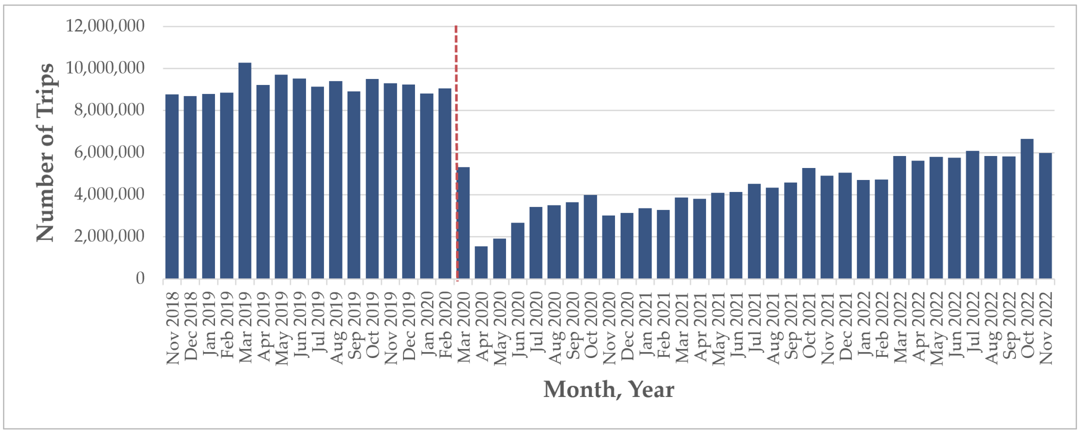


FIGURE 3.2: The total monthly ride-hailing trips over the last few years in Chicago. The red dashed line marks the beginning of COVID-19 pandemic in Chicago. Source: City of Chicago [139]

3.1 Taxes Related to Ride-hailing

In the effort to battle congestion, the city has imposed Ground Transportation Tax (GTT) which also applies to ride-hailing trips. See Table 3.1 for the values as of January 2018 and as of January 2020 (current). Note that the city imposes a much higher fee for trips to and from Special Zones which include airports. There is also another additional tax for the so-called Downtown Zone (refer to the highlighted areas in Figure 3.1) which applies on weekdays 6AM - 10PM. Moreover, a tax reduction was granted for shared trips to incentivise more ride-pooling. The listed fees also include the $0.02 TNP Administration Fee and $0.10 TNP Accessibility Fund Fee.

TABLE 3.1: Taxation of ride-hailing trips in Chicago.
Source: City of Chicago [140], [141]

		As of 2019	As of 2020	
			Trip without Downtown Zone Surcharge	**Trip with Downtown Zone Surcharge**
Non-Pooled	**Regular Trip**	$0.60	$1.25	$3.00
	Trip to/from Special Zone	$5.60	$6.25	$8.00
	Trip with Wheelchair Accessible Vehicle (WAV)	Not Differentiated	$0.55	$0.55
Pooled	**Regular Trip**	$0.60	$0.65	$1.25
	Trip to/from Special Zone	$5.60	$5.65	$6.25
	Trip with Wheelchair Accessible Vehicle (WAV)	Not Differentiated	$0.55	$0.55

3.2 Chicago's Ride-hailing Scene during the COVID-19 Pandemic

The very first case of COVID-19 in Chicago was recorded on the 4th of March 2020 [142]. Since then, Chicago went through multiple levels of restrictions before alleviating most of them at the end of February 2022. The Figure 3.3 below summarises Chicago's COVID-19 timeline based on the Orders issued by the City of Chicago [143] and the State of Illinois [144].

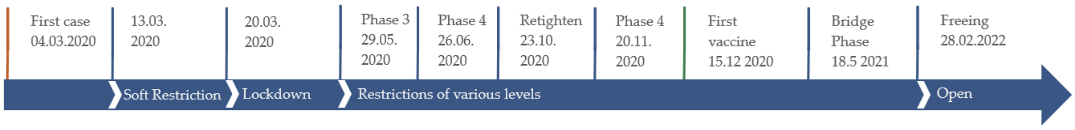


FIGURE 3.3: COVID-19 timeline in Chicago. Source: City of Chicago [143], State of Illinois [144]

Due to the pandemic, the main TNPs in Chicago, Uber and Lyft, suspended the ride-pooling services in the United States from March 2020 [145]. Lyft reintroduced the service in July 2021 for Chicago [146], while Uber only followed suit in February 2022 [147].

Chapter 4

Methodology

Due to the large size, the data that is involved in this study were always stored and processed in subsets. This was to make sure that no knowledge was lost and all of the available information could be put to use without requiring high-performance computer. The methodological workflow of this thesis is illustrated by Figure 4.1. In this thesis, 'WTP trips' refer to the trips which were authorised for pooling regardless if it was successfully matched or not, while 'pooled trips' refer to the ones that were successfully pooled.

4.1 Data Collection

The data that were employed in this thesis consisted of the main ride-hailing trip data and the exogenous factors data (e.g., demographic, spatial, weather, and pandemic progression) which were sourced from multiple open platforms. The data were split into two study periods: pre-outbreak (January 2019 - December 2019) and post-outbreak after the pooling services returned (August 2021 - November 2022). This section outlines the sources of each type of data.

4.1.1 Ride-hailing Trips

This data was obtained from the Chicago Data Portal[a], which is a part of the city's Open Data initiatives. The Transportation Network Providers [139] dataset comprises of the records of trips conducted by ride-hailing companies dating as early as November 2018. The overall data consists of 293,304,308 individual trips at the time of download. The information contained in the dataset includes the trip attributes, timestamps of the trips, and the approximate locations of the pickup and dropoff.

For privacy protection reasons, the locations are only precise to the level of corresponding census tract or community area. These information may also be empty for locations outside the city's border. For the purpose of this study, only the records that had both the pickup and dropoff census tract information were considered (i.e., intra-city trips). Moreover, the timestamps, fare and tip are rounded.

[a] data.cityofchicago.org

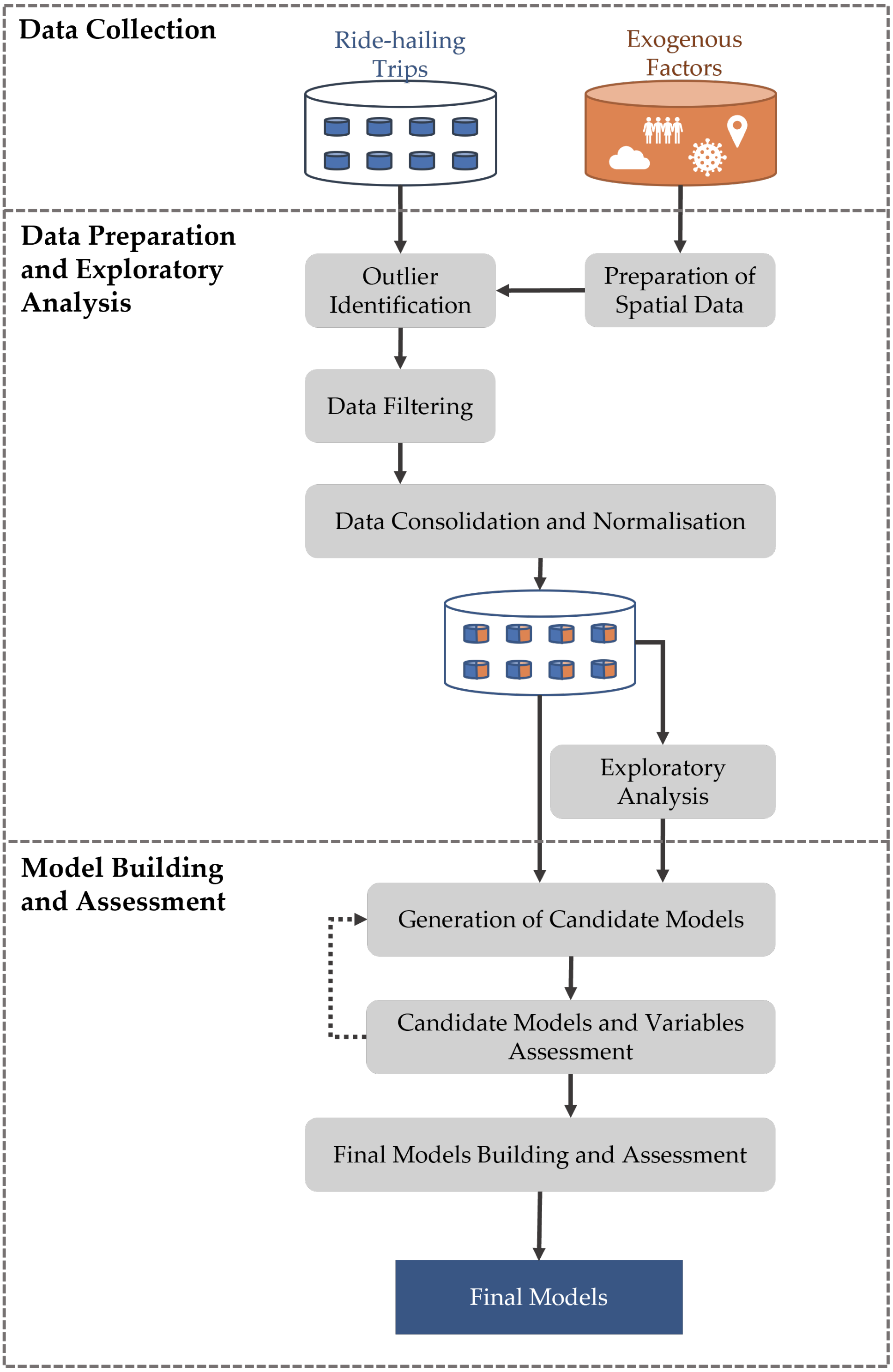

Data Collection
Ride-hailing Trips
Exogenous Factors
Data Preparation and Exploratory Analysis
Outlier Identification
Preparation of Spatial Data
Data Filtering
Data Consolidation and Normalisation
Exploratory Analysis
Model Building and Assessment
Generation of Candidate Models
Candidate Models and Variables Assessment
Final Models Building and Assessment
Final Models

FIGURE 4.1: Methodological framework

4.1.2 Demographic Data

This thesis gathered the information regarding Chicago's population number, employment, age, income, and vehicle ownership from the American Community Survey carried out by the US Census Bureau [137].

4.1.3 Spatial Data

Land Use Information

The zoning information of Chicago was obtained from the Chicago Data Portal in the form of a polygon *shapefile* [148]. Based on the zoning and land use ordinance in the Municipal Code of Chicago [149], this thesis classified land uses into nine categories:

- Residential (code R)
- Business (code B)
- Commercial (code C)
- Downtown (code D)
- Manufacturing (code M)
- Planned Manufacturing (code PMD)
- Planned Development (code PD)
- Transportation (code T)
- Parks and Open Spaces (code POS)

Boundaries and Point Locations

The following datasets were obtained from Chicago Data Portal.

1. **Boundaries:**
    - Individual census tracts [150]
    - Chicago's central business district (CBD)
2. **Point Locations:**
    - Public transit stops including buses, intracity 'L' trains, and 'Metra' commuter trains [151]–[153]
    - Locations of each crime reported within Chicago during the observation periods [154]. Pre- and post-COVID-19 outbreak periods were differentiated.

4.1.4 Weather Data

The daily average temperature, average wind speed, and total precipitation for each day over the study period were gathered from a weather reporter website [155].

4.1.5 Public Holidays in Chicago

Dates of Chicago's public holidays were sourced from Chicago's official publication [156] and historical public holidays website [157].

4.1.6 Pandemic Progression

The daily rolling averages of COVID-19 cases, hospitalisations, and deaths, along with the cumulative number of complete vaccination series associated with Chicago residents were sourced from Chicago Data Portal [142], [158]

4.2 Data Preparation

4.2.1 Outlier Identification and Data Filtering

The univariate distribution of each variable was visualised from a sample dataset to identify outliers. Subsequently, the outliers were filtered out to ensure a robust dataset which contained the information that was sought for by this study.

4.2.2 Preparation of Spatial Data

The spatial data and any information related to each census tract were processed and prepared using the software *QGIS*.

1. **Tract Characteristics**
   Geographical characteristics of each census tract including its area and its centroid distance to the central business district's centroid were computed.
2. **Land Use Information**
   The percentage area of each type within each census tract were calculated. Subsequently, the tract's prevailing land use were derived along with the tract's *Entropy Index* (Equation 4.1) which described the land use mix in the scale of 0 to 1.

   $$Land\ Use\ Entropy = -(\sum_{j=i}^{k} P_j \ln(P_j)) \,/\, \ln(k) \tag{4.1}$$

   P_j = the percentage area of land use j
   k = the total number of land use types
3. **Public Transit Access**
   The measure of access to public transit was defined by taking the number of public transit stops per unit area of each census tract.
4. **Crime Rate**
   The crime rate at each census tract were measured by counting the crime point locations that fell within each tract and taking the monthly average.
5. **Downtown Zone Boundary**
   The boundary to the specially-taxed zone was created in *QGIS* based on the given map description by the City of Chicago [141].

4.2.3 Consolidation of Data

The data preparation process included consolidating the trip data with the spatial/spatiodemographic information at the pickup and dropoff tracts of the trip as well as the temporally-varying data (e.g., weather and COVID-19 conditions) at the day of the trip.

4.2.4 Data Normalisation

Several spatial variables such as population and public transit stops were standardised to per unit area. Units were also converted to metric system. Moreover, to ensure that all of the continuous variables (and the associated regression coefficients) were in comparable scales, the values were normalised using the standard score method. However, as the data were stored in subsets, the weighted mean and the weighted standard deviation of each variable were utilised. Thus, for a certain variable:

$$\bar{x} = \frac{\sum_{i=1}^{N}(w_i x_i)}{\sum_{i=1}^{N} w_i} \tag{4.2}$$

$$\bar{\sigma} = \sqrt{\frac{\sum_{i=1}^{N} w_i (x_i - \bar{x})^2}{\frac{(M-1)}{M}\sum_{i=1}^{N} w_i}} \tag{4.3}$$

$$z = \frac{X - \bar{x}}{\bar{\sigma}} \tag{4.4}$$

$\bar{x}$ = weighted mean
x_i = the mean value subset i
w_i = the weight (i.e., the number of observations) of subset i
N = total number of subsets i
$\bar{\sigma}$ = weighted standard deviation
M = the number of non-zero weights
z = normalised score
X = raw score

4.3 Exploratory Analysis

4.3.1 Spatial and Temporal Variations

The spatial distribution of pickup rates, dropoff rates, spatiodemographic measures, land use entropy, public transit access, and crime rate were visualised in *QGIS* to identify any pattern. Similarly, the temporal variations of daily trips (both WTP and pooled trips) were plotted along with variables related to weather and COVID-19 pandemic. However, due to time constraint, the results of temporal variation analysis are not presented in this report.

4.3.2 Multivariate Distributions

As the data were stored in subsets, in order to achieve distributions involving the whole dataset, intermediary tables were generated to store summarised information from each subset. Subsequently, various charts were produced from these intermediary tables. This procedure was done automatically using the programming language *Python 3.10.4* with the Integrated Development Environment (IDE) *Spyder 5.4.0*. Trips which began or ended on a Special Zone were isolated and analysed separately to investigate whether differing patterns exist.

4.3.3 Correlation Analysis

As the variable set in this study involved both continuous and categorical data types, three different correlation matrices were produced:

1. **Pearson's correlation test between continuous variables**
   This study opted for Pearson's test instead of Spearman's due to the simultaneous use of Point Biserial test which is a Pearson-like test for a continuous-dichotomous variable pair. Thus, using Pearson's enabled comparability of the correlation coefficients. The test was done using the function *corr()* from the *Pandas* library of *Python*.

2. **Chi-square (χ_2) test of independence between categorical variables**
   The test was conducted by generating contingency tables between each pair of categorical variables through the use of *crosstab()* function from the *Pandas* library. Afterwards, the *chi2_contingency()* function from the library *Scipy* was employed to compute the χ_2 statistic and the p-value between each pair. The p-value matrix was then assessed.

3. **Point biserial test between categorical and continuous variables**
   Prior to applying the test, all of the non-dichotomous categorical variables were artificially dichotomised through the use of dummy variables. The function *pointbiserialr()* from the *Scipy* library was utilised to compute the correlation coefficient between each dichotomous-continuous variable pair.

4.4 Model Building and Assessment

The aim of this procedure was to obtain logistic models which sufficiently classify:

- whether a trip with its features would be authorised for pooling or not –> **'WTP' models**
- whether a trip with its features that was already authorised for pooling would be successfully pooled or not –> **'Pool' models**

This thesis proposes a methodology in which N subsets of data are fitted to a binary logistic function separately. This would result in the generation of N candidate models which may involve different combinations of variables and varying coefficient values. The variance and the significance of these coefficients across the whole set of candidate models lay the foundation to the final variables selection, and hence, the development of the final model. The whole modelling workflow is further detailed by Figure 4.2 below. This model building procedure was conducted separately on the pre-outbreak dataset and the post-outbreak dataset.

4.4.1 Random Subsampling

The trip data was initially obtained and stored in M subsets according to its timestamps. The processed data were then reorganised by splitting each subset into N equal portions without replacements. Each portion was randomly assigned to one of the new N subsets which would be utilised for modelling. This made sure that each new subset was representative of the whole dataset. The value N was dependent on the original dataset size, as the number of observations within each subset was kept approximately constant at one million. This constant sample size ensured that each model approximation in this study had a similar level of reliability [159]. Exception was the input data for the post-outbreak 'Pool' models where subsampling was not done, as the total observations of WTP trips during the post-outbreak period was significantly smaller.

4.4.2 Generation of Candidate Models

This thesis utilised and compared two variable selection techniques:

1. **Backward Stepwise Selection**
   When there is a certain degree of collinearity, the effect of one predictor may only be significant in the presence of another variable. Therefore, this study opted for backward elimination technique as recommended by Mantel [160] as it begins by considering the effects of all predictors simultaneously. Despite being a faster procedure, forward selection (also bidirectional selection) tends to exclude said correlated predictors altogether and miss to capture potentially significant effects [160].

   The procedure was performed by integrating the functions *glm()* and *step()* from the *stats* package of the statistical programming language *R*. The input sample was first split into 80% training data and 20% test data. The *glm()* function took input of the full logistic formula (Equation 2.2), the training data, and the parameter *family* set to "binomial" to return a *glm* object of the full model. The *glm* object was fed into the *step()* function along with the parameters *direction* set as "backward" and *k* as log of observation number to signify using BIC as selection criterion. BIC was used rather than AIC for stricter penalty. The remaining 20% of data was then used to test the performance of the fitted model (see Section 4.4.4).

2. **Lasso Regression**
   Similarly, 80% of the input sample was used to train the model, while 20% was used to test the candidate model. The package *glmnet* in *R* was utilised. The value of λ was calculated using the function *cv.glmnet()* which conducted *k-fold cross validation* to return the optimum λ value (*lambda.min*) and the *one-standard-error* λ (*lambda.1se*). To yield higher regularisation and avoid overfitting on a certain input sample, the *lambda.1se* was utilised to fit the model. This was done with the function *glmnet()* by setting the parameters *alpha* to 1 to signify Lasso method and *family* to "binomial".

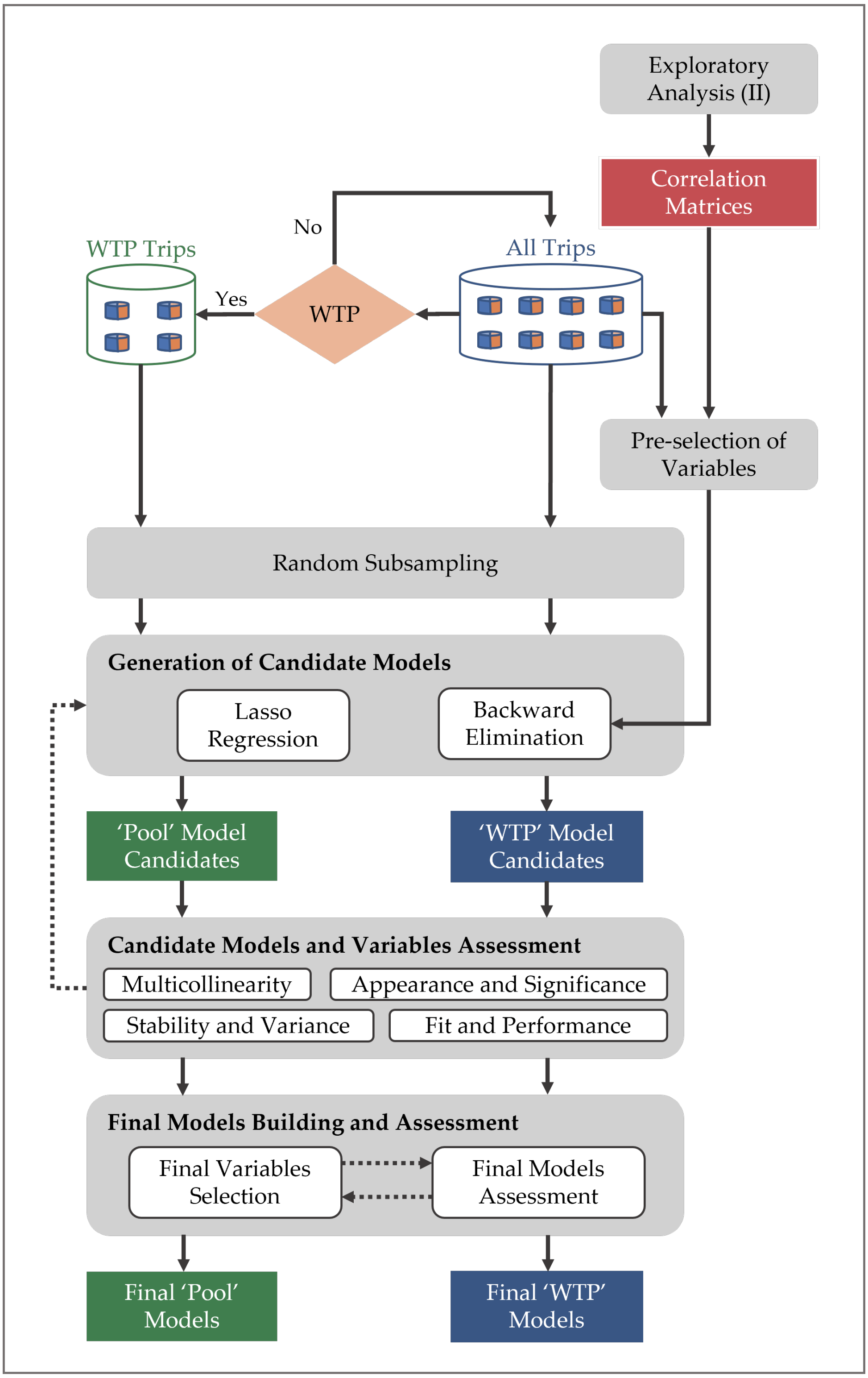


FIGURE 4.2: Modelling workflow

4.4.3 Pre-selection of Variables

Stepwise regression method is susceptible to collinearity problems as it exacerbates the effects [161]. As a consequence, collinearity was addressed prior to using this technique. Based on the correlation matrices between the variables, the predictor variables were pre-selected to only involve those with low degrees of collinearity.

For correlations between continuous-continuous and continuous-categorical variable pairs, this thesis adopted the correlation coefficient threshold of $|r| > 0.7$ for variable removal based on previous studies [123], [162], [163]. Taking notes from Gehrke et al. [84], when two variables were highly correlated, the one which has higher absolute coefficient correlation to the dependent variable of interest was kept.

In the case of correlations between categorical variables, as the relative strengths could not be measured, pre-selection was done based on domain knowledge.

4.4.4 Candidate Models and Variables Assessment

Multicollinearity

Each candidate model was subjected to multicollinearity test using the GVIF method. As mentioned in Section 2.3.2, the square of $(GVIF)^{(1/(2*Df))}$ is analogous to the regular VIF. Based on the literature [164]–[166], this thesis regarded variables with multicollinearity values greater than 5 to be potentially problematic. In such case, the corresponding full model composition would be amended.

Stability and Variance of the Coefficients

For each set of candidate models, the regression coefficients were plotted to show the mean, maximum, and minimum values across all N subsets. Thus, the variance of each variable and the stability of its value sign (positive or negative) could be observed. In the case of sign switching, the corresponding full model composition would be amended.

Proportion of Appearances and Significance of the Coefficients

This thesis assessed the proportion of times each variable was involved across the N candidate models. Furthermore, for candidate model sets generated through stepwise selection, the proportions at which the coefficient estimates were statistically significant were also investigated. Through these, the degree of importance of each predictor could be visualised.

Goodness-of-fit and Predictive Performance

The performance of each candidate model was also monitored through:

- **McFadden's Pseudo-R^2**
  The value ranging from 0.2 to 0.4 is generally regarded as excellent fit [167].

- **Out-of-Sample Accuracy**
- **Out-of-Sample AUC of ROC**
  In general the value of 0.7 - 0.8 indicates acceptable performance, while a greater value would be considered excellent [168].

4.4.5 Final Models Building and Assessment

Final Variables Selection

The final models would ideally consist of predictors which were stable and exhibit small variance. Appearance and relatively high significance of greater than 50% were also set as a selection criteria.

Final Models Assessment

The final models were re-fitted using regular binomial logistic regression (without penalisation) with the respective N subsets as inputs. Again, each subset was split 80:20 for training and testing respectively. The multicollinearity within the final models along with the out-of-sample predictive performance and goodness-of-fit across the corresponding dataset would be re-assessed as per Section 4.4.4.

Chapter 5

Results and Discussion

This section presents and discusses the results obtained from each step of the methodology elaborated in Chapter 4.

5.1 Data Collection and Data Preparation

The data were collected from open-sources as previously explained. The following charts give a brief descriptions of the trip data during the pre-outbreak and post-outbreak period. Table 5.1 shows the total and average monthly trips while Figure 5.1 and 5.2 show the proportions of WTP trips and pooled trips.

TABLE 5.1: Ride-hailing trips within the two study periods

Period	Total Ride-hailing Trips	Average Trips per Month
Pre-outbreak (01/2019-12/2019)	111,850,744	9,320,895
Post-outbreak (08/2021-11/2022)	86,953,825	5,796,922

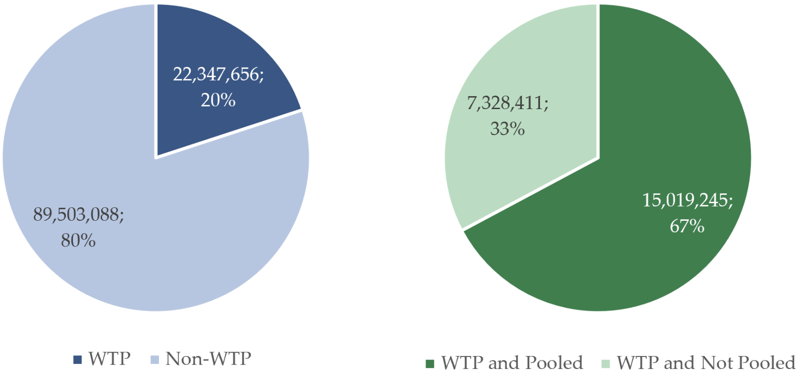


FIGURE 5.1: Pre-outbreak proportions of ride-hailing trips

A drastic drop in the proportions of trips authorised for pooling (WTP trips) among all of the ride-hailing trips between pre-outbreak and post-outbreak periods can be observed in these charts. Although more modest, there was also a drop in the proportion of successfully pooled trips among all of the trips authorised for pooling.

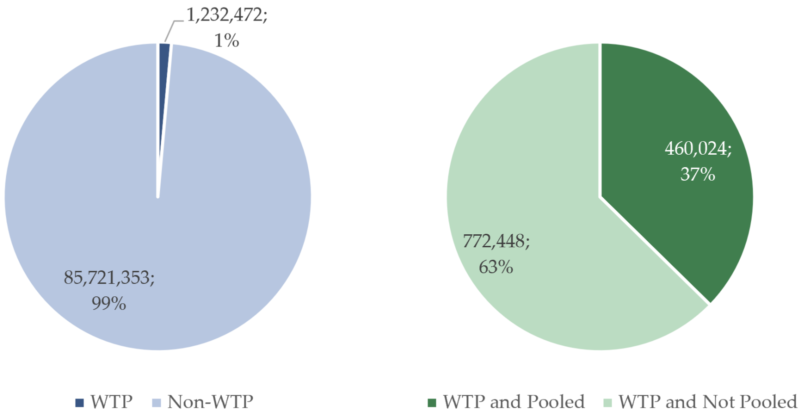


FIGURE 5.2: Post-outbreak proportions of ride-hailing trips

5.1.1 Outlier Identification and Data Filtering

Figure 5.3 and 5.4 plot the distributions of the trip attributes from a randomly sampled trip dataset.

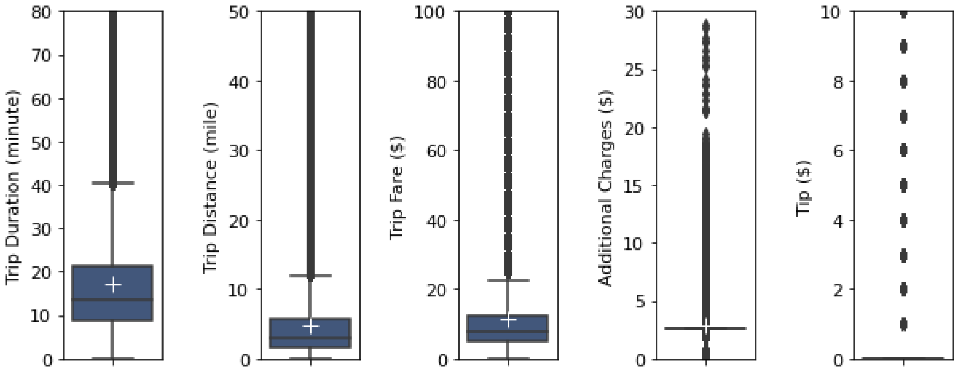


FIGURE 5.3: Boxplots of trip attributes

Based on these results, the data filtering was applied so that the trip dataset should only include:

- datapoints which have both the pickup and dropoff census tract information
- trips whose distance lies between 0.5 - 20 miles (0.8 - 32 km)
- trips whose duration lies between 60 - 3600 seconds (1 - 60 minutes)
- trips whose fare were at most $30 and additional charges at most $10

Consequently, 67,194,023 pre-outbreak ride-hailing trips (out of which 11,211,594 are WTP) and 39,420,596 post-outbreak trips (out of which 355,513 are WTP) remained.

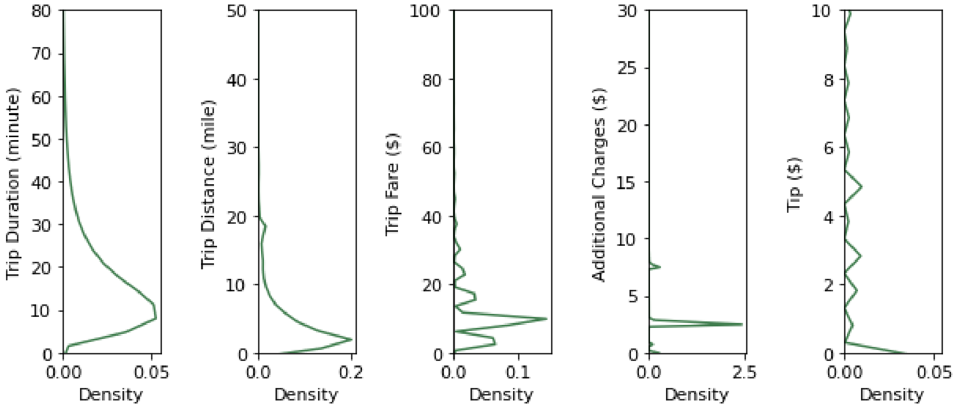


FIGURE 5.4: Density plots of trip attributes

5.2 Exploratory Analysis

This section presents the results of the exploratory analysis that was conducted on the prepared data.

5.2.1 Descriptive Statistics of the Variables

The variable descriptions could be found in Appendix A. Meanwhile, the complete table of the variables' descriptive statistics could be found in Appendix B. This descriptive statistic table compares the variables' statistics between the two study periods.

5.2.2 Spatial Variations

Figure 5.5 shows the spatial distribution of the trip pickups. It can be observed how the closer the tract to the central CBD/downtown area, the higher the pickup frequency. This persisted even after the COVID-19 outbreak, albeit at much lower magnitude. There is no significant difference between pickup and dropoff distribution. Meanwhile, Figure 5.6 shows the distribution of some spatiodemographic characteristics of Chicago.

From these figures, initial speculations were drawn regarding possible relationships between WTP or pooled trip probabilities and the corresponding tracts' characteristics. For example, it can be observed how pickup rates are higher at areas with higher income level, hinting on possible income's effect on WTP or pooling odds.

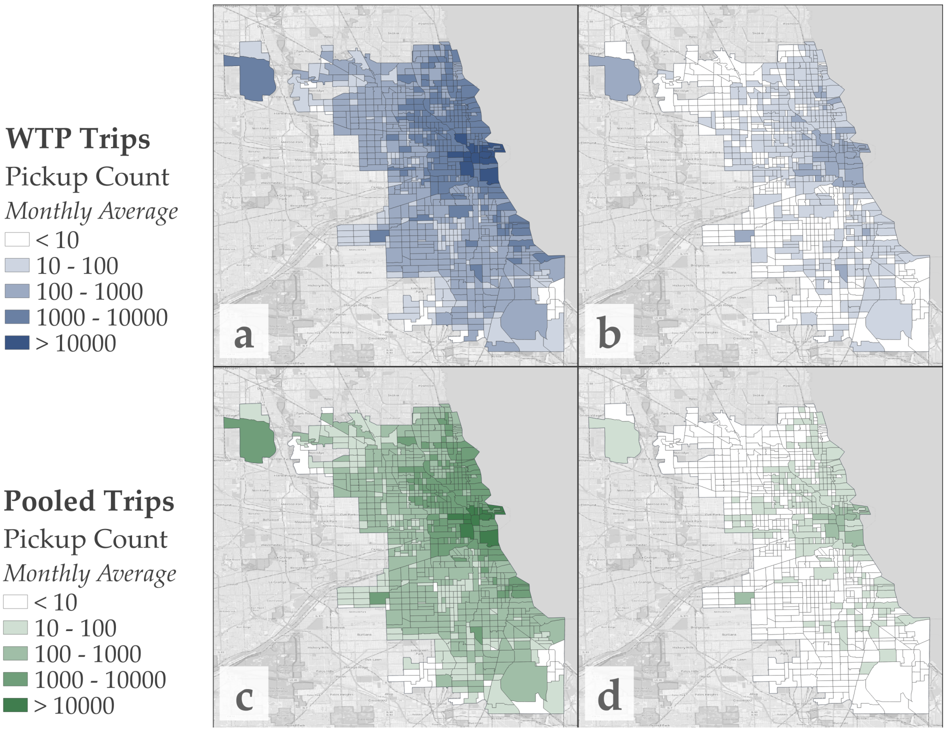


FIGURE 5.5: Average monthly pickup count of a) pre-outbreak WTP trips, b) post-outbreak WTP trips, c) pre-outbreak pooled trips, and d) post-outbreak pooled trips

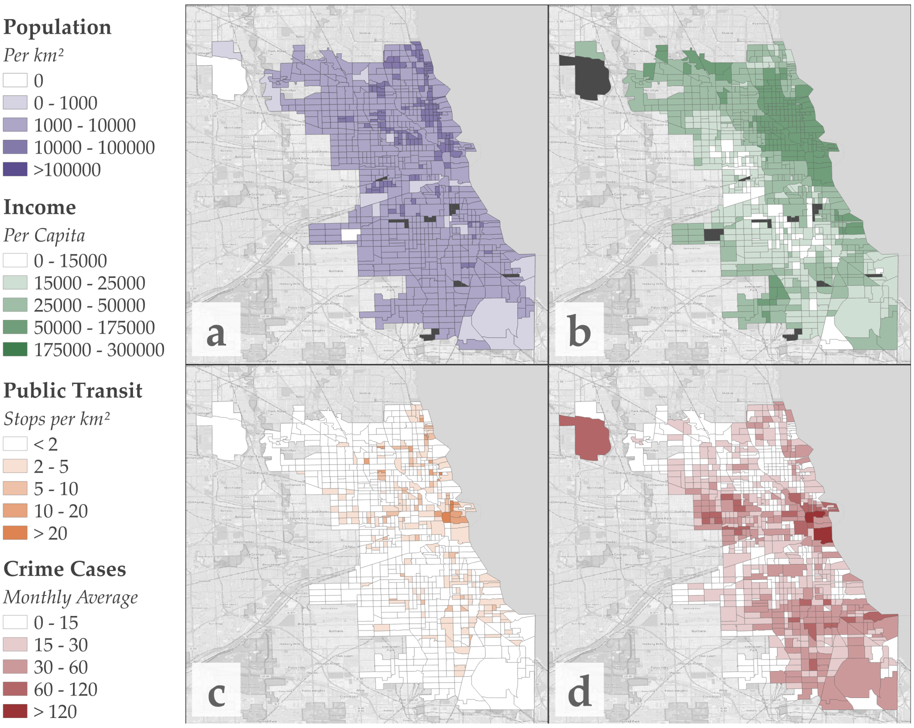


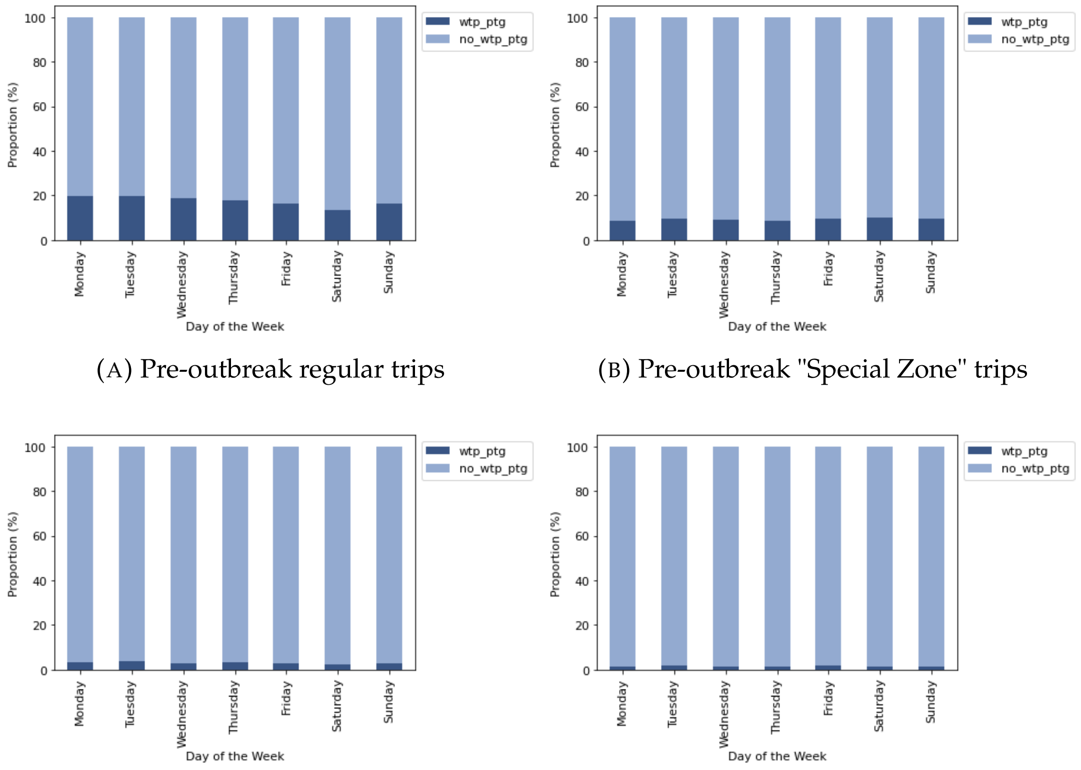
FIGURE 5.6: Spatial variation of a) population density, b) income level, c) public transit access, and d) crime rate in Chicago

5.2.3 Multivariate Distributions

During this analysis, the trips beginning or ending at a Special Zone were isolated to analyse separately. Pre-outbreak and post-outbreak data are also compared below. Distributions pertaining to demand proportions and fare rate variations are discussed below, while more of the results could be found in the Appendix C.

WTP Trip Proportion over Different Days of the Week

Figure 5.7 reiterates the mentioned finding about the big drop between the WTP proportions before and after the outbreak. Additionally, it also shows how across all week, there is similar proportion of WTP trips, while it is generally lower for "Special Zone" trips. Note that 'wtp_ptg' in the figures refers to proportion of WTP trips while 'no_wtp_ptg' refers to the proportion of non-WTP trips.



(A) Pre-outbreak regular trips

(B) Pre-outbreak "Special Zone" trips

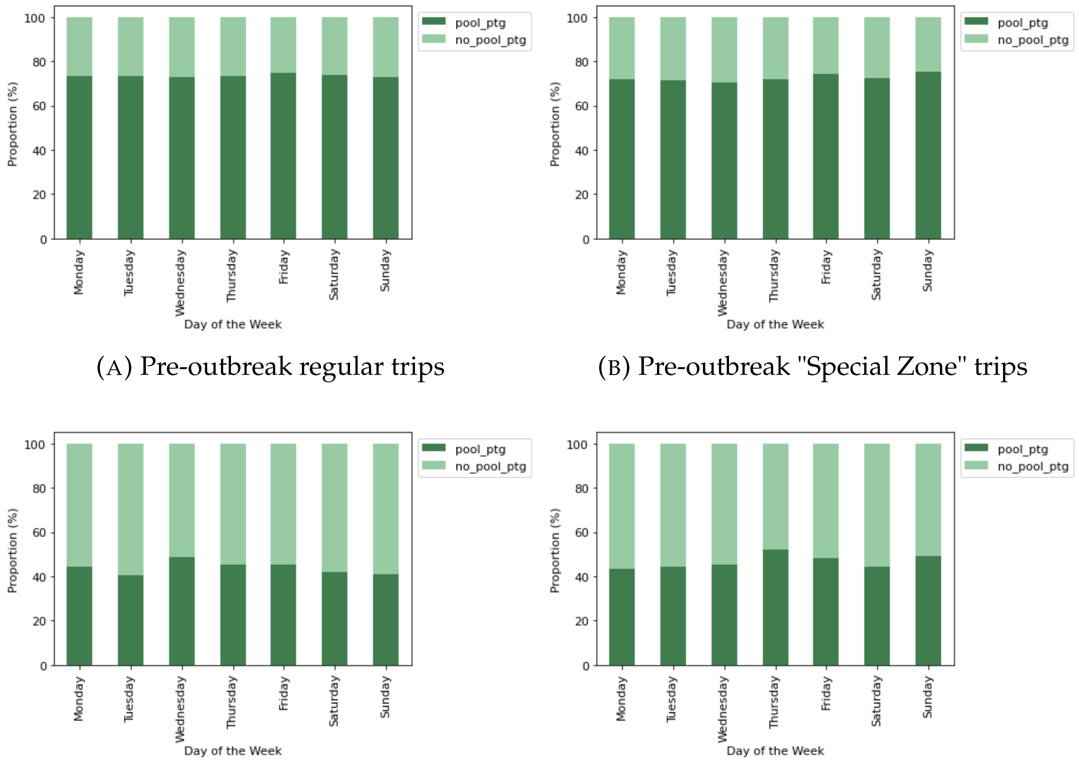
(C) Post-outbreak regular trips

(D) Post-outbreak "Special Zone" trips

FIGURE 5.7: Variations of WTP trip proportion over different days of the week

Pooled Trip Proportion over Different Days of the Week

Similarly, Figure 5.8 reiterates the mentioned finding about the considerable drop of pooled trip proportion before and after the outbreak. However, it doesn't show a significant difference between regular trips and "Special Zone" trips. Note that 'pool_ptg' in the figure refers to proportion of pooled trips among WTP trips while 'no_pool_ptg' refers to the proportion of non-pooled trips among WTP trips.



(A) Pre-outbreak regular trips

(B) Pre-outbreak "Special Zone" trips

(C) Post-outbreak regular trips

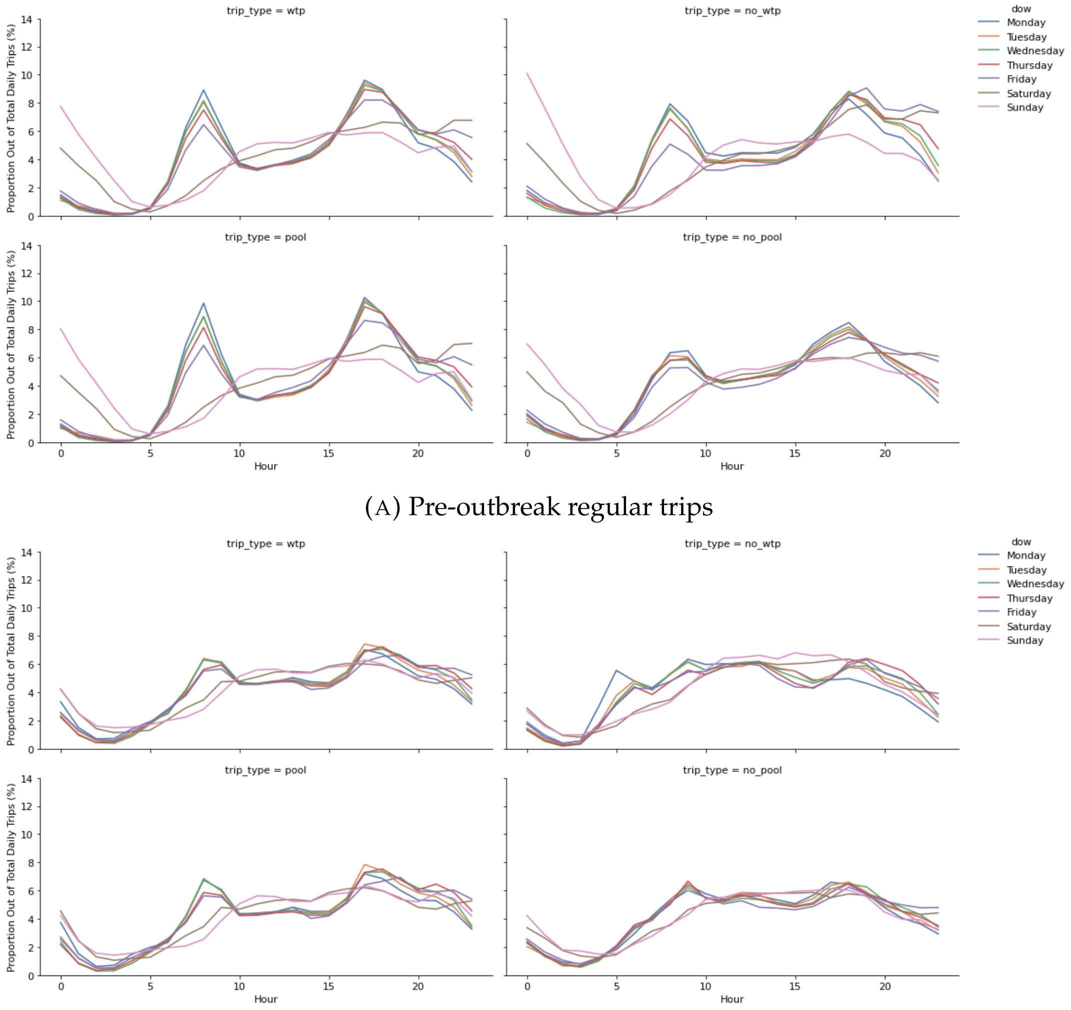
(D) Post-outbreak "Special Zone" trips

FIGURE 5.8: Variations of pooled trip proportion over different days of the week

Hourly Demand over Different Days of the Week

It can be observed in all cases in Figure 5.9 and 5.10 how during the weekdays, the demand peaks in the morning around 8AM and again in the early evening at around 5 - 6PM. Meanwhile, on the weekend, the demand peaks at midnight (12AM). The peaks seem to be less defined for "Special Zone" trips, indicating a more evenly distributed hourly demand for these trips. This pattern persists even in the post-outbreak dataset.

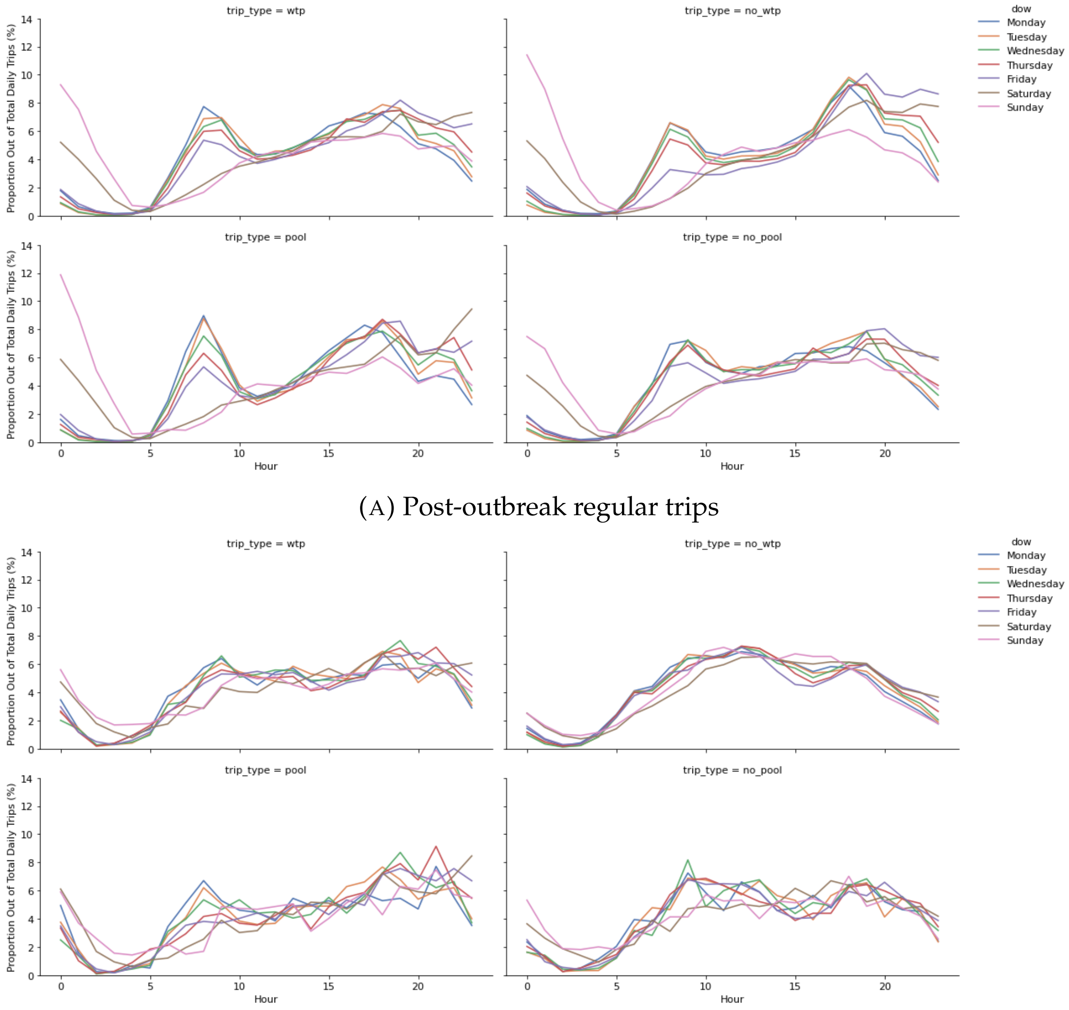
The distributions are differentiated by trip type. ’Wtp’ refers to trips authorised for pooling, ’no_wtp’ refers to trips not authorised for pooling, ’pool’ refers to trips successfully pooled, while ’no_pool’ refers to unsuccessfully pooled WTP trips. Note that the y-axis indicates the proportion out of the total trips in the corresponding day.



(A) Pre-outbreak regular trips

(B) Pre-outbreak "Special Zone" trips

FIGURE 5.9: Variations of hourly demand of different trip types differentiated by days of the week (pre-outbreak)



(A) Post-outbreak regular trips

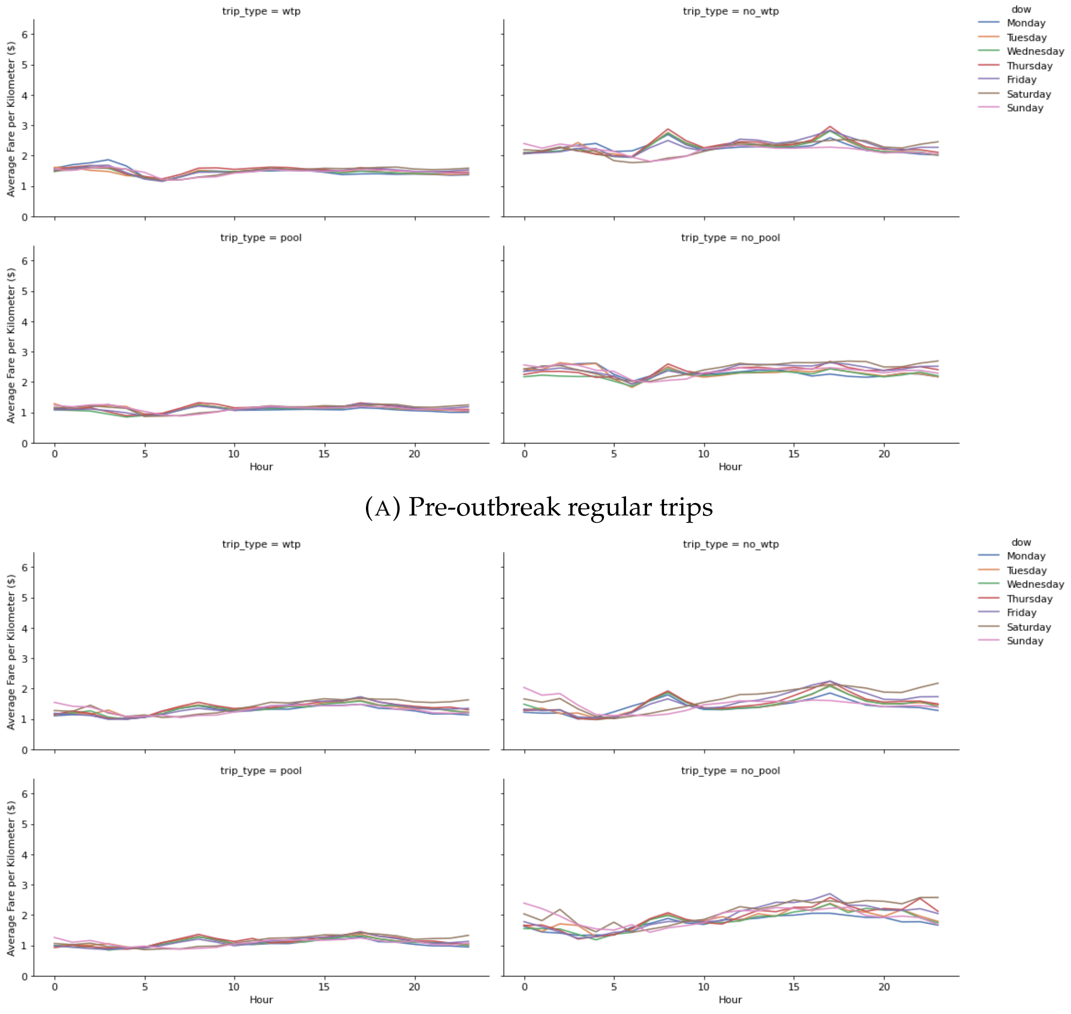
(B) Post-outbreak "Special Zone" trips

FIGURE 5.10: Variations of hourly demand of different trip types differentiated by days of the week (post-outbreak)

Hourly Fare Rate over Different Days of the Week

Figure 5.11 and 5.12 show an evidence of surge pricing during peak hours. Overall, pooled trips have cheaper fare rate compared to non-pooled trips. The cheaper rate only applies if the pooling is successful. Compared to pre-outbreak, the fare rate is higher post-outbreak and the hourly pattern is not as predictable. There is no significant difference between regular fare rate and the "Special Zone" fare rate which shows that "Special Zone" taxing is only applied through additional charges.

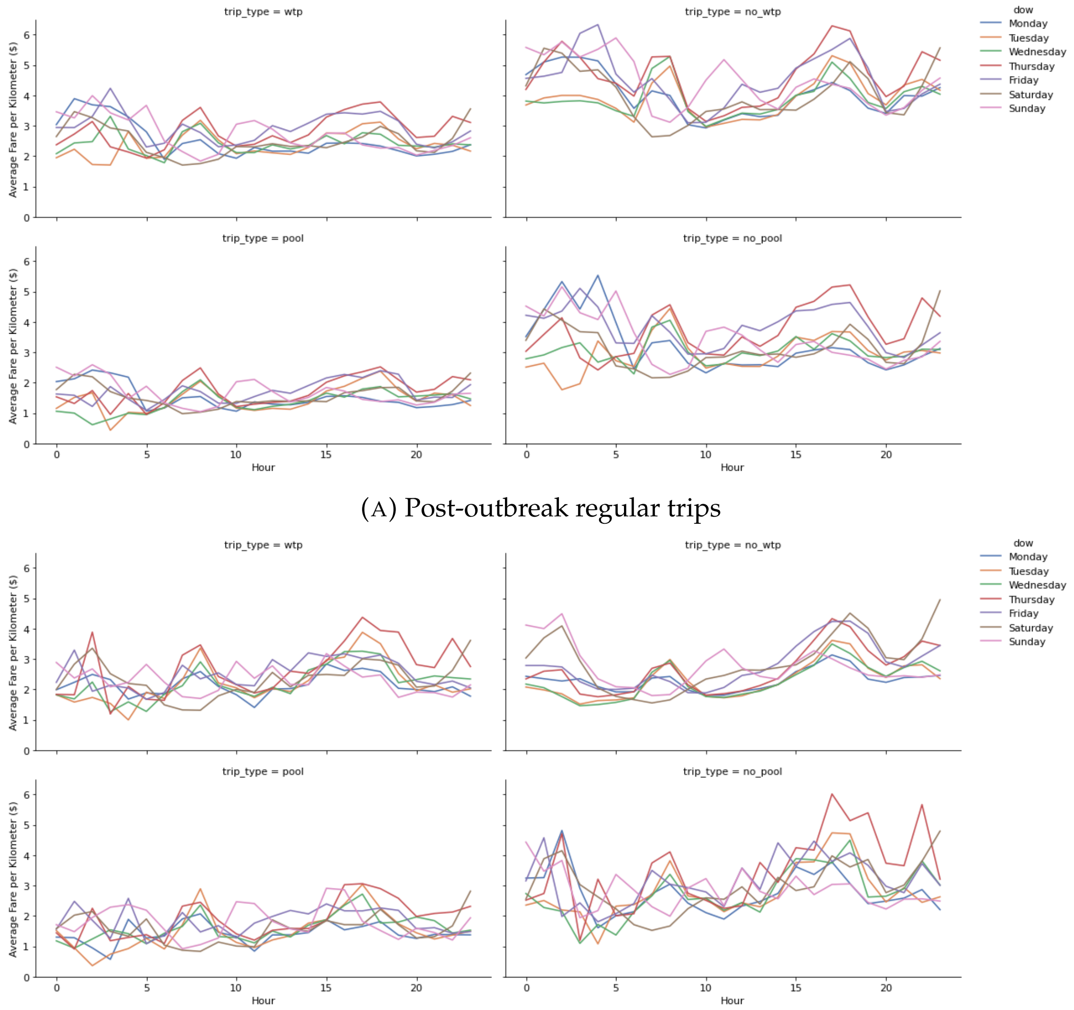
As before, the distributions are differentiated by trip type. 'Wtp' refers to trips authorised for pooling, 'no_wtp' refers to trips not authorised for pooling, 'pool' refers to trips successfully pooled, while 'no_pool' refers to unsuccessfully pooled WTP trips. Note that it plots the average fare per kilometer on the y-axis.



(A) Pre-outbreak regular trips

(B) Pre-outbreak "Special Zone" trips

FIGURE 5.11: Variations of hourly fare rate of different trip types differentiated by days of the week (pre-outbreak)



(A) Post-outbreak regular trips

(B) Post-outbreak "Special Zone" trips

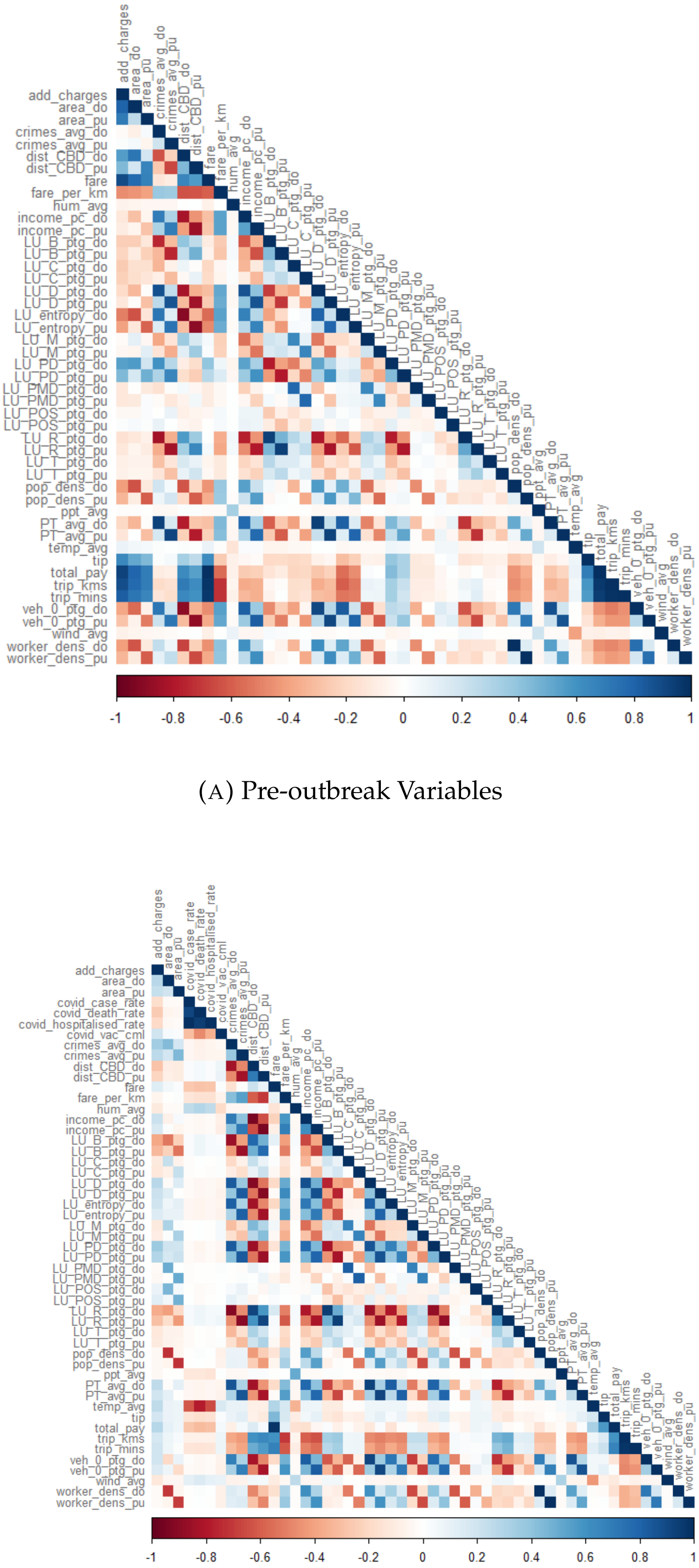
FIGURE 5.12: Variations of hourly fare rate of different trip types differentiated by days of the week (post-outbreak)

5.2.4 Correlation Analysis

The complete descriptions of variables could be found in Appendix A. Note that some variables are only valid for post-outbreak dataset, such as the ones relating to the pandemic and the newly instated Downtown Zone taxing.

Correlation Matrices between Continuous Variables

Figure 5.13 plots the strength of relationship between each continuous variables in terms of Pearson's coefficient. Red indicates negative, while blue indicates positive correlation. For better visibility, these matrices are also presented in the Appendix D.



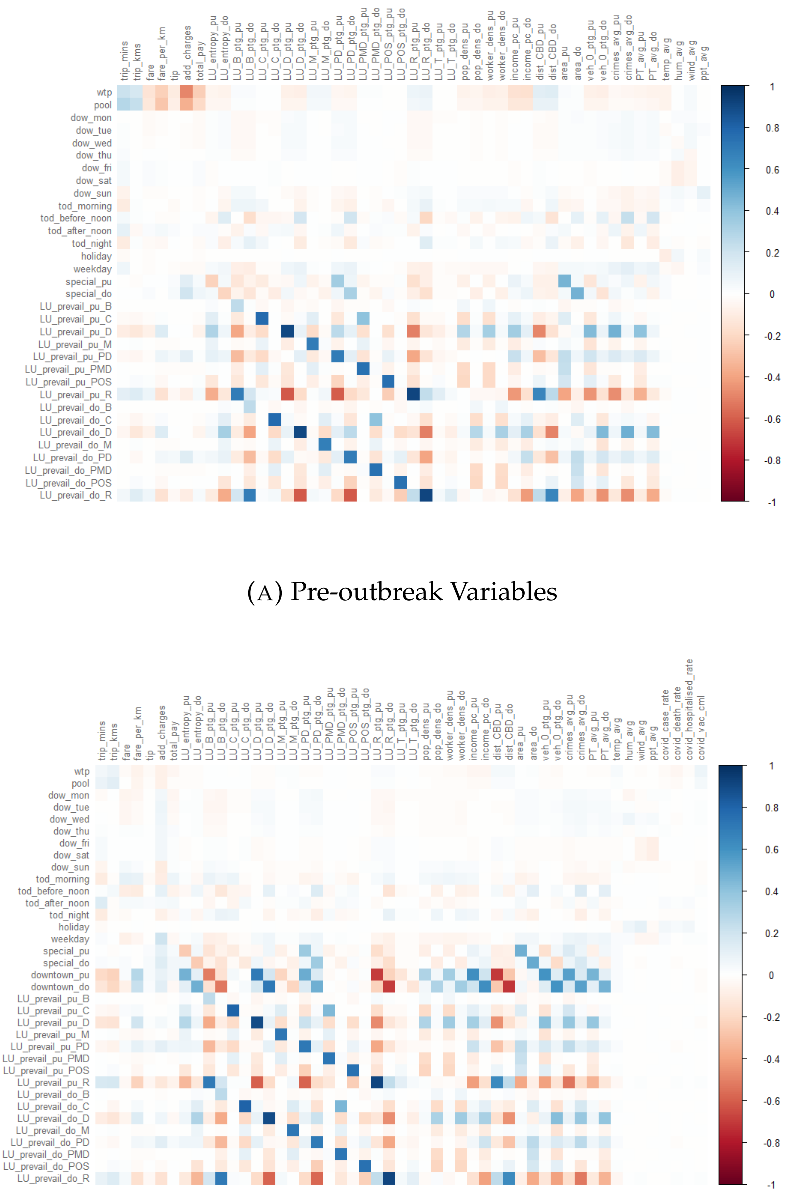
(A) Pre-outbreak Variables

(B) Post-outbreak Variables

FIGURE 5.13: Pearson's coefficient between continuous variables

Correlation Matrices between Continuous and Categorical Variables

Similarly, Figure 5.14 shows the correlation strengths between the continuous and categorical variables. Attention was especially given to variables that have stronger correlations to the dependent variables *wtp* and *pool*. For better visibility, these matrices are also presented in the Appendix D.



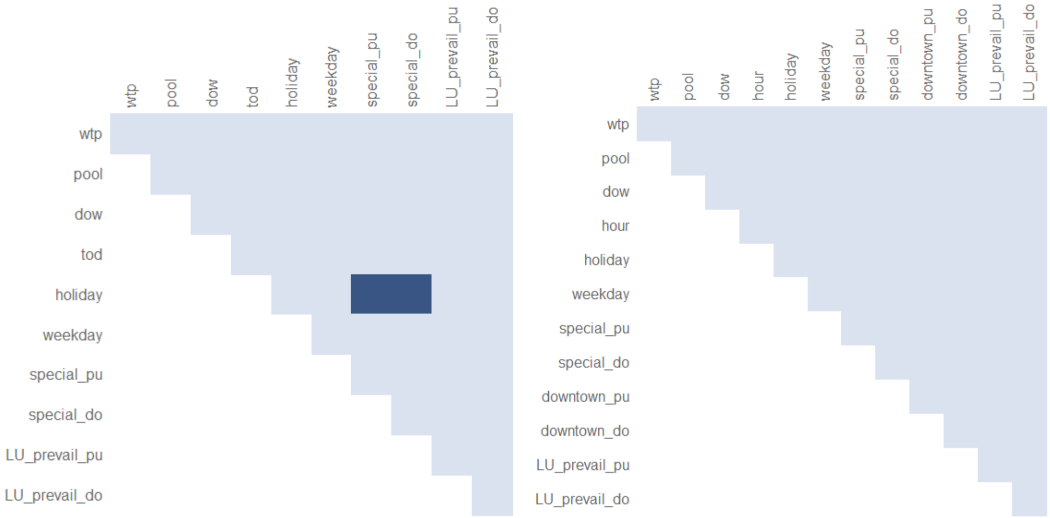
(A) Pre-outbreak Variables

(B) Post-outbreak Variables

FIGURE 5.14: Point biserial (Pearson's) coefficients between continuous and categorical variables

Correlation Matrices between Categorical Variables

Unlike the above matrices, Figure 5.15 plots the dependence between each categorical variables based on the p-value of the Chi-square test. Dark blue indicates independence while light blue the otherwise.



(A) Pre-outbreak Variables (B) Post-outbreak Variables

FIGURE 5.15: Dependence between categorical variables

5.3 Model Building and Assessment

5.3.1 Pre-selection of Variables

Based on the correlation matrices presented above, the candidate model variables were pre-selected as per the method outlined in Section 4.4.3. Table 5.2 lists the remaining variables. Note that in both periods, the variable candidates for 'WTP' and 'Pool' models were identical. At this stage, it was also decided that *LU_entropy_..*[*] and *LU_prevail_..*[*] were better descriptors for each tract's land use compared to the percentages of each land use type due to the collinearity the latter imposed. Similarly, the variable *dow* which indicated the day of the week of the trip was aggregated to the variable *weekday* instead.

[*] referring to both 'pu' and 'du'

TABLE 5.2: Pre-selected candidate variables

Period					
Pre-outbreak			Post-outbreak		
Pre-selected Variables					
add_charges	LU_entropy_do	tip	add_charges	hum_avg	temp_avg
area_do	LU_entropy_pu	tod	area_do	income_pc_do	tip
area_pu	LU_prevail_do	trip_kms	area_pu	income_pc_pu	tod
dist_CBD_do	LU_prevail_pu	user_age_ptg_pu	covid_case_rate	LU_entropy_do	total_pay
dist_CBD_pu	ppt_avg	veh_0_ptg_do	covid_death_rate	LU_entropy_pu	trip_kms
fare_per_km	PT_avg_do	veh_0_ptg_pu	covid_vac_cml	LU_prevail_do	veh_0_ptg_do
holiday	PT_avg_pu	weekday	crimes_avg_do	LU_prevail_pu	veh_0_ptg_pu
hum_avg	special_do	wind_avg	crimes_avg_pu	ppt_avg	weekday
income_pc_do	special_pu	worker_dens_do	dist_CBD_do	PT_avg_do	wind_avg
income_pc_pu	temp_avg	worker_dens_pu	dist_CBD_pu	PT_avg_pu	worker_dens_do
			fare_per_km	special_do	worker_dens_pu
			holiday	special_pu	

5.3.2 Generation of Candidate Models

This section presents the summaries of 8 sets of candidate models in total:

1. Pre-outbreak ’WTP’ models generated through Stepwise selection
2. Pre-outbreak ’WTP’ models generated through Lasso selection
3. Post-outbreak ’WTP’ models generated through Stepwise selection
4. Post-outbreak ’WTP’ models generated through Lasso selection
5. Pre-outbreak ’Pool’ models generated through Stepwise selection
6. Pre-outbreak ’Pool’ models generated through Lasso selection
7. Post-outbreak ’Pool’ models generated through Stepwise selection
8. Post-outbreak ’Pool’ models generated through Lasso selection

Each of the model sets presented below have gone through an iterative process of assessments and amendments such that the signs of the coefficients are stable, no multicollinearity equating to VIF > 5 is present, and have satisfactory fit and predictive performance.

To reiterate, removal of variables from a Stepwise model set was credit to the pre-selection and the Stepwise elimination procedure itself. On the other hand, while pre-selection was not conducted on a Lasso set, removal of variables was due to elimination by Lasso regression (forcing the coefficients to be zero) and manual removal when multicollinearity were found in the resulting model candidates.

Note that for sets generated through Stepwise selection, the variable’s importance was measured by the proportion of significance—the proportions at which the coefficient has high significance, moderate significance, low significance, and non-significance/does not appear in the model at all across the whole N subsets. Meanwhile, for the reason mentioned in Section 2.3.3, for sets generated through Lasso selection, the proportion of appearance—the proportion of times the variable is involved across the whole N subsets–was assessed to measure the variable’s importance.

Candidate 'WTP' Models

In these models, the dependent variable is the binary variable *wtp*, i.e., the willingness to pool. The input data corresponded to all ride-hailing trips conducted in the respective study periods.

- **Pre-outbreak Models**

For the following model sets, *N* equals to 60 subsets. The generated model candidates for pre-outbreak 'WTP' models using Stepwise selection are summarised by Figure 5.16 and 5.17.

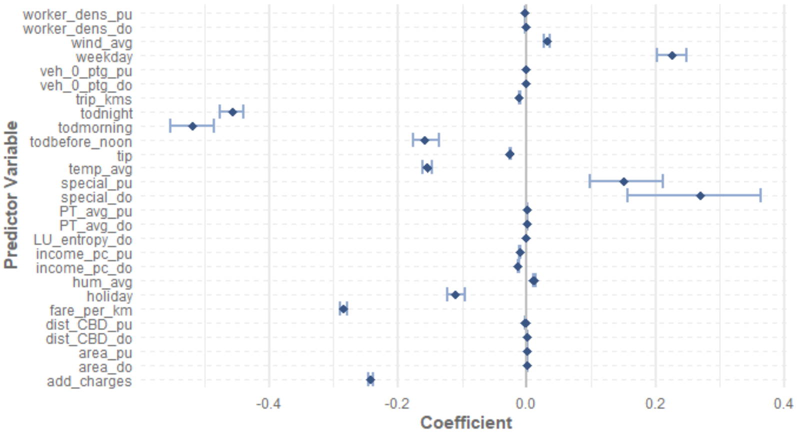


FIGURE 5.16: Variance of coefficients (pre-outbreak 'WTP' with Stepwise)

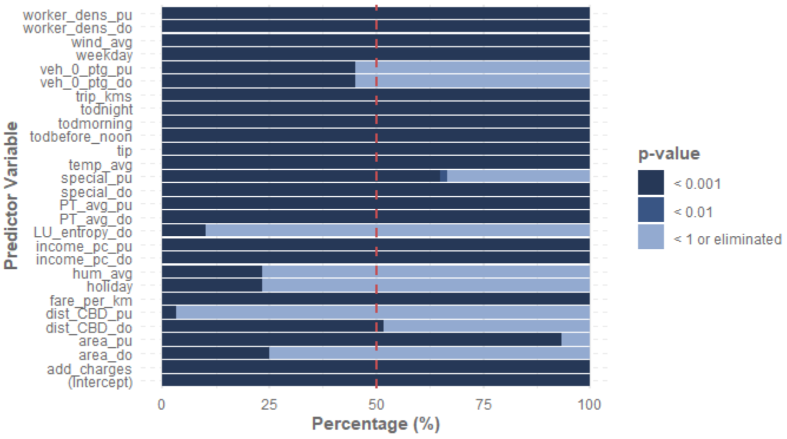


FIGURE 5.17: Proportion of significance (pre-outbreak 'WTP' with Stepwise)

The generated model candidates for pre-outbreak 'WTP' models using Lasso regression are summarised by Figure 5.18 and 5.19.

In the Stepwise model sets, the variables *crimes_avg_..** and *pop_dens_..** were already removed in the pre-selection step, while *LU_entropy_pu* and *ppt_avg* were eliminated by the Stepwise procedure. This was justified, as these variables were

*referring to both 'pu' and 'do'

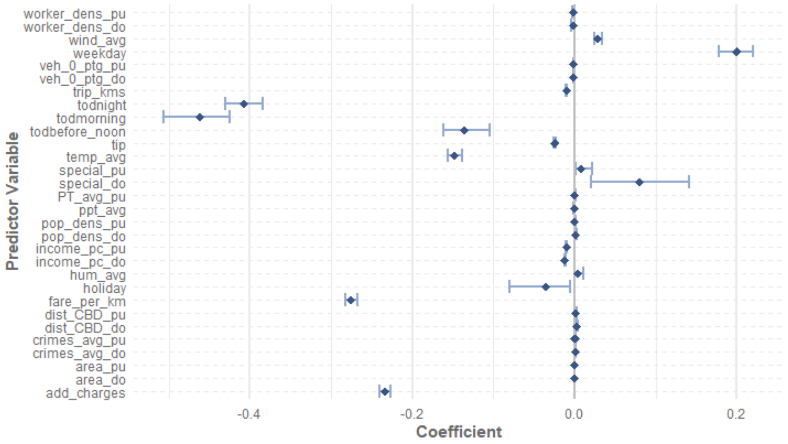


FIGURE 5.18: Variance of coefficients (pre-outbreak 'WTP' with Lasso)

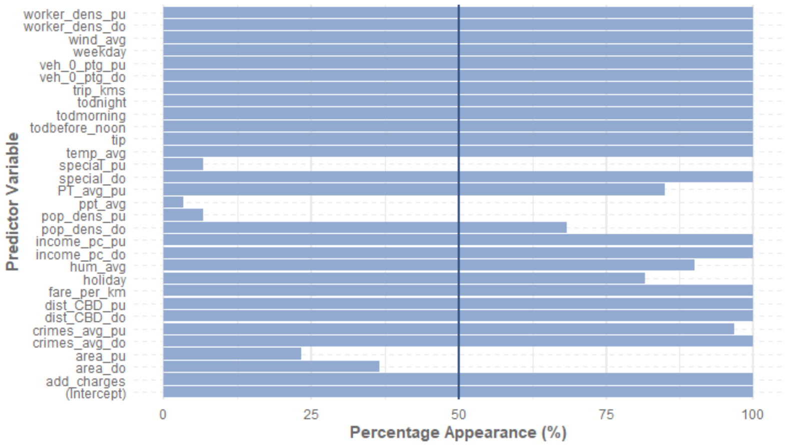


FIGURE 5.19: Proportion of appearance (pre-outbreak 'WTP' with Lasso)

also found to be removed (zero coefficient) or negligible in the Lasso regression result. For the most part, both the Stepwise and Lasso model sets agree with each other. The variables *trip_mins*, *fare*, and *total_pay* were removed from both sets due to multicollinearity, while *LU_prevail_..*[*] were removed due to sign instability.

- **Post-outbreak Models**

For the following two sets of models, the number N equals to 35 subsets. The generated model candidates for post-outbreak 'WTP' models using Stepwise selection are summarised by Figure 5.20 and 5.21.

[*] referring to both 'pu' and 'do'

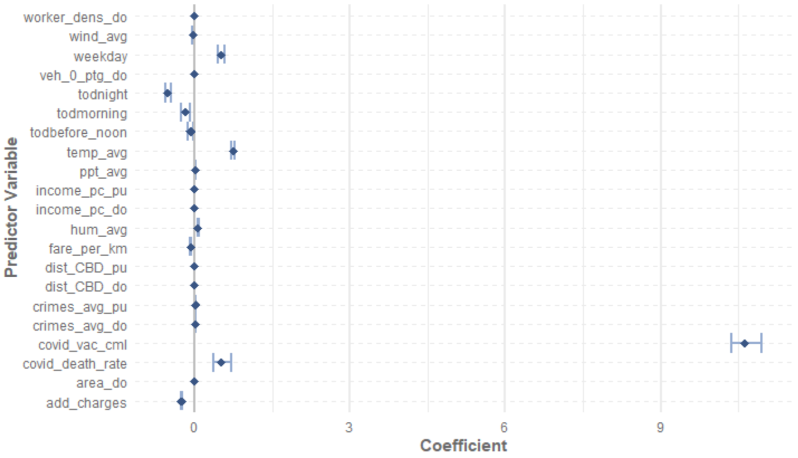


FIGURE 5.20: Variance of coefficients (post-outbreak 'WTP' with Stepwise)

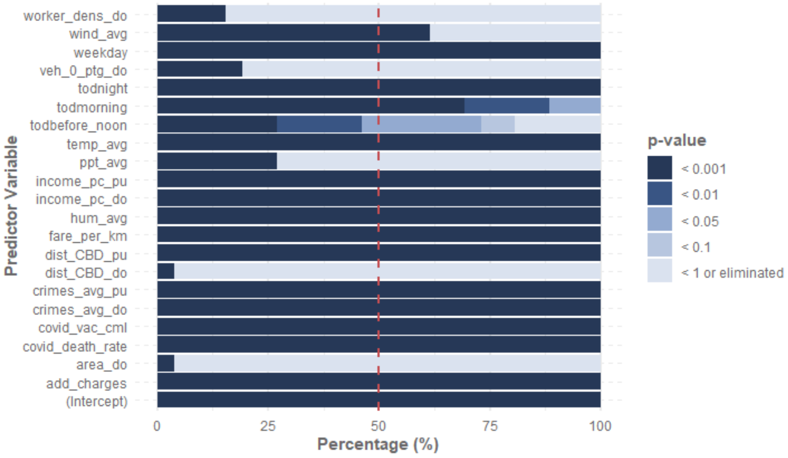


FIGURE 5.21: Proportion of significance (post-outbreak 'WTP' with Stepwise)

The generated model candidates for post-outbreak 'WTP' models using Lasso regression are summarised by Figure 5.22 and 5.23.

As before, most of the variables that were eliminated in the Stepwise procedure also ended up eliminated in the Lasso procedure due to coefficient shrinkage to zero/negligible or removed due to multicollinearity. Exceptions were for the variables *downtown_..*[*] which were pre-eliminated in the Stepwise procedure, as well as the variable *holiday*.

[*] referring to both 'pu' and 'du'

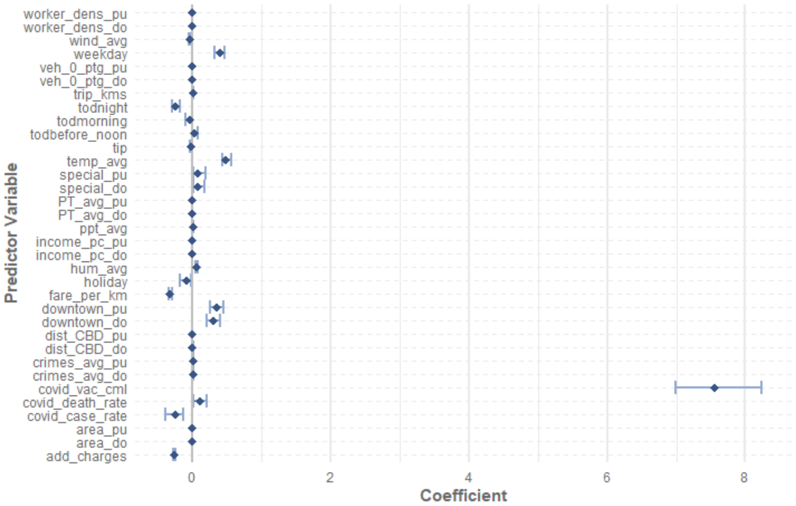


FIGURE 5.22: Variance of coefficients (post-outbreak 'WTP' with Lasso)

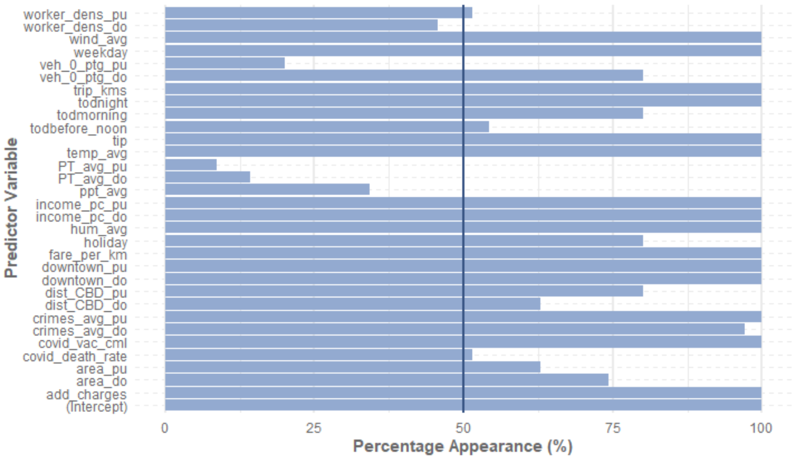


FIGURE 5.23: Proportion of appearance (post-outbreak 'WTP' with Lasso)

Candidate 'Pool' Models

In these models, the dependent variable is the binary variable *pool*, i.e., the success of a pooling request. The input data corresponded to all ride-hailing trips that were authorised for pooling conducted in the respective study periods.

• Pre-outbreak Models

The generated candidate model for pre-outbreak 'Pool' models using Stepwise selection are summarised by Figure 5.24 and 5.25. The number of subsets N is equal to 10 for these models.

The generated model candidates for pre-outbreak 'Pool' models using Lasso regression are summarised by Figure 5.26 and 5.27.

In the case of pre-outbreak 'Pool' model sets, all of the variables that are not present in the Stepwise set as a result of pre-selection, Stepwise elimination, or removal due

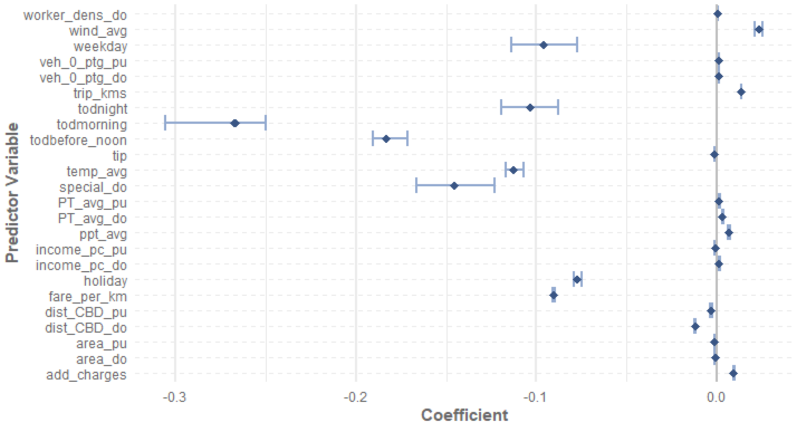


FIGURE 5.24: Variance of coefficients (pre-outbreak 'Pool' with Stepwise)

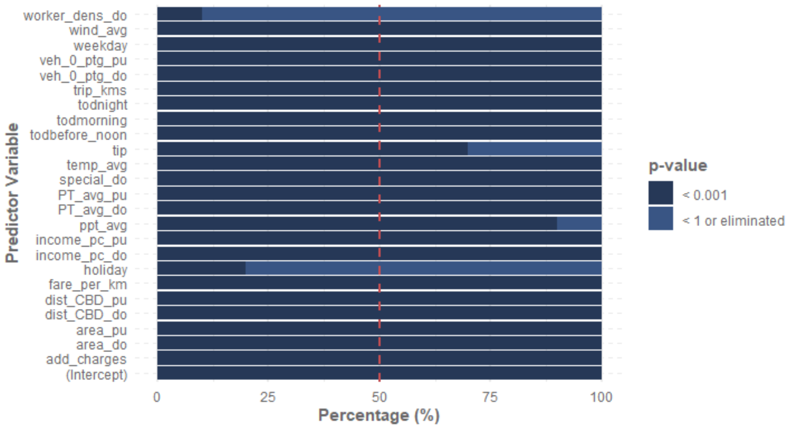


FIGURE 5.25: Proportion of significance (pre-outbreak 'Pool' with Stepwise)

to multicollinearity/instability were found to also be eliminated or had low appearance in the Lasso set.

The variables *pop_dens_..*[*], *trip_mins*, *fare*, and *total_pay* were removed in both sets due to multicollinearity. The variables *LU_prevail_..*[*] were removed from both sets due to instability.

[*] referring to both 'pu' and 'du'

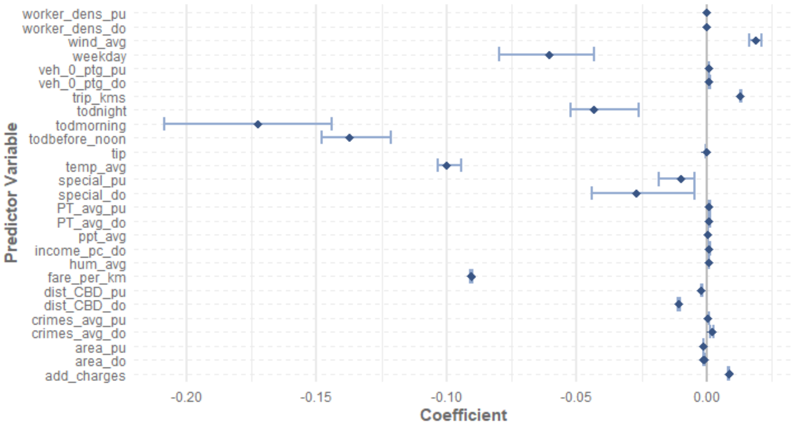


FIGURE 5.26: Variance of coefficients (pre-outbreak 'Pool' with Lasso)

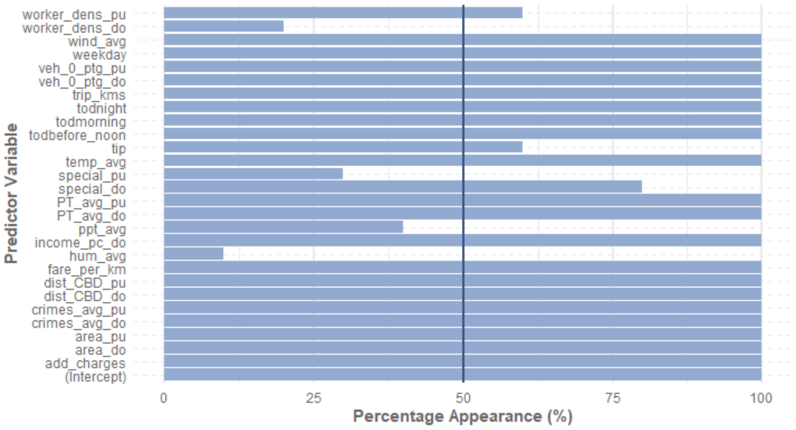


FIGURE 5.27: Proportion of appearance (pre-outbreak 'Pool' with Lasso)

Post-outbreak Models

The generated model candidates for post-outbreak 'Pool' models using Stepwise selection are summarised by Figure 5.28 and 5.29. Note that due to mach smaller dataset of 'WTP' trips post-outbreak (less than one million trips), the subset size N for the post-outbreak 'Pool' models is equal to 1.

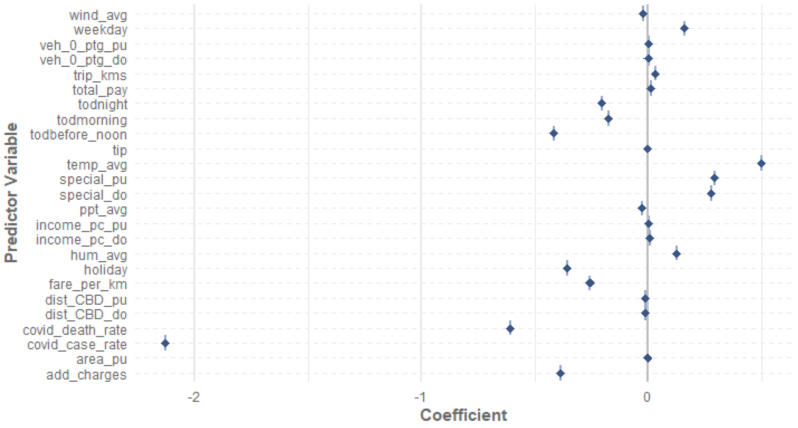


FIGURE 5.28: Variance of coefficients (post-outbreak 'Pool' with Stepwise)

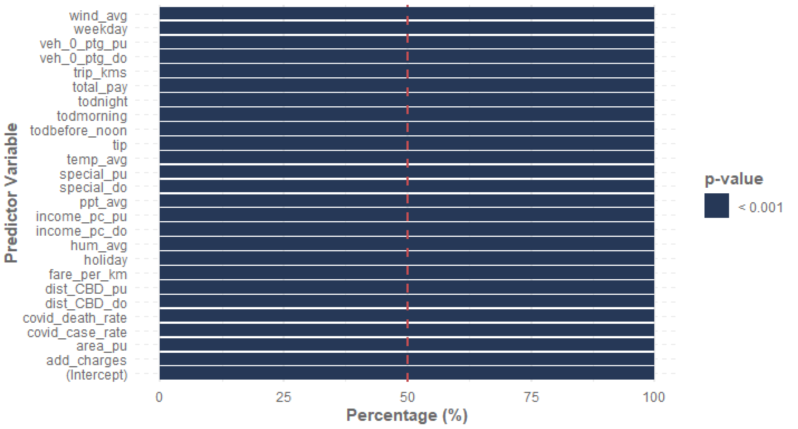


FIGURE 5.29: Proportion of significance (post-outbreak 'Pool' with Stepwise)

The generated model candidates for post-outbreak 'Pool' models using Lasso regression are summarised by Figure 5.30 and 5.31.

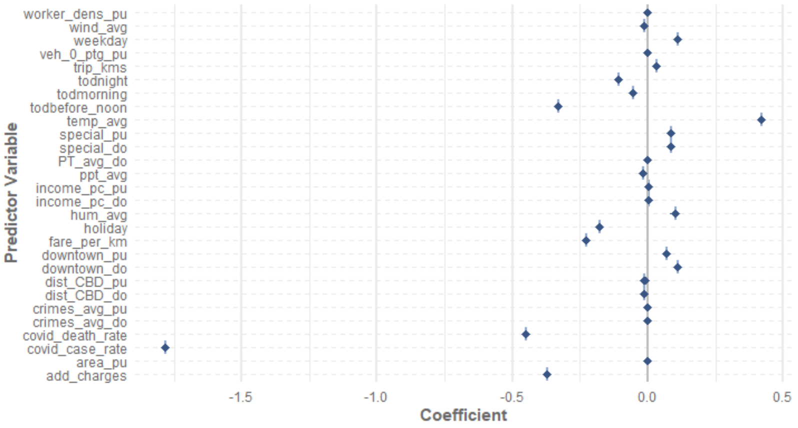


FIGURE 5.30: Variance of coefficients (post-outbreak 'Pool' with Lasso)

Again, the removal of variables from one model set were mostly justified as the same variables were also eliminated on the other set due to Stepwise or Lasso elimination.

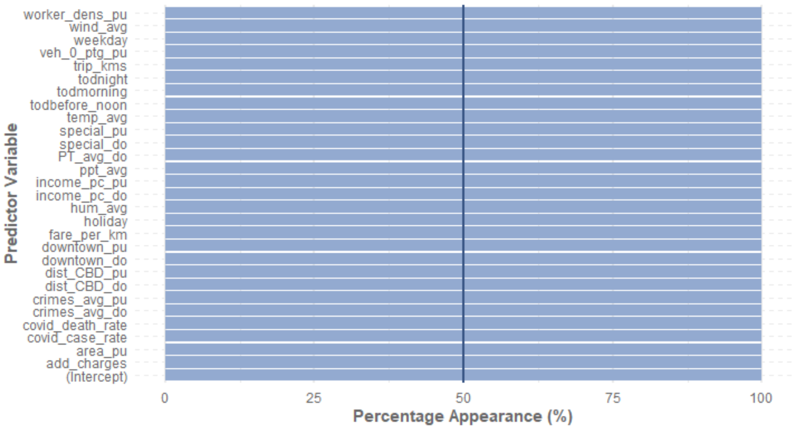


FIGURE 5.31: Proportion of appearance (post-outbreak 'Pool' with Lasso)

Except for the variables *downtown_..*[*] which were pre-eliminated from the Stepwise set due to high correlation with other variables. Unfortunately, the importance of these variables could not be assessed with the post-outbreak 'Pool' model only having one input data subset.

5.3.3 Final Models Building and Assessment

Based on the results of the previous section, final candidate variables were selected for each model. The requirements include non-negligible stable coefficients and more than 50% proportion of high significance (p-value < 0.01) or more than 50% appearance rate. Subsequently, the following models were re-fitted to the binomial logistic function (without penalty) using the respective dataset. Note that the performance measures listed are the minimum across the whole dataset, while the VIF values are the maximum.

Final 'WTP' Models

• Pre-outbreak Model (Stepwise Selection)

The model formula for pre-outbreak 'WTP' model obtained through Stepwise procedure is as follows:

$$wtp \sim add\_charges + fare\_per\_km + income\_pc\_do + income\_pc\_pu + special\_do + \\ special\_pu + temp\_avg + tip + tod + trip\_kms + weekday + wind\_avg$$

Figure 5.32, Figure 5.33, Table 5.3, and Table 5.4 show the assessment results of the above logistic model across the 60 subsets of pre-outbreak ride-hailing trips.

This model suggests that:

- The odds of riders willing to pool is decreased with increasing additional charge, fare rate, income level of pickup and dropoff tracts, temperature, tip, and trip distance

[*] referring to both 'pu' and 'du'

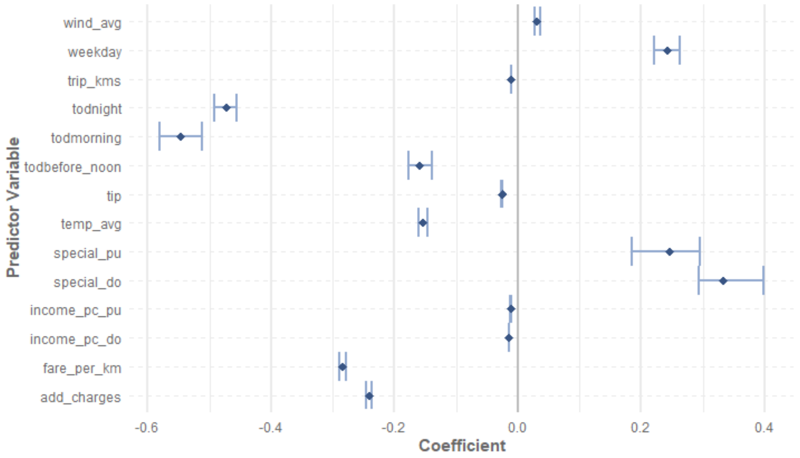


FIGURE 5.32: Variance of coefficients of the final pre-outbreak 'WTP' model (Stepwise)

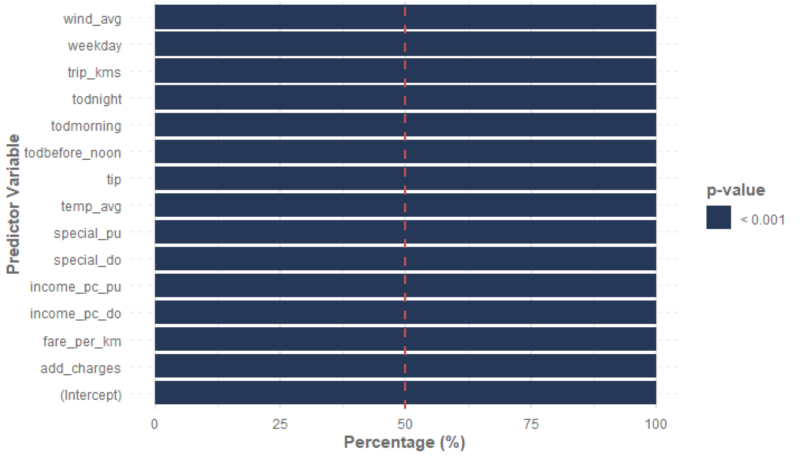


FIGURE 5.33: Proportion of significance in the final pre-outbreak 'WTP' model (Stepwise)

TABLE 5.3: Performance and goodness-of-fit of the final pre-outbreak 'WTP' (Stepwise) model

Model	Accuracy	Pseudo-R^2		AUC ROC
		McFadden	McFadden (Adj)	
Pre-outbreak 'WTP' (Stepwise)	0.92996	0.51162	0.51158	0.93306

- The odds of riders willing to pool is increased with each increase in average wind speed of that day
- Having the pickup or dropoff location within the Special Zones increases the odds of WTP compared to when the pickup or dropoff location is within a regular area
- Compared to the afternoon (1PM - 7PM), having the trip in the morning (12AM - 5AM), before noon (6AM - 12PM), or night (8PM - 11PM) decreases the odd of riders willing to pool
- The trip occurring on the weekday has higher odds of riders willing to pool in comparison to the weekend

TABLE 5.4: Multicollinearity of the variables in the final pre-outbreak 'WTP' (Stepwise) model

VIF/Square of $(GVIF)^{(1/(2*Df))}$					
add_charges	**fare_per_km**	**income_pc_pu**	**income_pc_do**	**special_do**	**special_pu**
1.5482	1.9827	1.1031	1.0926	1.0177	1.0182
temp_avg	**tip**	**tod**	**trip_kms**	**weekday**	**wind_avg**
1.0488	1.0050	1.0327	1.4036	1.0572	1.0461

- The highest effect on the odds of WTP is imposed by the trip occurring in the morning, followed by the trip occurring in the night, and the trip destination being within the Special Zones
- All of the above effects are significant

This model also has excellent predictive performance, indicated by the high out-of-sample prediction accuracy and AUC of ROC. The fit is also excellent with both McFadden and Adjusted McFadden pseudo-R^2 being higher than 0.2. There is also a low level of multicollinearity among the variables.

• Pre-outbreak Model (Lasso Selection)

The following is the model formula for pre-outbreak 'WTP' obtained through Lasso procedure:

$$wtp \sim add\_charges + fare\_per\_km + holiday + income\_pc\_do + income\_pc\_pu + \\ special\_do + temp\_avg + tip + tod + trip\_kms + weekday + wind\_avg$$

The logistic model above was assessed against the whole pre-outbreak ride-hailing dataset (N = 60 subsets) and yielded the results shown in Figure 5.34, Figure 5.35, Table 5.5, and Table 5.6.

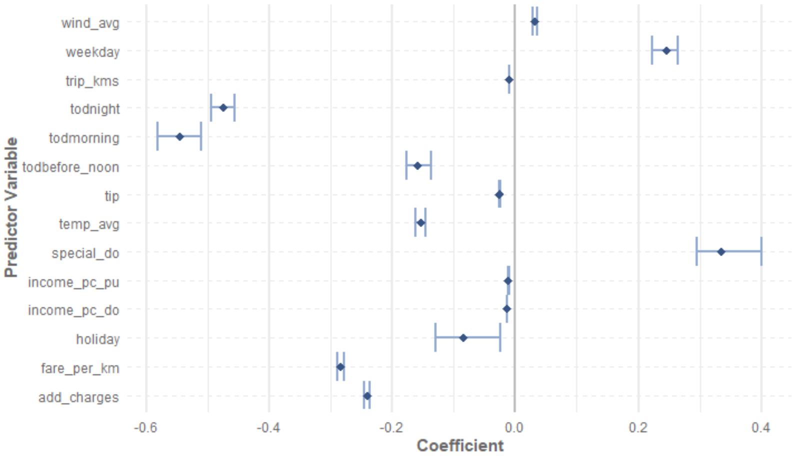


FIGURE 5.34: Variance of coefficients of the final pre-outbreak 'WTP' model (Lasso)

This model suggests that:

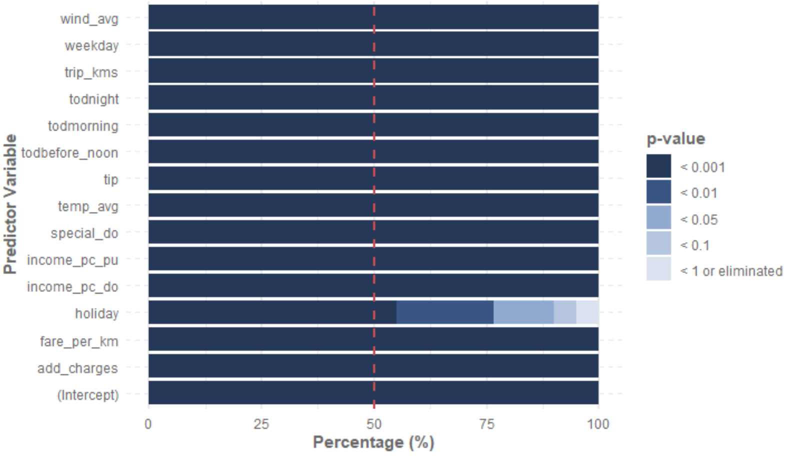


FIGURE 5.35: Proportion of significance of the final pre-outbreak 'WTP' model (Lasso)

TABLE 5.5: Performance and goodness-of-fit of the final pre-outbreak 'WTP' (Lasso) model

Model	Accuracy	Pseudo-R^2		AUC ROC
		McFadden	McFadden (Adj)	
Pre-outbrak 'WTP' (Lasso)	0.92994	0.51154	0.51150	0.93305

- The odds of riders willing to pool decreases with increasing additional charges, fare rate, income level of the pickup or dropoff tracts, average temperature of the day, tip, and trip distance
- The odds of riders willing to pool increases with increasing average wind speed of the day
- In comparison to non-holiday dates, the odds of riders willing to pool decreases on holiday dates. This effect is more than 80% of the times statistically significant under 5% margin of error
- The odds of riders willing to pool is higher when the destination is within the Special Zones compared to when the destination is within a non-Special Zone
- In comparison to the afternoon, having the trip in the morning, before noon, or night decreases the odds of riders willing to pool
- Compared to weekend, the odds of riders willing to pool is increased when the trip is on a weekday

TABLE 5.6: Multicollinearity of the variables in the final pre-outbreak 'WTP' (Lasso) model

VIF/Square of $(GVIF)^{(1/(2*Df))}$					
add_charges	**fare_per_km**	**holiday**	**income_pc_pu**	**income_pc_do**	**special_do**
1.5481	1.9829	1.0298	1.0862	1.0924	1.0177
temp_avg	**tip**	**tod**	**trip_kms**	**weekday**	**wind_avg**
1.0601	1.0050	1.0338	1.4034	1.0735	1.0461

- The highest effect on the odds of WTP is imposed by the trip occurring in the morning, followed by the trip occurring in the night, and the trip destination being within the Special Zones
- Except for the effect of public holidays, all of the above effects are highly significant

This model has excellent predictive performance and fit with the high out-of-sample accuracy and AUC of ROC, as well as both pseudo-R^2s passing 0.2. The multicollinearity level between the variables is also low.

• Post-outbreak Model (Stepwise Selection)

The model formula for post-outbreak 'WTP' model obtained through Stepwise procedure is as follows:

$$wtp \sim add\_charges + covid\_death\_rate + covid\_vac\_cml + fare\_per\_km + \\ hum\_avg + temp\_avg + tod + weekday + wind\_avg$$

Figure 5.36, Figure 5.37, Table 5.7, and Table 5.8 show the assessment results of the above logistic model.

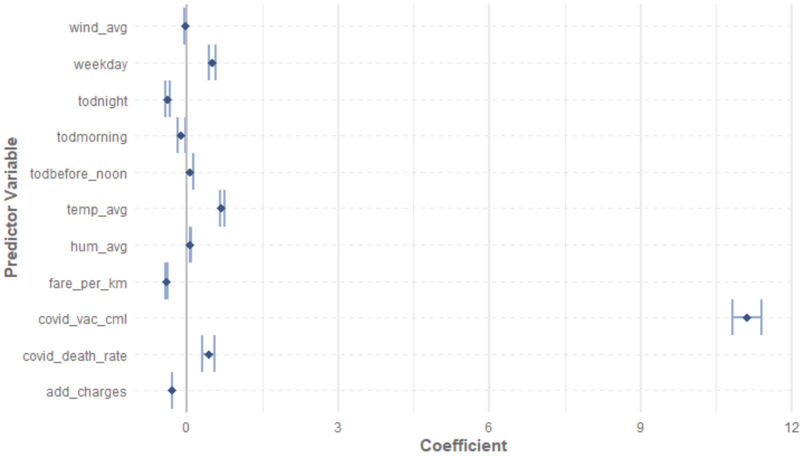


FIGURE 5.36: Variance of coefficients of the final post-outbreak 'WTP' model (Stepwise)

TABLE 5.7: Performance and goodness-of-fit of the final post-outbreak 'WTP' (Stepwise) model

Model	Accuracy	Pseudo-R^2		AUC ROC
		McFadden	McFadden (Adj)	
Post-outbreak 'WTP' (Stepwise)	0.99119	0.35505	0.35479	0.93554

This model suggests that:

- The increase in additional charges, fare rate, and average wind speed of the day decreases the odds of riders willing to pool

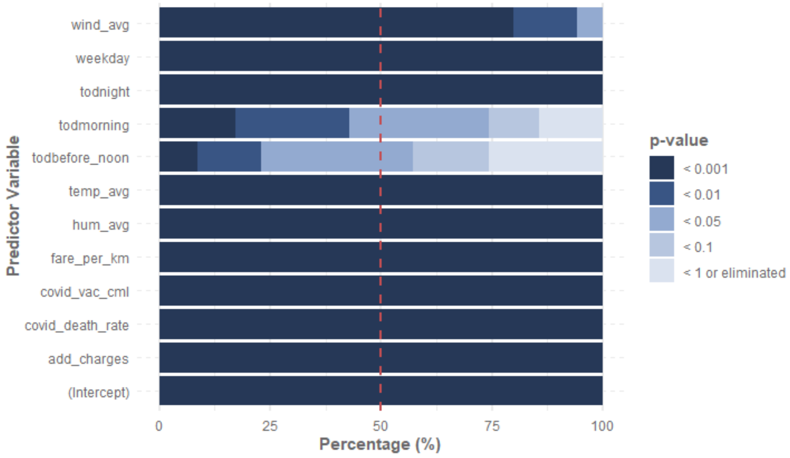


FIGURE 5.37: Proportion of significance of the final post-outbreak 'WTP' model (Stepwise)

TABLE 5.8: Multicollinearity of the variables in the final post-outbreak 'WTP' (Stepwise) model

VIF/Square of $(GVIF)^{(1/(2*Df))}$				
add_charges	**covid_death_rate**	**covid_vac_cml**	**fare_per_km**	**hum_avg**
1.0440	1.3056	1.8251	1.0239	1.0971
temp_avg	**tod**	**weekday**	**wind_avg**	
1.9294	1.0203	1.0805	1.2169	

- The increase in the COVID-19 death rate or cumulative vaccination increases the odds of WTP
- The increase in average humidity or average temperature of the day increases the odds of WTP
- In comparison to the afternoon, trip conducted before noon increases the odds of riders willing to pool. This effect, however, has a lower significance proportion. It is significant under the threshold of p-value < 0.05 only lightly above 50% across the whole data subsets
- In comparison to the afternoon, trip conducted in the morning or night decreases the WTP odds, although the former has lower significance proportion
- In comparison to weekend, the trip conducted on a weekday yields higher odds of WTP
- COVID-19 cumulative vaccination has a dominating effect compared to other variables
- Most of the effects above are statistically highly significant

The predictive performance and fit of this model is excellent, while the multicollinearity level among the variables are generally low as it stays below 5.

- **Post-outbreak Model (Lasso Selection)**

The following is the model formula for post-outbreak 'WTP' obtained through Lasso procedure:

$$\begin{aligned} wtp \sim & add\_charges + covid\_death\_rate + covid\_vac\_cml + downtown\_pu + \\ & downtown\_do + fare\_per\_km + holiday + hum\_avg + \\ & temp\_avg + tod + weekday \end{aligned}$$

The logistic model above was assessed against the whole post-outbreak ride-hailing dataset and yielded the results shown in Figure 5.38, Figure 5.39, Table 5.9, and Table 5.10.

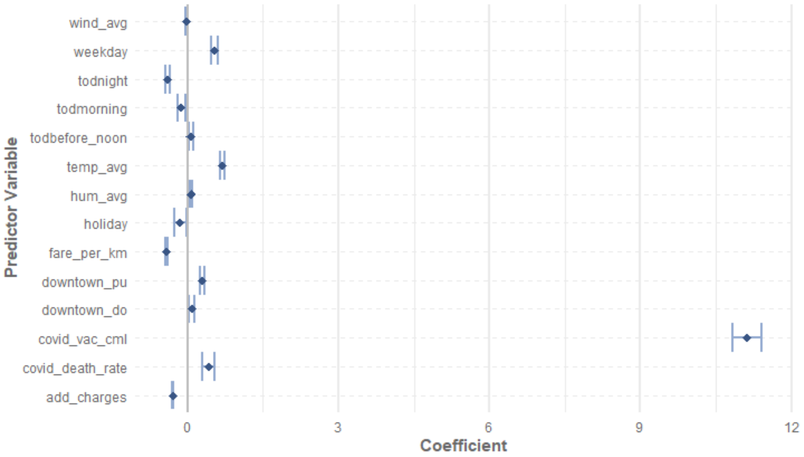


FIGURE 5.38: Variance of coefficients of the final post-outbreak 'WTP' model (Lasso)

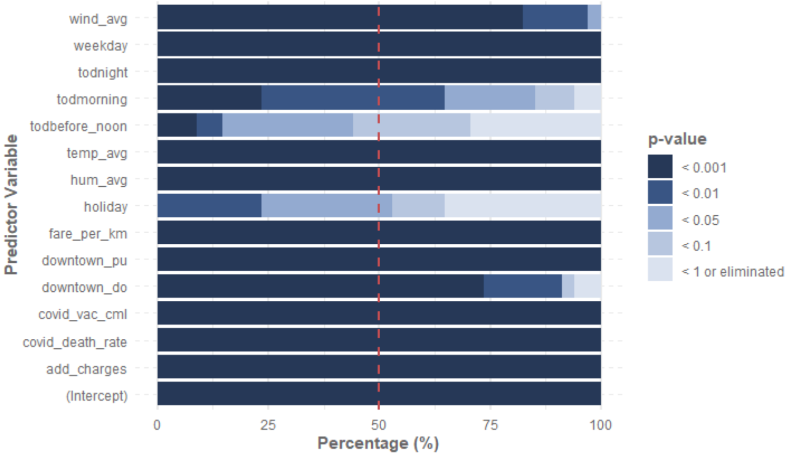


FIGURE 5.39: Proportion of significance of the final post-outbreak 'WTP' model (Lasso)

The above results suggest that:

- The odds of WTP decreases with the increase of additional charges, fare rate, or average wind speed

TABLE 5.9: Performance and goodness-of-fit of the final post-outbreak 'WTP' (Lasso) model

Model	**Accuracy**	**Pseudo-R^2**		**AUC ROC**
		McFadden	**McFadden (Adj)**	
Post-outbreak WTP (Lasso)	0.99124	0.35647	0.35614	0.93585

TABLE 5.10: Multicollinearity of the variables in the final post-outbreak 'WTP' (Lasso) model

VIF/Square of $(GVIF)^{(1/(2*Df))}$					
add_charges	**covid_death_rate**	**covid_vac_cml**	**downtown_do**	**downtown_pu**	**fare_per_km**
1.0438	1.3204	1.8260	1.0348	1.0685	1.0977
holiday	**hum_avg**	**temp_avg**	**tod**	**weekday**	**wind_avg**
1.0394	1.0975	1.9297	1.0252	1.0959	1.2189

- The odds of WTP increases with the increase in COVID-19 death rate or cumulative vaccination
- The odds of riders willing to pool also increases should the average humidity or the temperature of the day rises
- When the trip begins or ends within the Downtown Zone, the odds of WTP is higher compared to when it begins or ends outside of Downtown Zone
- In comparison to non-holiday dates, a trip conducted during a holiday decreases the odds of WTP. This effect, however, has lower significance proportion where it is significant under the threshold of p-value < 0.05 only slightly more than 50% across the data subsets
- In contrast to the afternoon, a trip conducted before noon increases the odds of WTP. However, this effect has low significance proportion
- In contrast to the afternoon, a trip conducted in the morning or night decreases the odds of WTP
- In contrast to the weekend, the odds of riders willing to pool increases should the trip is conducted on a weekday
- COVID-19 cumulative vaccination has a dominating effect compared to other variables
- Most of the effects above are statistically highly significant.

As before, this model yields excellent fit and predictive performance as indicated by pseudo-R^2 higher than 0.2 and high out-of-sample prediction accuracy and AUC of ROC. The multicollinearity between the variables is also within the threshold of VIF < 5.

Final 'Pool' Models

- **Pre-outbreak Models**

The final variables for pre-outbreak 'Pool' model obtained through Stepwise and Lasso were identical. The model formula is as follows:

$$\begin{aligned} pool \sim & add\_charges + dist\_CBD\_do + fare\_per\_km + special\_do + temp\_avg + \\ & tod + trip\_kms + weekday + wind\_avg \end{aligned}$$

This logistic model was tested against the WTP trip dataset of the pre-outbreak study period. The results are presented in Figure 5.40, Figure 5.41, Table 5.11, and Table 5.12.

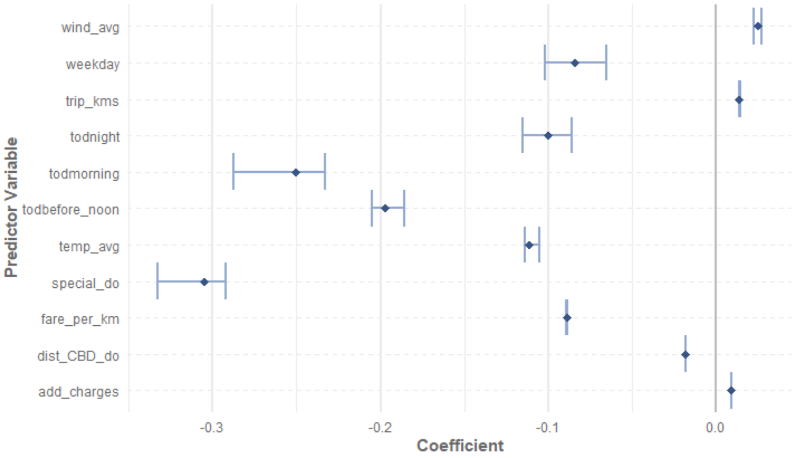


FIGURE 5.40: Variance of coefficients of the final pre-outbreak 'Pool' model

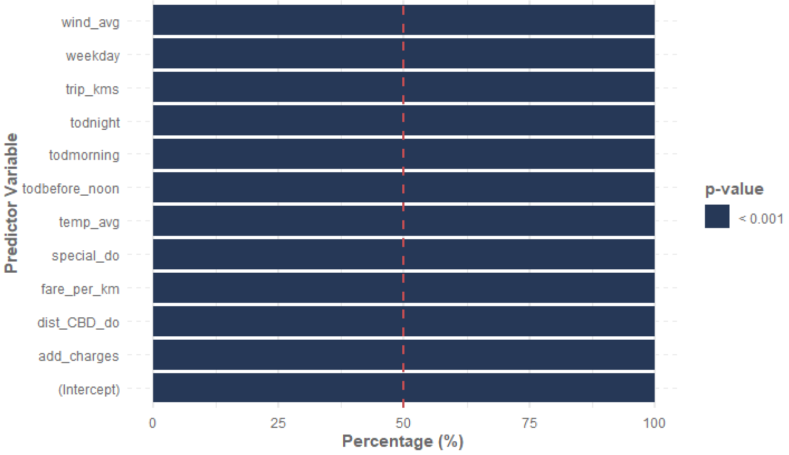


FIGURE 5.41: Proportion of significance of the final pre-outbreak 'Pool' model

This above results suggest that:

- The odds of a pooling request successfully matched increases with each increase in additional charges, trip distance, and the average wind of the day

TABLE 5.11: Performance and goodness-of-fit of the final pre-outbreak 'Pool' model

Model	Accuracy	Pseudo-R^2		AUC ROC
		McFadden	McFadden (Adj)	
Pre-outbrak 'Pool'	0.78475	0.19009	0.19006	0.79466

TABLE 5.12: Multicollinearity of the variables in the final pre-outbreak 'Pool' model

VIF/Square of $(GVIF)^{(1/(2*Df))}$				
add_charges	**dist_CBD_do**	**fare_per_km**	**special_do**	**temp_avg**
1.2004	1.1507	1.5012	1.0164	1.0595
tod	**trip_kms**	**weekday**	**wind_avg**	
1.0327	1.4353	1.0753	1.0532	

- The odds of pooling success decreases should the destination's distance to the CBD increases
- The odds of pooling success decreases with the rise in fare rate and average temperature of the day
- Having the destination within the Special Zones decreases the odds of pooling success compared to when the destination is outside the Special Zones
- Compared to the afternoon, trip with pooling request before noon, in the morning, or in the night decreases the odds of pooling/matching success
- Relative to the weekend, a pooling request on a weekday decreases the odds of pooling success
- Among the above effects, the destination within the Special Zones has the highest magnitude towards the odds of pooling success. It is followed by morning time of day and before noon
- All of these effects are statistically highly significant

Although lower than other models, the pre-outbreak 'Pool' model has a satisfactory fit with the pseudo-R^2 measures fall just under 0.2. The out-of-sample accuracy and AUC of ROC are also acceptable as they stay above 0.7.

• Post-outbreak Model (Stepwise Selection)

Meanwhile, the final post-outbreak 'Pool' model obtained through Stepwise selection is as follows:

$$\begin{aligned} pool \sim & add\_charges + covid\_case\_rate + covid\_death\_rate + fare\_per\_km + \\ & holiday + hum\_avg + special\_do + special\_pu + temp\_avg + tod + \\ & trip\_kms + weekday \end{aligned}$$

After the model above was the re-fitted with the post-outbreak WTP trip dataset, the results in Figure 5.42, Figure 5.43, Table 5.13, and Table 5.14 were obtained.

This model suggests that:

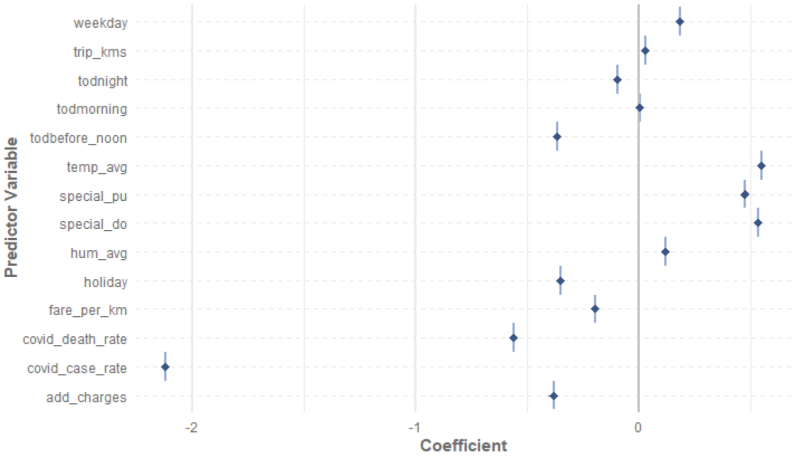


FIGURE 5.42: Variance of coefficients of the final post-outbreak 'Pool' (Stepwise) model

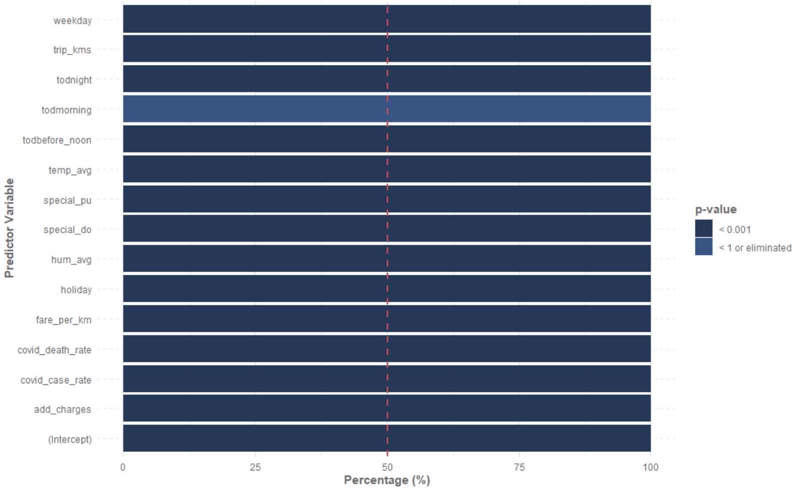


FIGURE 5.43: Proportion of significance of the final post-outbreak 'Pool' (Stepwise) model

- The odds of pooling success decreases with the increase of additional charges, COVID-19 case rate, COVID-19 death rate, and fare rate
- The odds of pooling success increases with the increase of trip distance, average humidity, and average temperature of the day
- In contrast to non-holiday dates, the odds of pooling success decreases during public holidays
- Having the pickup or dropoff locations within the Special Zones increases the odds of pooling success compared to when the pickup or dropoff locations are outside the Special Zones

TABLE 5.13: Performance and goodness-of-fit of the final post-outbreak 'Pool' (Stepwise) model

Model	Accuracy	Pseudo-R^2		AUC ROC
		McFadden	McFadden (Adj)	
Post-outbrak 'Pool' (Stepwise)	0.86683	0.51565	0.51550	0.93491

TABLE 5.14: Multicollinearity of the variables in the final post-outbreak ’Pool’ (Stepwise) model

VIF/Square of $(GVIF)^{(1/(2*Df))}$					
add_charges	**covid_case_rate**	**covid_death_rate**	**fare_per_km**	**holiday**	**hum_avg**
1.2698	1.7628	1.1268	1.5888	1.0428	1.0971
special_pu	**special_do**	**temp_avg**	**tod**	**trip_kms**	**weekday**
1.0040	1.0034	1.8692	1.0388	1.6582	1.1272

- Compared to the afternoon, requesting for pooling before noon and in the night decreases the odds of pooling success. Meanwhile, the effect that morning time of day has on the odds of pooling success is negligible and statistically insignificant
- Compared to the weekend, requesting for pooling on a weekday increases the odds of pooling success
- COVID-19 case rate has the highest magnitude of effect, followed by the COVID-19 death rate
- Aside from the effect of morning time of day, the above effects are statistically highly significant

This logistic model has an excellent fit and predictive performance. The multicollinearity among the predictors is also low.

• Post-outbreak Model (Lasso Selection)

The Lasso technique yielded a similar model formula shown below:

$$\begin{aligned} pool \sim & add\_charges + covid\_case\_rate + covid\_death\_rate + downtown\_pu + \\ & downtown\_do + fare\_per\_km + holiday + hum\_avg + special\_do + \\ & special\_pu + temp\_avg + tod + trip\_kms + weekday \end{aligned}$$

The re-fitting results of this logistic model are shown in Figure 5.44, Figure 5.45, Table 5.15 and Table 5.16.

TABLE 5.15: Performance and goodness-of-fit of the final post-outbreak ’Pool’ (Lasso) model

Model	**Accuracy**	**Pseudo-R^2**		**AUC ROC**
		McFadden	**McFadden (Adj)**	
Post-outbreak ’Pool’ (Lasso)	0.86934	0.52025	0.52009	0.93618

TABLE 5.16: Multicollinearity of the variables in the final post-outbreak ’Pool’ (Lasso) model

VIF/Square of $(GVIF)^{(1/(2*Df))}$						
add_charges	**covid_case_rate**	**covid_death_rate**	**downtown_pu**	**downtown_do**	**fare_per_km**	**holiday**
1.2755	1.7626	1.1260	1.1227	1.0994	1.7046	1.0434
hum_avg	**special_pu**	**special_do**	**temp_avg**	**tod**	**trip_kms**	**weekday**
1.0974	1.0070	1.0081	1.8617	1.0491	1.6996	1.1270

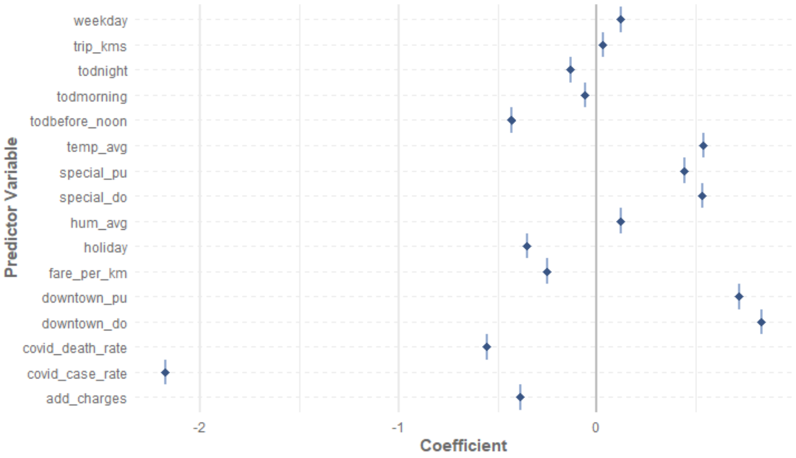


FIGURE 5.44: Variance of coefficients of the final post-outbreak 'Pool' (Lasso) model

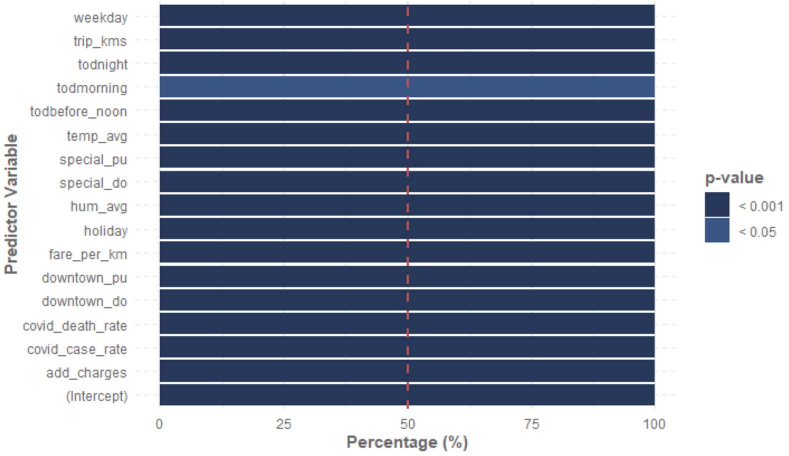


FIGURE 5.45: Proportion of significance of the final post-outbreak 'Pool' (Lasso) model

The results above suggest that:

- The odds of pooling success decreases with the increase in additional charges, COVID-19 case rate, COVID-19 death rate, and fare rate
- The odds of pooling success increases with the increase in trip distance, average humidity, and average temperature of the day
- Having the pickup or dropoff location within the Downtown Zone increases the odds of pooling success in comparison to if the pickup or dropoff location is outside the Downtown Zone
- Requesting for pooling during public holidays decreases the odds of pooling success in comparison to non-holiday dates
- Having the pickup or dropoff location within the Special Zones increases the odds of pooling success in comparison to if the pickup or dropoff location is outside the Special Zones
- In comparison to the afternoon, requesting for pooling before noon (6AM - 12PM), in the morning (12AM - 5AM), or in the night (8PM - 11PM) decreases

the odds of pooling success

- Compared to the weekend, requesting for pooling on a weekday increases the odds of pooling success
- Among the above effects on pooling success, COVID-19 case rate yields the biggest magnitude followed by when the trip destination is within the Downtown Zone
- All of the above effects are statistically significant.

This model yields excellent fit and predictive performance, while the multicollinearity between the predictors is also low.

Chapter 6

Discussion

Following the presentation of final models in the previous chapter, this chapter discusses the main factors affecting willingness to pool and pooling success in the two study periods, lays a comparison between non-pandemic and pandemic periods, and reviews the variable selection techniques employed in this thesis.

6.1 Influencing Factors

6.1.1 Factors Affecting Willingness to Pool

From the final 'WTP' models, it could be observed how some effects are repeated in both Stepwise-selected and Lasso-selected models. These include the effects of additional charges, fare rate, weekday/weekend, and morning and before noon time of day compared to the afternoon. Albeit at lower magnitudes, these effects also persist in the post-outbreak models, indicating that these factors may have high relevance to willingness to pool.

Outside of pandemic context, whether or not the the pickup or dropoff location are within the Special Zones seems to possibly play a role in riders' willingness to pool. However, during a pandemic, these factors may no longer be relevant. The same goes to income levels of the pickup and dropoff tracts, tip, and trip distance.

Moreover, in the pre-pandemic models, average temperature has negative coefficients, while average wind speed has positive coefficients. This indicates that the odds of willingness-to-pool may increase with colder/"worse" weather in non-pandemic context. This is intuitive as colder weather is commonly associated with disruptions of public transportation. People are also generally less willing to spend time outdoors during "bad" weather which may lead to raise in the demand for door-to-door transportation such as ride-hailing. In this scenario, being open for pooling would rise the chance of getting a ride. On the other hand, these weather-related effects are reversed in the post-outbreak models. This is consistent with the belief during the COVID-19 pandemic that the spread of the virus was weaker during warm weather. Moreover, given a pandemic, the vaccination level of the population seems to possibly play the biggest role in riders' willingness to pool.

6.1.2 Factors Affecting Pooling Success

Looking at the 'Pool' models, the effects of fare rate and trip distance are common in all models. This indicates that a decrease in fare rate and an increase in trip distance may rise the odds of pooling success regardless of a pandemic. Likewise, outside of afternoon times, the odds of pooling success may decrease. These effects persist in both non-pandemic or pandemic context.

In contrast to the pre-outbreak model, however, the effect of additional charges flips and gains in magnitude in the post-outbreak models, meaning that this factor may become more influential to pooling success given a pandemic.

Similarly, the effects of weekday and average temperature also reverse. The latter is consistent with the previous finding mentioned in Section 6.1.1 that "bad" weather may encourage willingess-to-pool and hence subsequently increase the chance of pooling success. The effect of wind speed, however, is present in the pre-outbreak model but may not be relevant in the post-outbreak context, while average humidity may only be relevant in the post-outbreak models.

Interestingly, in normal times, having the destination outside of Special Zones may increase the odds of pooling success, while it is the otherwise in pandemic times. Moreover, the effects of holiday and the pickup tract being a Special Zone may be present in post-outbreak context but not in the pre-outbreak. Finally, both the post-outbreak Stepwise-selected model and Lasso-selected model agree that the case rate may have the most dominating influence towards pooling success given a pandemic.

6.1.3 Pandemic's Relevance in Willingness to Pool and Pooling Success

As highlighted in the Section 6.1.1 and 6.1.1, there are discrepancies between the found influencing factors in non-pandemic and pandemic context. This suggests that a pandemic-like situation does alter riders behaviour towards ride-pooling and alter the factors to pooling success. Not to mention that in both 'WTP' and 'Pool' model sets, pandemic-related measures may have dominating influence towards willingness to pool or pooling success given a pandemic.

6.1.4 The Impacts of Downtown Zone Taxing in Willingness to Pool and Pooling Success

The recently instated Downtown Zone taxing (refer to Section 3.1) was also discovered through this research to potentially have impacts on willingness to pool and pooling success. This is based on the fact that the pickup or dropoff location within the Downtown Zone mostly have significant coefficients in the Lasso-selected post-outbreak 'WTP' and 'Pool' models. The positive coefficients mean that the odds of WTP and pooling success may increases when the trip starts or ends in the Downtown Zone.

6.2 Variable Selection Methodology Review

This section outlines the pros and contras of the Backward Elimination (Stepwise) and Lasso variable selection methods and compares the experiences of utilising the two techniques in this thesis.

Automated methods such as Stepwise selection is beneficial when one is faced with a large number of candidate predictors. However, some of the known drawbacks of Stepwise regression include invalid p-value as well as biased coefficient estimations, confidence intervals, and R^2 [161]. Moreover, as Stepwise does not consider all of the possible combinations of variables, it may not pick the best one [169]. In some cases, it may even leave out actual good predictors, leading to poor out-of-sample fit [170]. Efforts have been done in this thesis to mitigate these drawbacks of Stepwise by conducting the selection procedure over multiple separate sample sets, having the selected variables re-fitted with regular MLE, and validating with out-of-sample test sets. However, Stepwise's tendency to exacerbate collinearity problems necessitated a manual pre-selection of variables which itself may impose subjective biases.

The Lasso technique is more robust against the variance inflating issue of multicollinearity. This suggests that pre-elimination of correlated variables is not necessary. However, when faced with a pair of highly correlated variables, Lasso arbitrarily chooses to keep one and eliminate the other. This may lead to less intuitive variables being chosen. Moreover, due to the shrinkage that Lasso introduces, the resulting coefficients are biased towards zero and could not be inferred as the true magnitude of the variable's effect. Other disadvantages of Lasso include generally unstable estimates which may lead to different sets of chosen variables given different dataset. For the above reasons, this thesis utilised Lasso over separate training sets and only for variable selection.

Reflecting on the experiences during this thesis, the Backward Stepwise and Lasso eliminations generally agreed with each other on the variables to keep or eliminate and both resulted with models with excellent fit. In terms of practicality, Lasso excelled with much faster computation speed and since it did not require prior variable pre-selection (which could also introduce bias). However, the true importance of the chosen variables by Lasso is a question mark due to the reasons mentioned above. Moreover, due to coefficient shrinkage bias by Lasso, the Stepwise procedure could detect variables with sign instability earlier in the process, while such variables might elude the Lasso elimination process to only be detected in the re-fitting step. This was similar to the significance level of Lasso-selected variables that—due to Lasso's nature—could only be checked after re-fitting with regular MLE.

Chapter 7

Conclusion

Upon realising the benefits of ride-pooling over regular single ride-hailing, this thesis set out to investigate the influencing factors behind people's willingness to pool (WTP) and pooling request successfully matched. The ride-hailing industry in the City of Chicago was taken as a case study, and building on past studies, a wide range of potential factors were considered. This thesis also considered the recent COVID-19 pandemic as an opportunity to investigate the impacts of a pandemic on the mentioned factors. A systematic methodology was established to enable handling, processing, and analysing large-sized ride-hailing trip data integrated with various exogenous factors. Statistical modelling of willingness to pool and pooling success at a trip level were carried out involving as much available trip data as possible to avoid information loss by sampling or aggregating. To identify the most influential factors, the modelling methodology incorporated two different variable selection methods: (Backward) Stepwise selection and penalised regression with Lasso. The final models yielded excellent predictive performance, suggesting the ability to explain the factors to willingness to pool and pooling success well.

To answer the research questions of this thesis, the main potential factors influencing rider's WTP mostly consist of the attributes of the trip itself, i.e., the additional charges, fare rate, and the temporal details of the trip (the time of day and whether it is on a weekday or weekend). Aside from these factors, there are discrepancies between non-pandemic and pandemic context. In the former, the amount of tip, the trip distance, income level of the pickup and dropoff census tract, and whether or not the trip begins or ends in a Special Zone possibly play a role in WTP. This may not be the case in pandemic context where vaccination level of the population potentially pose the dominating effect. Weather attributes also have potential significance in rider's WTP, although the effects may be opposing in non-pandemic and pandemic context. The final models suggest that other exogenous and demographic factors such as crime rate, land use, population density, level of vehicle ownership, and public transit access do not pose significant effects on WTP.

Similarly, the main potential factors influencing pooling success were also identified. The trip attributes which may remain influential in both pandemic and non-pandemic scenarios are the fare rate, trip distance, and the time of day of the trip. The factors which may be present in both scenarios but have opposing effects are average temperature, weekday, additional charge, and the destination being within the Special Zones. Meanwhile, case rate potentially has the greatest influence on pooling success in a pandemic scenario. As in WTP, weather may play some role in pooling success, but the models suggest that factors such as land use, crime rate, and other exogenous or demographic variables do not have a part.

This thesis also concludes that pandemic indeed potentially alters the influence of various factors on WTP and pooling success. Meanwhile, policies such as special taxing for downtown areas may contribute to WTP and pooling success.

Based on the above results, to encourage riders to pool and enhance pooling success, this thesis recommends:

- **TNPs as the service providers to optimise the pool matching algorithm.** A review on the literature revealed the role of matching algorithm and its parameters play on matching success
- **TNPs to optimise the fee structure and the authorities to subsidise for ride-pooling.** This thesis found that lower price encourages WTP. On top of that, Romeo et al. claimed that ride-pooling decision is price-sensitive [96] and a survey revealed that one of the top reasons riders opted for not pooling was that the price was not significantly cheaper than the single-occupant counterpart [83]
- **TNPs to ensure sufficient supply of pooling-suitable vehicles** during time windows where the odds of WTP and pooling success is higher
- **TNPs to maximise cleanliness and health measures during and post-pandemic** to alleviate the concerns of virus spread during ride-pooling
- **Authorities to implement taxing policies** for cars entering certain zones, especially congested ones. This thesis found that having the origin or destination within specially-taxed area (Special Zones or Downtown Zone) possibly affect the odds of WTP and pooling success. This could also push down congestion

7.1 Limitations and Future Work

Ultimately, the methodology proposed in this thesis is advantageous for works involving large-sized data. However, as previously outlined, it is not without a flaw. Further limitations and the subsequent recommendations for future work are discussed below.

The Stepwise procedure relied on the correlation coefficients to rule out highly correlated variables, however this thesis missed to check the significance of each coefficient. Moreover, this step is prone to error due to subjectivity. This thesis also discovered that should the Spearman's coefficients were utilised instead of Pearson's, the resulting pre-selected variables could vary.

The Lasso technique is less prone to the mentioned issue. However, the nature of Lasso selection makes the true importance of the resulting variables questionable. Moreover, both Stepwise and Lasso eliminations treat multiple levels of a categorical variable as a separate individual predictor and not as a one related unit.

Consequently, for works involving large number of variables, this thesis recommends exploring other variable selection methods such as Principal Component Logistic Regression (PCLR) or Elastic Net Regression which may perform better than Lasso depending on the level of multicollinearity [171], [172], as well as Group Lasso which enables grouped selection of variables [173]. Moreover, this thesis could have also benefited from larger post-outbreak WTP trip dataset which would have enabled assessment of the post-outbreak 'Pool' models over multiple sample sets.

Finally, this thesis inherited the limitation of Chicago's TNP trip dataset where fees and pickup/dropoff locations are aggregated. The lack of riders' characteristics information also meant that socioeconomic/demographic data were at neighbourhood-level, whereas finer details (e.g., at individual level) of these data could reveal more knowledge on the effects. Moreover, there was no data on the trip purpose and waiting time, whereas these factors have been claimed to affect travel decisions [49], [50], [84]. Lastly, this thesis cannot guarantee the validity of the above results in other cities.

Appendix A

Descriptions of Variables

A.1 Weather Variables

TABLE A.1: Descriptions of the variables related to weather

No	Variable Name (Short)	Description	Data Type	Unit	Pre-Outbreak	Post-Outbreak
1	temp_avg	The average temperature on the day of the trip	Numerical	°C	●	●
2	hum_avg	The average humidity on the day of the trip	Numerical	%	●	●
3	wind_avg	The average wind speed on the day of the trip	Numerical	km/h	●	●
4	ppt_avg	The average precipitation on the day of the trip	Numerical	cm	●	●

A.2 Trip Attributes Variables

TABLE A.2: Description of the trip attribute variables

No	Variable Name (Short)	Description	Data Type	Unit	Pre-Outbreak	Post-Outbreak
1	wtp	1 = trip was authorised by the passenger to be pooled 0 = trip was not authorised for pooling	Categorical (Binary)	-	●	●
2	pool	1 = trip was successfully pooled 0 = the trip was not pooled	Categorical (Binary)	-	●	●
3	dow	The day of the week on which the trip was started. Contains 7 levels from ’Monday’ to ’Sunday’	Categorical	-	●	●
4	tod	The time of the day at which the trip was started. Aggregated into 4 levels: ’morning’ (0-5), ’before_noon’ (6-12), ’afternoon’ (13-19) and ’night’ (20-23)	Categorical	-	●	●
5	trip_mins	The total duration of the trip	Numerical	minutes	●	●
6	trip_kms	The total length of the trip	Numerical	km	●	●
7	fare	The fare of the trip excluding additional charges and tip. Rounded to the nearest 2.50	Numerical	$	●	●
8	fare_per_km	The fare of the trip per kilometer	Numerical	$/km	●	●
9	tip	The tip given by the passenger. Not including cash tip. Rounded to the nearest 1.00	Numerical	$	●	●
10	add_charges	Additional charges on top of the fare (e.g. taxes)	Numerical	$	●	●
11	total_pay	The summation of fare, tip, and additional charges	Numerical	$	●	●
12	holiday	1 = the trip was during a public holiday 0 = the trip was during a regular day	Categorical (Binary)	-	●	●
13	weekday	1 = the trip occurred during a weekday 0 = the trip occurred during the weekend	Categorical (Binary)	-	●	●

A.3 Spatial/Demographic Variables

TABLE A.3: Descriptions of the spatial/demographic variables related to the tracts involved in the trips

No	Variable Name (Short)	Description	Data Type	Unit	Pre-Outbreak	Post-Outbreak
1	special_..*	1 = the tract is considered as Special Zone	Categorical (Binary)	-	●	●
		0 = the tract is a regular area				
2	downtown_..*	1 = the tract is considered as Special Zone	Categorical (Binary)	-		●
		0 = the tract is a regular area				
3	LU_entropy_..*	The landuse entropy of the tract	Numerical	-	●	●
4	LU_prevail_..*	The prevailing land-use type of the tract	Categorical	-	●	●
5	LU_B_ptg_..*	The proportion of 'Business' land-use in the tract	Numerical	-	●	●
6	LU_C_ptg_..*	The proportion of 'Commercial' land-use in the tract	Numerical	-	●	●
7	LU_D_ptg_..*	The proportion of 'Downtown' land-use in the tract	Numerical	-	●	●
8	LU_M_ptg_..*	The proportion of 'Manufacture' land-use in the tract	Numerical	-	●	●
9	LU_PD_ptg_..*	The proportion of 'Planned Development' land-use in the tract	Numerical	-	●	●
10	LU_PMD_ptg_..*	The proportion of 'Planned Manufacture Development' land-use in the tract	Numerical	-	●	●
11	LU_POS_ptg_..*	The proportion of 'POS' land-use in the tract	Numerical	-	●	●
12	LU_R_ptg_..*	The proportion of 'Residential' land-use in the tract	Numerical	-	●	●
13	LU_T_ptg_..*	The proportion of 'Transportation' land-use in the tract	Numerical	-	●	●
14	pop_dens_..*	Population density of the tract	Numerical	$/km^2$	●	●
15	worker_dens_..*	Density of the employed population at the tract	Numerical	$/km^2$	●	●
16	income_pc_..*	Income per capita of the tract	Numerical	$	●	●
17	dist_CBD_..*	Distance of the tract centroid to the centroid of the central business district	Numerical	km	●	●
18	area_..*	Area of the tract	Numerical	km^2	●	●

B.3 Spatial/Demographic Variables (continued)

TABLE A.4: Descriptions of the spatial/demographic variables related to the tracts involved in the trips (continued)

No	Variable Name (Short)	Description	Data Type	Unit	Pre-Outbreak	Post-Outbreak
19	PT_avg_..*	Density of the public transit stops in the tract		km^2	●	●
20	veh_0_ptg_..*	The proportion of the population at the tract who do not own a vehicle	Numerical	-	●	●
21	crimes_avg_..*	The average monthly crime cases which occurred in the tract	Numerical	-	●	●

* The suffixes ′pu′ and ′do′ indicate the measure in the pickup and dropoff tract respectively

A.4 Pandemic Variables

TABLE A.5: Descriptions of variables related to the progression of the COVID-19 pandemic

No	Variable Name (Short)	Description	Data Type	Unit	Pre-Outbreak	Post-Outbreak
1	covid_case_rate	The 7-day rolling average of the COVID-19 cases in Chicago on the day of the trip	Numerical	-		●
2	covid_death_rate	The 7-day rolling average of the COVID-19 deaths in Chicago on the day of the trip	Numerical	-		●
3	covid_hospitalised_rate	The 7-day rolling average of the COVID-19 hospitalisation in Chicago on the day of the trip	Numerical	-		●
4	covid_vac_cml	The cumulative completed vaccine series in Chicago on the day of the trip	Numerical	-		●

Appendix B

Descriptive Statistics

TABLE B.1: Descriptive statistics of the complete variables in two study periods

Variable	Pre-outbreak				Post-outbreak			
	Min	Mean	Max	Std. Dev.	Min	Mean	Max	Std. Dev.
add_charges	0.000	2.449	10.00	0.074	0.000	3.316	10.00	0.324
area_do	0.069	0.783	21.64	0.026	0.069	0.810	21.64	0.033
area_pu	0.069	0.768	21.64	0.021	0.069	0.792	21.64	0.026
covid_case_rate	-	-	-	-	5.000	29.28	257.3	36.77
covid_death_rate	-	-	-	-	0.000	0.141	1.200	0.228
covid_hospitalised_rate	-	-	-	-	0.200	1.424	9.100	1.388
covid_vac_cml	-	-	-	-	1432473	1785642	1901146	132498
crimes_avg_do	0.250	76.14	388.3	5.243	0.333	54.18	195.7	2.528
crimes_avg_pu	0.250	70.81	388.3	3.885	0.333	52.30	195.7	2.140
dist_CBD_do	0.495	4.442	26.05	0.197	0.495	4.358	26.05	0.221
dist_CBD_pu	0.495	4.505	26.05	0.189	0.495	4.481	26.05	0.218
fare	0.000	8.549	30.00	0.564	0.000	14.97	30.00	1.954
fare_per_km	0.000	2.022	31.07	0.126	0.000	3.815	31.07	0.502
hum_avg	28.90	69.92	95.40	7.522	32.00	60.88	89.70	5.814
income_pc_do	1629	77542	171616	1319	1629	79611	171616	1600
income_pc_pu	1629	77287	171616	1302	1629	78816	171616	1603
LU_B_ptg_do	0.000	0.071	0.495	0.005	0.000	0.068	0.495	0.005
LU_B_ptg_pu	0.000	0.073	0.495	0.004	0.000	0.071	0.495	0.004
LU_C_ptg_do	0.000	0.043	0.534	0.001	0.000	0.045	0.534	0.002
LU_C_ptg_pu	0.000	0.045	0.534	0.001	0.000	0.046	0.534	0.001
LU_D_ptg_do	0.000	0.180	0.698	0.011	0.000	0.185	0.698	0.010
LU_D_ptg_pu	0.000	0.172	0.698	0.010	0.000	0.177	0.698	0.009
LU_entropy_do	3.E-04	0.673	0.998	0.006	0.000	0.678	0.998	0.006
LU_entropy_pu	3.E-04	0.666	0.998	0.005	0.000	0.672	0.998	0.005
LU_M_ptg_do	0.000	0.029	0.799	0.001	0.000	0.027	0.799	0.001
LU_M_ptg_pu	0.000	0.030	0.799	0.001	0.000	0.028	0.799	0.001
LU_PD_ptg_do	0.000	0.241	1.000	0.011	0.000	0.252	1.000	0.009
LU_PD_ptg_pu	0.000	0.233	1.000	0.009	0.000	0.244	1.000	0.008
LU_PMD_ptg_do	0.000	0.031	0.707	0.001	0.000	0.032	0.707	0.001
LU_PMD_ptg_pu	0.000	0.033	0.707	0.001	0.000	0.033	0.707	0.001
LU_POS_ptg_do	0.000	0.053	0.885	0.003	0.000	0.057	0.885	0.003
LU_POS_ptg_pu	0.000	0.053	0.885	0.002	0.000	0.056	0.885	0.003
LU_R_ptg_do	0.000	0.293	0.977	0.018	0.000	0.269	0.977	0.019
LU_R_ptg_pu	0.000	0.303	0.977	0.016	0.000	0.282	0.977	0.017
LU_T_ptg_do	0.000	6E-04	0.060	0.000	0.000	0.001	0.060	3.E-05
LU_T_ptg_pu	0.000	6E-04	0.060	0.000	0.000	0.000	0.060	3.E-05

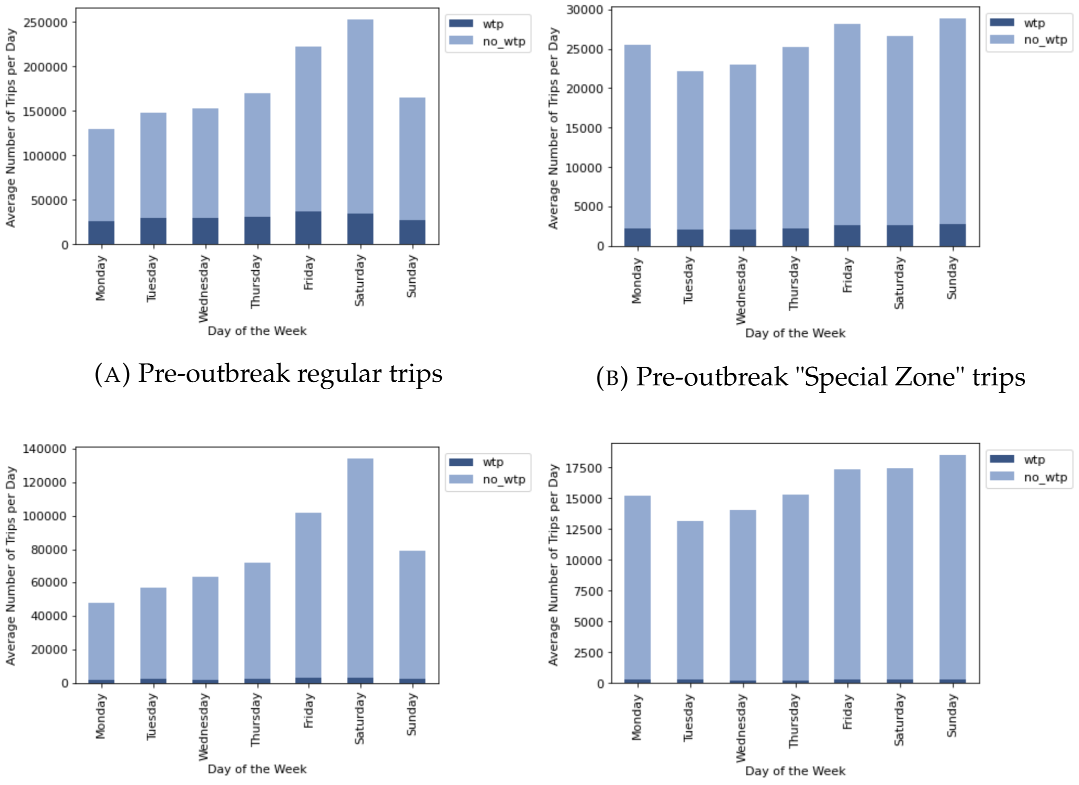
TABLE B.2: Descriptive statistics of the complete variables in two study periods (continued)

Variable	Pre-outbreak				Post-outbreak			
	Min	**Mean**	**Max**	**Std. Dev.**	**Min**	**Mean**	**Max**	**Std. Dev.**
pop_dens_do	140.2	10050	43029	150.3	140.2	10198	43029	202.1
pop_dens_pu	140.2	10173	43029	125.5	140.2	10294	43029	188.7
ppt_avg	0.000	2.292	58.17	2.765	0.000	2.070	71.88	2.151
PT_avg_do	0.000	5.485	41.25	0.336	0.000	5.401	41.25	0.273
PT_avg_pu	0.000	5.142	41.25	0.251	0.000	5.186	41.25	0.242
temp_avg	-26.56	10.24	30.61	10.85	-14.94	13.45	33.00	10.03
tip	0.000	0.460	100.0	0.059	0.000	0.981	200.0	0.128
total_pay	0.000	11.46	130.6	0.657	0.000	19.26	217.5	2.172
trip_kms	0.966	5.729	32.19	0.120	0.966	5.519	32.19	0.133
trip_mins	1.017	14.27	60.00	0.545	1.017	13.14	60.00	0.581
veh_0_ptg_do	0.000	28.80	72.44	0.458	0.000	29.67	72.44	0.457
veh_0_ptg_pu	0.000	28.39	72.44	0.387	0.000	29.32	72.44	0.442
wind_avg	5.950	16.20	42.49	3.110	4.180	16.43	34.92	3.165
worker_dens_do	49.91	6834	31344	116.7	49.91	6978	31344	163.2
worker_dens_pu	49.91	6919	31344	104.9	49.91	7036	31344	158.3

Appendix C

Multivariate Distributions

C.1 Average Counts of WTP Trips and Non-WTP Trips over Different Days of the Week



(A) Pre-outbreak regular trips

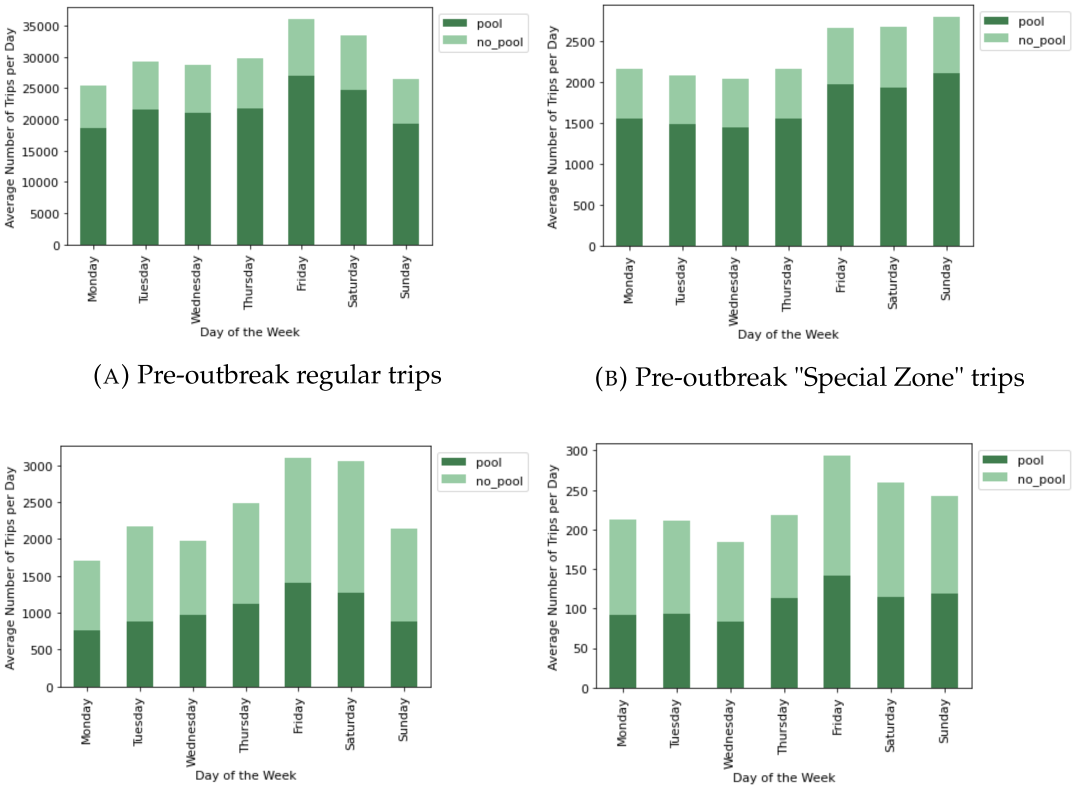
(B) Pre-outbreak "Special Zone" trips

(C) Post-outbreak regular trips

(D) Post-outbreak "Special Zone" trips

FIGURE C.1: Variations of average trip counts of WTP and non-WTP trips over different trip types over different days of the week

C.2 Average Counts of Pooled Trips and Non-Pooled WTP Trips over Different Days of the Week



(A) Pre-outbreak regular trips

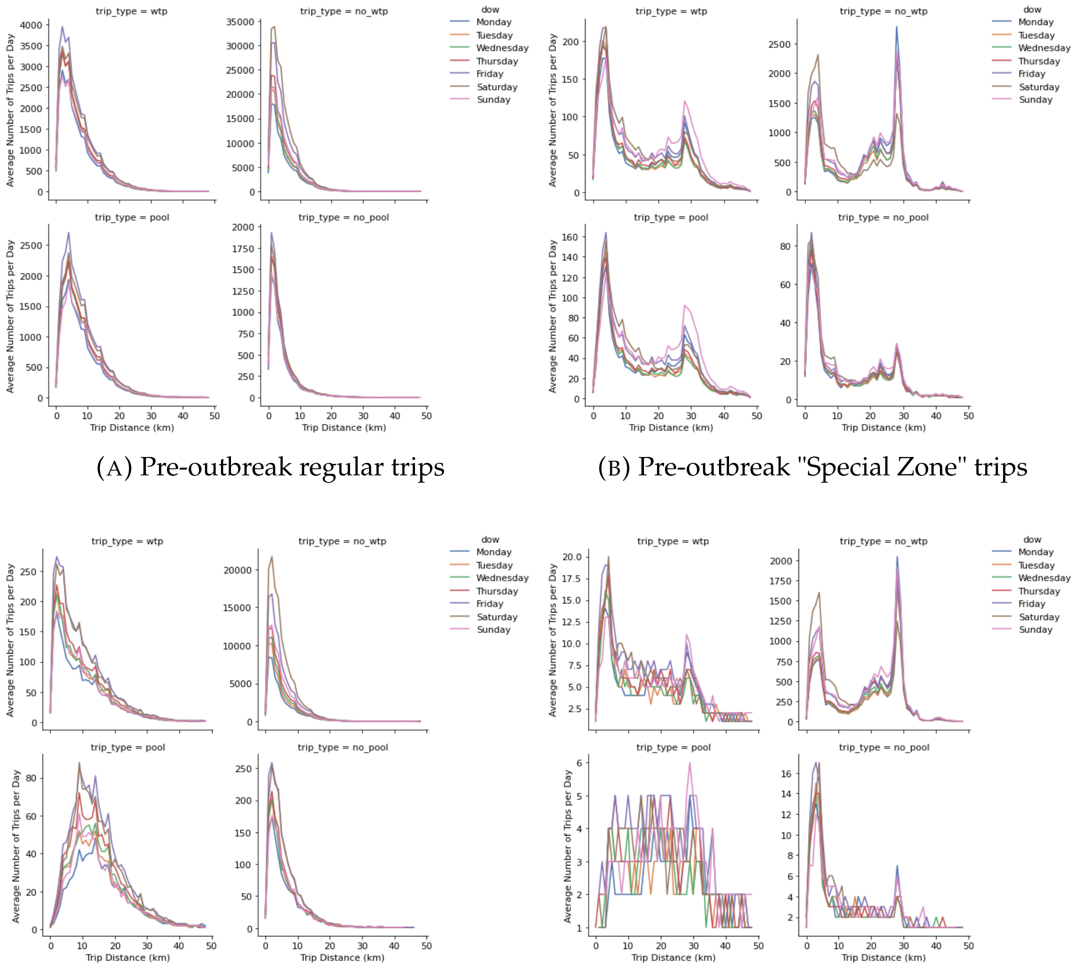
(B) Pre-outbreak "Special Zone" trips

(C) Post-outbreak regular trips

(D) Post-outbreak "Special Zone" trips

FIGURE C.2: Variations of average trip counts of pooled and non-pooled WTP trips over different trip types over different days of the week

C.3 Trip Duration over Different Days of the Week



(A) Pre-outbreak regular trips

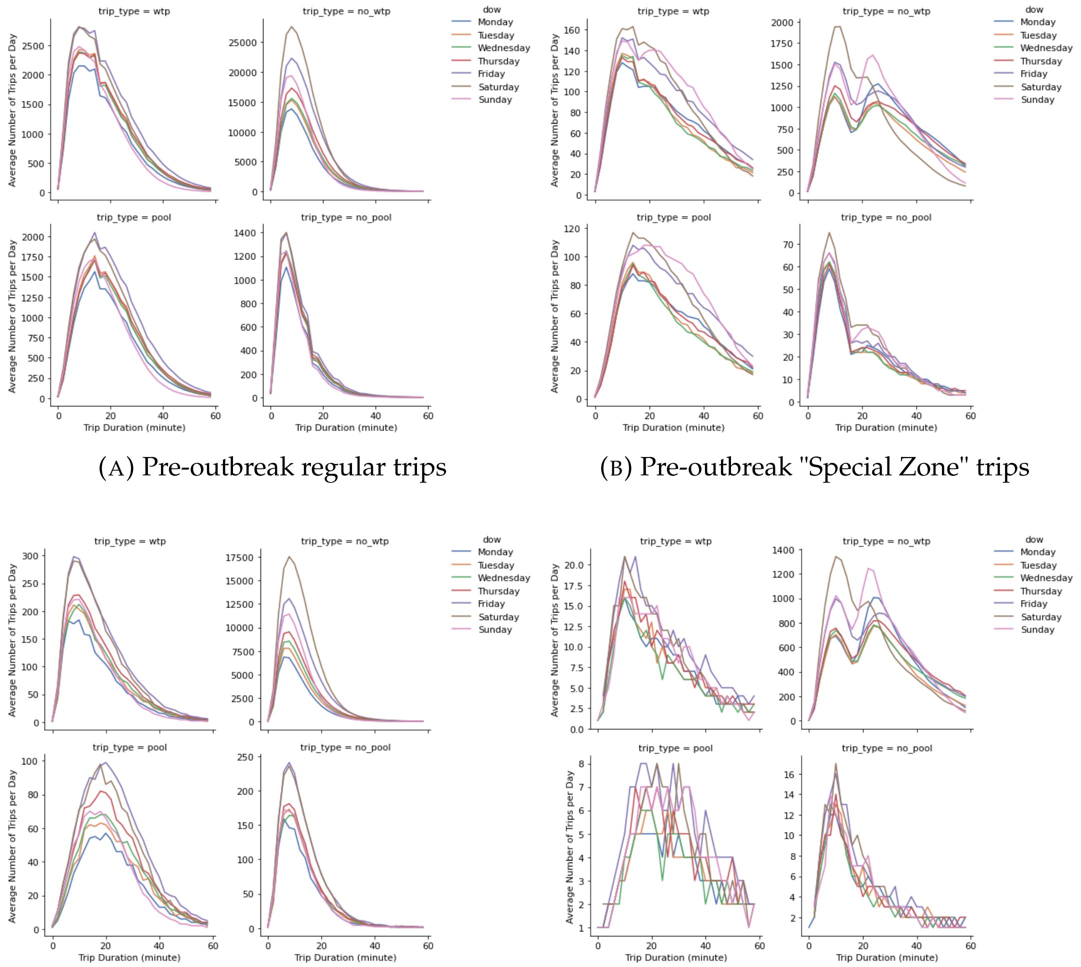
(B) Pre-outbreak "Special Zone" trips

(C) Post-outbreak regular trips

(D) Post-outbreak "Special Zone" trips

FIGURE C.3: Variations of trip duration of different trip types over different days of the week

C.4 Trip Distance over Different Days of the Week



(A) Pre-outbreak regular trips

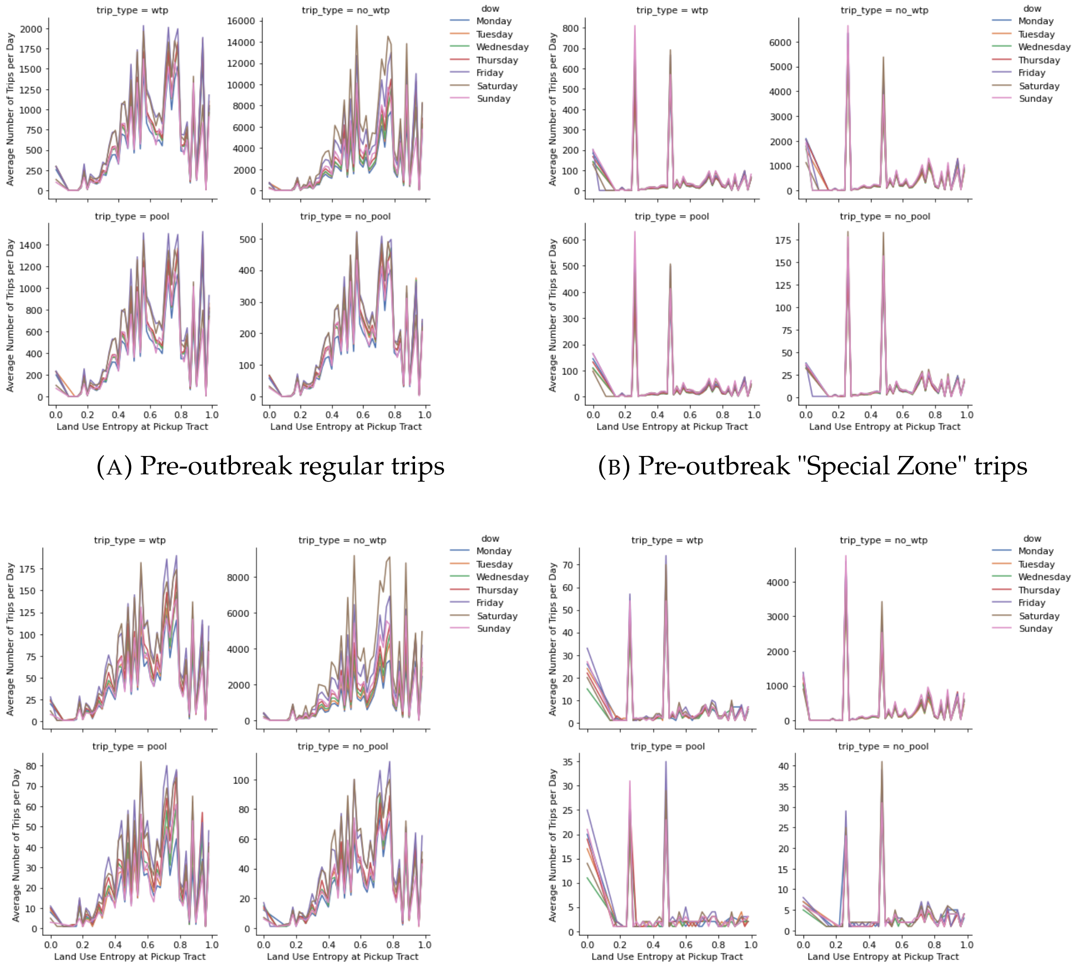
(B) Pre-outbreak "Special Zone" trips

(C) Post-outbreak regular trips

(D) Post-outbreak "Special Zone" trips

FIGURE C.4: Variations of trip distance of different trip types over different days of the week

C.5 Land Use Entropy of Pickup Tract over Different Days of the Week



(A) Pre-outbreak regular trips

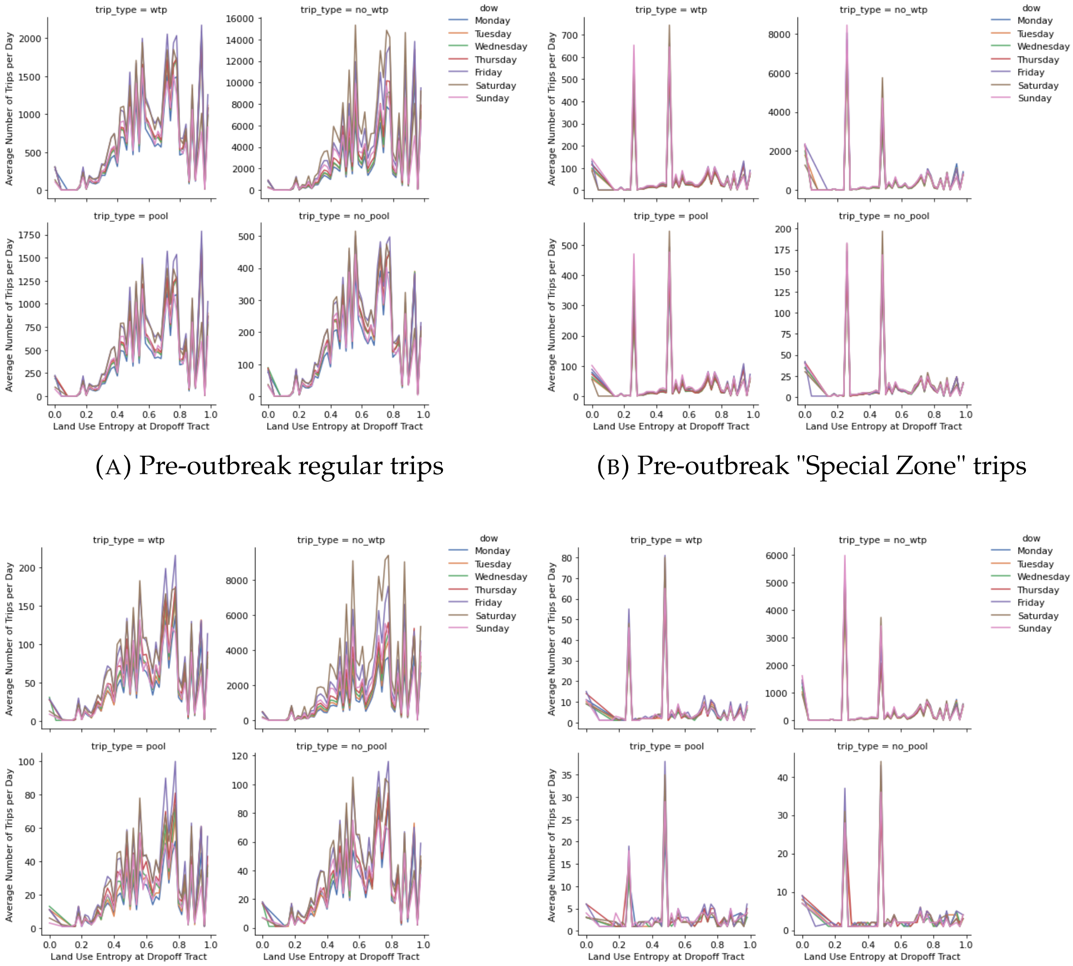
(B) Pre-outbreak "Special Zone" trips

(C) Post-outbreak regular trips

(D) Post-outbreak "Special Zone" trips

FIGURE C.5: Variations of land use entropy at the pickup tract of different trip types over different days of the week

C.6 Land Use Entropy of Pickup Tract over Different Days of the Week



(A) Pre-outbreak regular trips

(B) Pre-outbreak "Special Zone" trips

(C) Post-outbreak regular trips

(D) Post-outbreak "Special Zone" trips

FIGURE C.6: Variations of land use entropy at the dropoff tract of different trip types over different days of the week

Appendix D

Correlation Matrices

D.1 Correlation Matrices between Continuous Variables

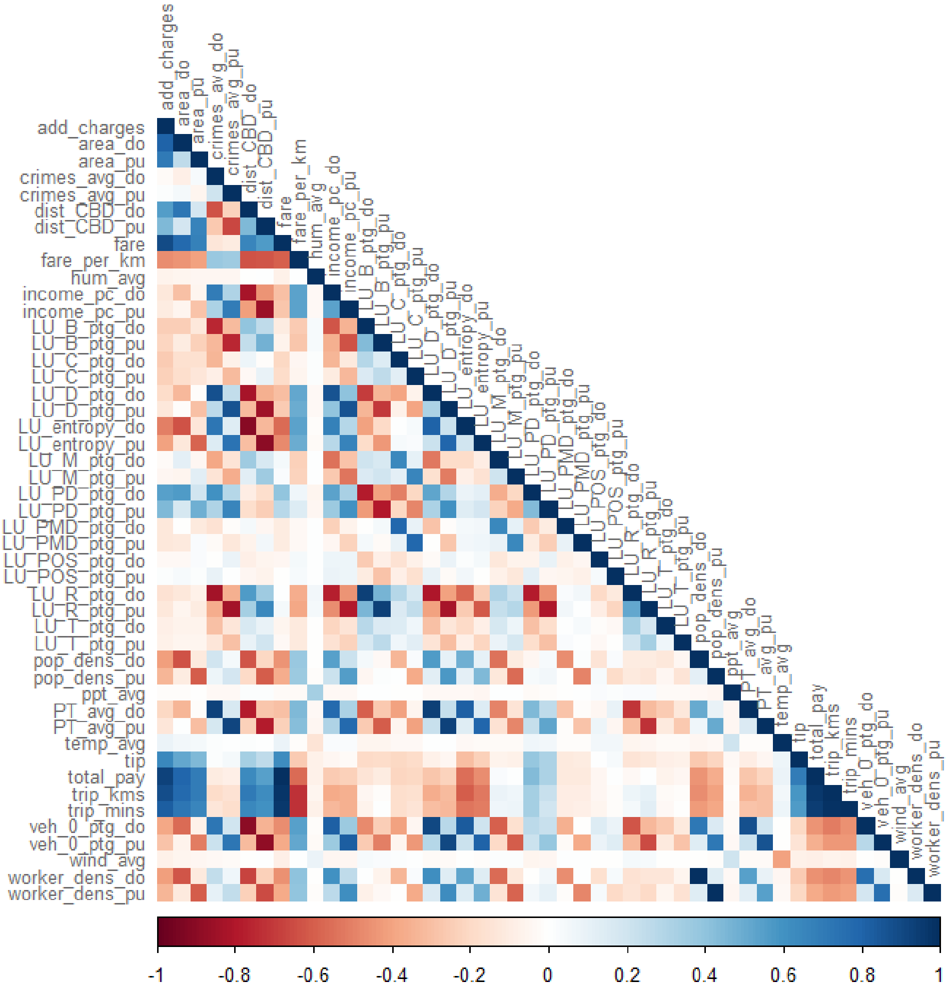


FIGURE D.1: Pearson's coefficient between continuous variables (pre-outbreak)

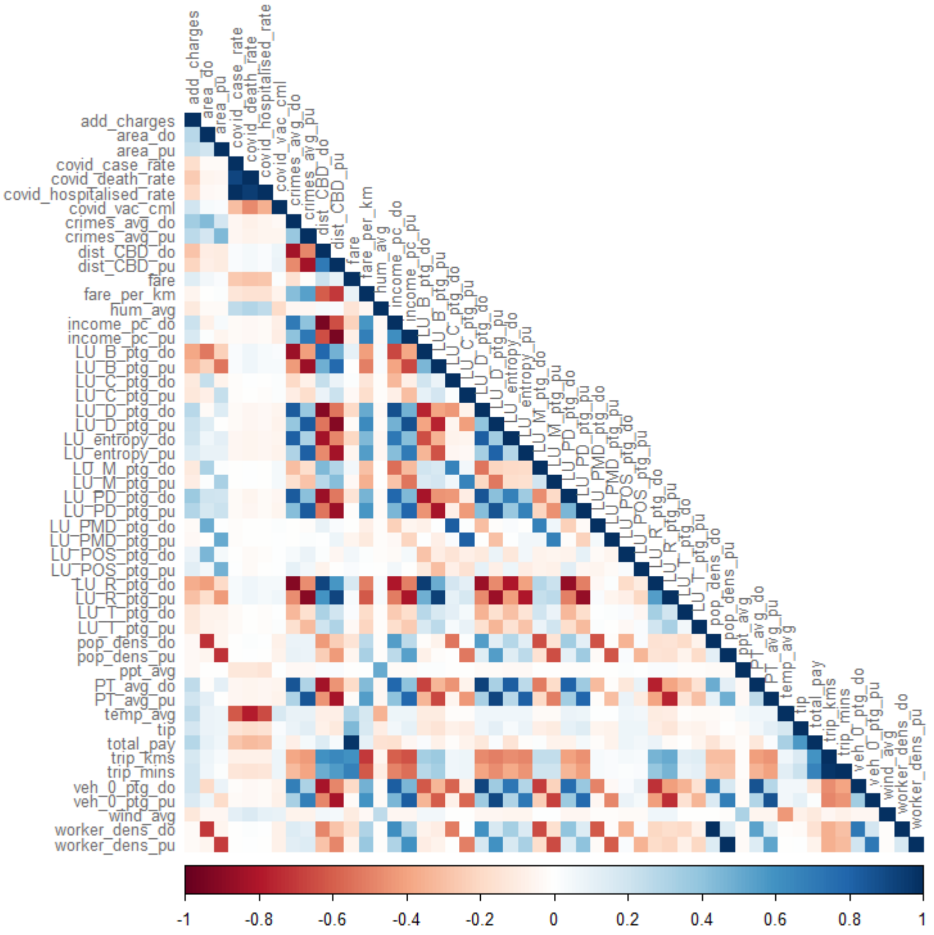


FIGURE D.2: Pearson's coefficient between continuous variables (post-outbreak)

D.2 Correlation Matrices between Continuous and Categorical Variables

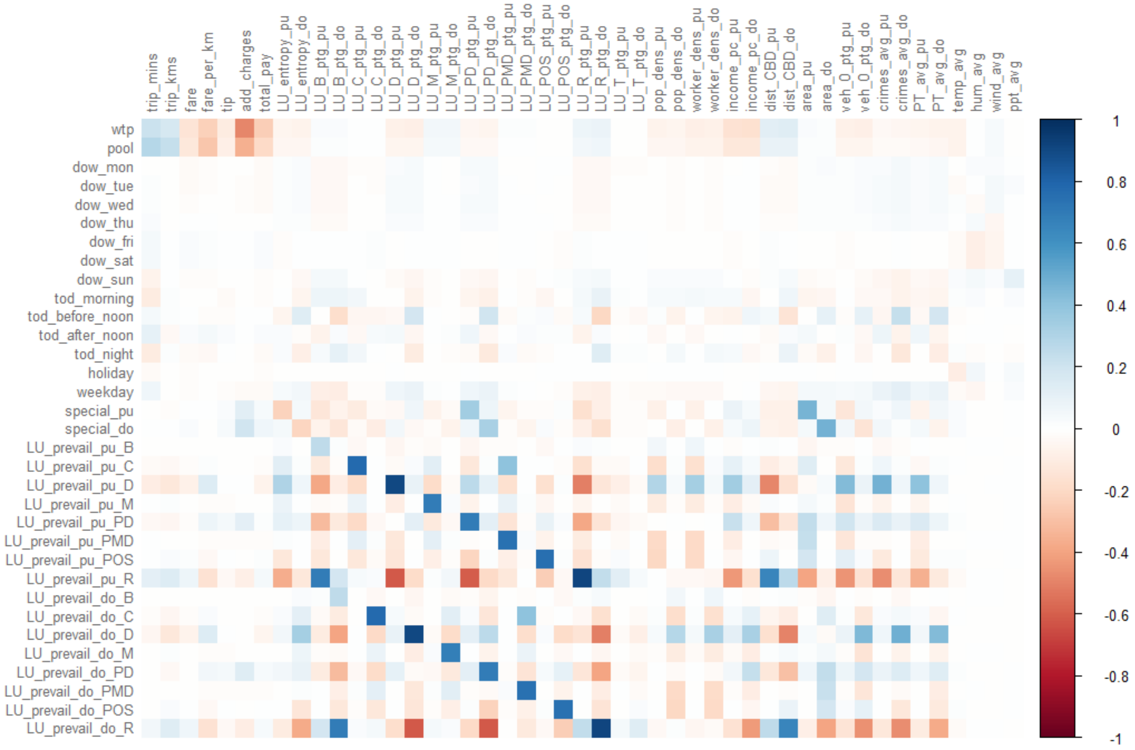


FIGURE D.3: Point biserial (Pearson's) coefficients between continuous and categorical variables (pre-outbreak)

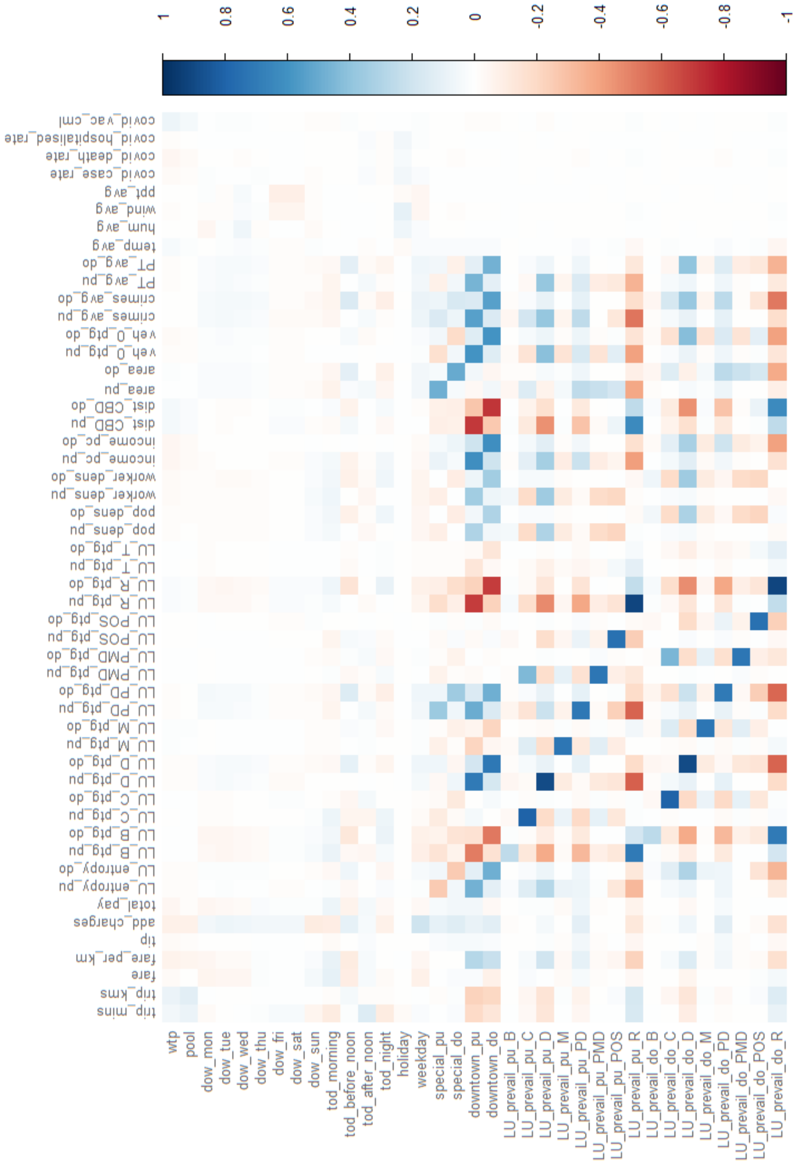


FIGURE D.4: Point biserial (Pearson's) coefficients between continuous and categorical variables (post-outbreak)

Bibliography

[1] H. Ritchie, *Sector by sector: Where do global greenhouse gas emissions come from? - our world in data*, Sep. 2020. [Online]. Available: https://ourworldindata.org/ghg-emissions-by-sector.

[2] IEA, "Key world energy statistics 2020", Aug. 2020. [Online]. Available: https://www.iea.org/reports/key-world-energy-statistics-2020.

[3] J. Conti, P. Holtberg, J. Diefenderfer, A. LaRose, J. T. Turnure, and L. Westfall, "International energy outlook 2016 with projections to 2040", U.S. Energy Information Administration, May 2016. DOI: 10.2172/1296780. [Online]. Available: https://www.osti.gov/biblio/1296780/.

[4] OECD, *Passenger transport (indicator)*, 2022. DOI: 10.1787/463da4d1. [Online]. Available: https://data.oecd.org/transport/passenger-transport.htm.

[5] C. A. S. Machado, N. Patrick, M. D. S. Hue, F. T. Berssaneti, and J. A. Quintanilha, "An overview of shared mobility", *Sustainability*, vol. 10, p. 4342, 12 2018. DOI: 10.3390/su10124342.

[6] Precedence Research, "Shared mobility market size, share, trends, growth 2022-2030", 2021. [Online]. Available: https://www.precedenceresearch.com/shared-mobility-market.

[7] Statista, *Shared mobility - worldwide*, Aug. 2022. [Online]. Available: https://www.statista.com/outlook/mmo/shared-mobility/worldwide#revenue.

[8] Grand View Research, "Global shared mobility market size report, 2022 - 2030", 2020. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/shared-mobility-market.

[9] Statista, *Shared mobility - united states*, Aug. 2022. [Online]. Available: https://www.statista.com/outlook/mmo/shared-mobility/united-states.

[10] ——, *Shared rides - united states*, Aug. 2022. [Online]. Available: https://www.statista.com/outlook/mmo/shared-mobility/shared-rides/united-states#revenue.

[11] T. Dong, L. Jiao, G. Xu, L. Yang, and J. Liu, "Towards sustainability? analyzing changing urban form patterns in the united states, europe, and china", *Science of The Total Environment*, vol. 671, pp. 632–643, Jun. 2019, ISSN: 0048-9697. DOI: 10.1016/J.SCITOTENV.2019.03.269.

[12] G. Mattioli, C. Roberts, J. K. Steinberger, and A. Brown, "The political economy of car dependence: A systems of provision approach", *Energy Research and Social Science*, vol. 66, Aug. 2020, ISSN: 22146296. DOI: 10.1016/J.ERSS.2020.101486.

[13] S. Shaheen, *Shared mobility: The potential of ridehailing and pooling*, D. Sperling, Ed., 2018. DOI: 10.5822/978-1-61091-906-7_3. [Online]. Available: https://link.springer.com/chapter/10.5822/978-1-61091-906-7_3.

[14] M. H. Chen, A. Jauhri, and J. P. Shen, "Data driven analysis of the potentials of dynamic ride pooling", Association for Computing Machinery, Nov. 2017, pp. 7–12, ISBN: 1595930361. DOI: 10.1145/3151547.3151549.

[15] Research and Innovative Technology Administration, “Ridesharing options analysis and practitioners’ toolkit”, U.S. Department of Transportation, Dec. 2010. [Online]. Available: https://www.planning.dot.gov/documents/ridesharingoptions_toolkit.pdf.

[16] S. R. Gehrke, A. Felix, and T. G. Reardon, “Substitution of ride-hailing services for more sustainable travel options in the greater boston region”, *Transportation Research Record*, vol. 2673, pp. 438–446, 1 2019. DOI: 10.1177/0361198118821903.

[17] A. Henao and W. E. Marshall, “The impact of ride-hailing on vehicle miles traveled”, *Transportation*, vol. 46, pp. 2173–2194, 6 Dec. 2019, ISSN: 15729435. DOI: 10.1007/S11116-018-9923-2/TABLES/6. [Online]. Available: https://link.springer.com/article/10.1007/s11116-018-9923-2.

[18] A. O.-L. Rosa, E. G. Chuquichambi, and G. P. Ingram, “Keep your (social) distance: Pathogen concerns and social perception in the time of covid-19”, *Personality and individual differences*, vol. 166, Nov. 2020, ISSN: 0191-8869. DOI: 10.1016/J.PAID.2020.110200. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32834278/.

[19] M. Marani, G. G. Katul, W. K. Pan, and A. J. Parolari, “Intensity and frequency of extreme novel epidemics”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 118, e2105482118, 35 Aug. 2021, ISSN: 10916490. DOI: 10.1073/PNAS.2105482118/SUPPL_FILE/PNAS.2105482118.SAPP.PDF. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2105482118.

[20] J. Attard, F. Orlandi, S. Scerri, and S. Auer, “A systematic review of open government data initiatives”, *Government Information Quarterly*, vol. 32, pp. 399–418, 4 Oct. 2015, ISSN: 0740-624X. DOI: 10.1016/J.GIQ.2015.07.006.

[21] OECD, *Open government data*. [Online]. Available: https://www.oecd.org/gov/digital-government/open-government-data.htm.

[22] M. Hossain, “Sharing economy: A comprehensive literature review”, *International Journal of Hospitality Management*, vol. 87, p. 102 470, May 2020, ISSN: 0278-4319. DOI: 10.1016/J.IJHM.2020.102470.

[23] R. Belk, “You are what you can access: Sharing and collaborative consumption online”, *Journal of Business Research*, vol. 67, pp. 1595–1600, 8 Aug. 2014, ISSN: 0148-2963. DOI: 10.1016/J.JBUSRES.2013.10.001.

[24] S. Shaheen, A. Cohen, N. Chan, and A. Bansal, “Sharing strategies: Carsharing, shared micromobility (bikesharing and scooter sharing), transportation network companies, microtransit, and other innovative mobility modes”, *Transportation, Land Use, and Environmental Planning*, pp. 237–262, Jan. 2019. DOI: 10.1016/B978-0-12-815167-9.00013-X.

[25] F. T. Administration, *Shared mobility definitions*, Feb. 2020. [Online]. Available: https://www.transit.dot.gov/regulations-and-guidance/shared-mobility-definitions.

[26] S. Castellanos, S. Grant-Muller, and K. Wright, “Technology, transport, and the sharing economy: Towards a working taxonomy for shared mobility”, *Transport Reviiews*, vol. 42, 3 2021. DOI: 10.1080/01441647.2021.1968976. [Online]. Available: https://www.tandfonline.com/action/journalInformation?journalCode=ttrv20.

[27] Y. Ke-fei, “Parking demand forecasting based on parking space sharing”, *Urban Transport of China*, 2009.

[28] M. Lai, X. Cai, and Q. Hu, “Market design for commute-driven private parking lot sharing”, *Transportation Research Part C: Emerging Technologies*, vol. 124, p. 102 915, Mar. 2021, ISSN: 0968-090X. DOI: 10.1016/J.TRC.2020.102915.

[29] R. Hahn and R. Metcalfe, *The ridesharing revolution: Economic survey and synthesis*, S. D. Kominers and A. Teytelboym, Eds., Jan. 2017. [Online]. Available: https://en.wikipedia.org/wiki/Carma.

[30] Mordor Intelligence, *Ride-hailing maret - growth, trends, covid-19 impact, and forcasts (2022-2027)*, 2022. [Online]. Available: https://www.mordorintelligence.com/industry-reports/ride-hailing-market.

[31] Cognitive Market Research, *Global ride hailing market report 2022*, 2022. [Online]. Available: https://www.cognitivemarketresearch.com/ride-hailing-market-report.

[32] S. Stasha, *Ride-sharing industry statistics to get you going in 2022*, Sep. 2022. [Online]. Available: https://policyadvice.net/insurance/insights/ride-sharing-industry-statistics/.

[33] E. B. Salas, *Market share of the leading ride-hailing companies in the united states from september 2017 to july 2021*, Sep. 2022. [Online]. Available: https://www.statista.com/statistics/910704/market-share-of-rideshare-companies-united-states/.

[34] B. Dean, *Uber statistics 2022: How many people ride with uber?*, Mar. 2021. [Online]. Available: https://backlinko.com/uber-users.

[35] L. Rayle, D. Dai, N. Chan, R. Cervero, and S. Shaheen, "Just a better taxi? a survey-based comparison of taxis, transit, and ridesourcing services in san francisco", *Transport Policy*, vol. 45, pp. 168–178, Jan. 2016, ISSN: 0967-070X. DOI: 10.1016/J.TRANPOL.2015.10.004.

[36] R. R. Clewlow and G. S. Mishra, "Disruptive transportation: The adoption, utilization, and impacts of ride-hailing in the united states", UC Davis Institute of Transportation Studies, Oct. 2017.

[37] R. A. Acheampong, A. Siiba, D. K. Okyere, and J. P. Tuffour, "Mobility-on-demand: An empirical study of internet-based ride-hailing adoption factors, travel characteristics and mode substitution effects", *Transportation Research Part C: Emerging Technologies*, vol. 115, p. 102 638, Jun. 2020, ISSN: 0968-090X. DOI: 10.1016/J.TRC.2020.102638.

[38] A. Tirachini and A. Gomez-Lobo, "Does ride-hailing increase or decrease vehicle kilometers traveled (vkt)? a simulation approach for santiago de chile", *International Journal of Sustainable Transportation*, vol. 14, pp. 187–204, 3 Mar. 2020, ISSN: 15568334. DOI: 10.1080/15568318.2018.1539146. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/15568318.2018.1539146.

[39] B. Schaller, "The new automobility: Lyft, uber and the future of american cities", Schaller Consulting, Jul. 2018. [Online]. Available: http://www.schallerconsult.com/rideservices/automobility.pdf.

[40] S. Sikder, "Who uses ride-hailing services in the united states?", *Transportation Research Record*, vol. 2673, pp. 40–54, 12 Jun. 2019, ISSN: 21694052. DOI: 10.1177/0361198119859302. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/0361198119859302.

[41] F. F. Dias, P. S. Lavieri, T. Kim, C. R. Bhat, and R. M. Pendyala, "Fusing multiple sources of data to understand ride-hailing use", *Transportation Research Record*, vol. 2673, pp. 214–224, 6 May 2019, ISSN: 21694052. DOI: 10.1177/0361198119841031. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0361198119841031.

[42] O. Sabogal-Cardona, D. Oviedo, L. Scholl, A. Crotte, and F. Bedoya-Maya, "Not my usual trip: Ride-hailing characterization in mexico city", *Travel Behaviour and Society*, vol. 25, pp. 233–245, Oct. 2021, ISSN: 2214-367X. DOI: 10.1016/J.TBS.2021.07.010.

[43] M. Young and S. Farber, "The who, why, and when of uber and other ride-hailing trips: An examination of a large sample household travel survey", *Transportation Research Part A: Policy and Practice*, vol. 119, pp. 383–392, Jan. 2019, ISSN: 0965-8564. DOI: 10.1016/J.TRA.2018.11.018.

[44] M. J. Mohamed, T. Rye, and A. Fonzone, "The utilisation and user characteristics of uber services in london", *Transportation Planning and Technology*, vol. 43, pp. 424–441, 4 May 2020, ISSN: 10290354. DOI: 10.1080/03081060.2020.1747205. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/03081060.2020.1747205.

[45] S. Li, W. Zhai, J. Jiao, and C. K. Wang, "Who loses and who wins in the ride-hailing era? a case study of austin, texas", *Transport Policy*, vol. 120, pp. 130–138, May 2022, ISSN: 0967-070X. DOI: 10.1016/J.TRANPOL.2022.03.009.

[46] S. Qiao and A. G.-O. Yeh, "Is ride-hailing competing or complementing public transport? a perspective from affordability", *Transportation Research Part D: Transport and Environment*, vol. 114, p. 103 533, Jan. 2023, ISSN: 1361-9209. DOI: 10.1016/J.TRD.2022.103533.

[47] C. Rodier, "The effects of ride hailing services on travel and associated greenhouse gas emissions", UC Davis Institute of Transportation Studies, Apr. 2018.

[48] F. Alemi, G. Circella, S. Handy, and P. Mokhtarian, "What influences travelers to use uber? exploring the factors affecting the adoption of on-demand ride services in california", *Travel Behaviour and Society*, vol. 13, pp. 88–104, Oct. 2018, ISSN: 2214-367X. DOI: 10.1016/J.TBS.2018.06.002.

[49] J. E. J. Evans, *Transit scheduling and frequency*, May 2004. DOI: 10.17226/23433.

[50] K. F. Turnbull and R. H. Pratt, *Transit information and promotion*, Mar. 2004. DOI: 10.17226/23386.

[51] B. J. Tang, X. Y. Li, B. Yu, and Y. M. Wei, "How app-based ride-hailing services influence travel behavior: An empirical study from china", *International Journal of Sustainable Transportation*, vol. 14, pp. 554–568, 7 Jul. 2020, ISSN: 15568334. DOI: 10.1080/15568318.2019.1584932.

[52] C. Bialik, A. Flowers, R. Fischer-Baum, and D. Mehta, *Uber is serving new york's outer boroughs more than taxis are*, Aug. 2015. [Online]. Available: https://fivethirtyeight.com/features/uber-is-serving-new-yorks-outer-boroughs-more-than-taxis-are/.

[53] R. Fischer-Baum and C. Bialik, *Uber is taking millions of manhattan rides away from taxis*, Oct. 2015. [Online]. Available: https://fivethirtyeight.com/features/uber-is-taking-millions-of-manhattan-rides-away-from-taxis/.

[54] S. T. Jin, H. Kong, R. Wu, and D. Z. Sui, "Ridesourcing, the sharing economy, and the future of cities", *Cities*, vol. 76, pp. 96–104, Jun. 2018, ISSN: 0264-2751. DOI: 10.1016/J.CITIES.2018.01.012.

[55] A. Brown, "Redefining car access: Ride-hail travel and use in los angeles", *Journal of the American Planning Association*, vol. 85, pp. 83–95, 2 Apr. 2019, ISSN: 01944363. DOI: 10.1080/01944363.2019.1603761. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01944363.2019.1603761.

[56] A. Tirachini, *Plataformas ridesourcing (tipo uber y cabify) en chile: Impactos en movilidad y recomendaciones para su regulación*, Jul. 2017. [Online]. Available: https://www.researchgate.net/publication/318429681_Plataformas_ridesourcing_tipo_Uber_y_Cabify_en_Chile_impactos_en_movilidad_y_recomendaciones_para_su_regulacion.

[57] ——, "Ride-hailing, travel behaviour and sustainable mobility: An international review", *Transportation*, vol. 47, pp. 2011–2047, 4 Aug. 2020, ISSN: 15729435. DOI: 10.1007/S11116-019-10070-2/TABLES/5. [Online]. Available: https://link.springer.com/article/10.1007/s11116-019-10070-2.

[58] A. P. T. Association, "Shared mobility and the transformation of public transit", Mar. 2016. [Online]. Available: www.sharedusemobilitycenter.org.

[59] T. Litman, "Well measured: Developing indicators for comprehensive and sustainable transport planning", *Transportation Research Record*, vol. 2017, pp. 10–15, 1 2007, ISSN: 03611981. DOI: 10.3141/2017-02.

[60] Y. Wang, W. Shi, and Z. Chen, "Impact of ride-hailing usage on vehicle ownership in the united states", *Transportation Research Part D: Transport and Environment*, vol. 101, p. 103 085, Dec. 2021, ISSN: 1361-9209. DOI: 10.1016/J.TRD.2021.103085.

[61] S. Agarwal, D. Mani, and R. Telang, "The impact of ride-hailing services on congestion: Evidence from indian cities", *Manufacturing Service Operations Management (Forthcoming)*, Jun. 2019. DOI: 10.2139/SSRN.3410623. [Online]. Available: https://papers.ssrn.com/abstract=3410623.

[62] J. A. Parrott and M. Reich, "An earnings standard for new york city's app-based drivers: Economic analysis and policy assessment", UC Berkely Center on Wage and Employment Dynamics, Jul. 2018. [Online]. Available: https://irle.berkeley.edu/an-earnings-standard-for-new-york-citys-app-based-drivers/.

[63] K. J. Wells, "The uber workplace in d.c.", Georgetown University Kalmanovitz Initiative for Labor and the Working Poor.

[64] Z. Wadud, "The effects of e-ridehailing on motorcycle ownership in an emerging-country megacity", *Transportation Research Part A: Policy and Practice*, vol. 137, pp. 301–312, Jul. 2020, ISSN: 0965-8564. DOI: 10.1016/J.TRA.2020.05.002.

[65] A. Henao, "Impacts of ridesourcing - lyft and uber - on transportation including vmt, mode replacement, parking, and travel behavior", University of Colorado Denver, May 2017. [Online]. Available: https://www.proquest.com/openview/5486ff6cc229889a3cdf2df1cd3993cb/1?pq-origsite=gscholar&cbl=18750.

[66] J. Cramer and A. B. Krueger, "Disruptive change in the taxi business: The case of uber", *American Economic Review*, vol. 106, pp. 177–82, 5 May 2016, ISSN: 0002-8282. DOI: 10.1257/AER.P20161002.

[67] T. Wenzel, C. Rames, E. Kontou, and A. Henao, "Travel and energy implications of ridesourcing service in austin, texas", *Transportation Research Part D: Transport and Environment*, vol. 70, pp. 18–34, May 2019, ISSN: 1361-9209. DOI: 10.1016/J.TRD.2019.03.005.

[68] D. Anair, J. Martin, M. C. P. de Moura, and J. Goldman, "Ride-hailing's climate risks: Steering a growing industry toward a clean transportation future", Union of Concerned Scientists, 2020. [Online]. Available: https://www.ucsusa.org/resources/ride-hailing-climate-risks.

[69] X. Wu and D. MacKenzie, "Assessing the vmt effect of ridesourcing services in the us", *Transportation Research Part D: Transport and Environment*, vol. 94, p. 102 816, May 2021, ISSN: 1361-9209. DOI: 10.1016/J.TRD.2021.102816.

[70] G. D. Erhardt, S. Roy, D. Cooper, B. Sana, M. Chen, and J. Castiglione, "Do transportation network companies decrease or increase congestion?", *Science Advances*, vol. 5, 5 2019, ISSN: 23752548. DOI: 10.1126/SCIADV.AAU2670. [Online]. Available: https://www.science.org/doi/10.1126/sciadv.aau2670.

[71] X. Feng, Q. Lin, N. Jia, and J. Tian, *The actual impact of ride-splitting: An empirical study based on large-scale gps data*. [Online]. Available: https://ssrn.com/abstract=4132998.

[72] X. Liu, W. Li, Y. Li, J. Fan, and Z. Shen, "Quantifying environmental benefits of ridesplitting based on observed data from ridesourcing services", *Transportation Research Record*, vol. 2675, pp. 355–368, 8 Mar. 2021, ISSN: 21694052. DOI: 10.1177/0361198121997827. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/0361198121997827.

[73] X. Chen, H. Zheng, Z. Wang, and X. Chen, "Exploring impacts of on-demand ridesplitting on mobility via real-world ridesourcing data and questionnaires", *Transportation*, vol. 48, pp. 1541–1561, 2021. DOI: 10.1007/s11116-018-9916-1. [Online]. Available: https://doi.org/10.1007/s11116-018-9916-1.

[74] W. Li, Z. Pu, Y. Li, and M. Tu, "How does ridesplitting reduce emissions from ridesourcing? a spatiotemporal analysis in chengdu, china", *Transportation Research Part D: Transport and Environment*, vol. 95, p. 102 885, Jun. 2021, ISSN: 1361-9209. DOI: 10.1016/J.TRD.2021.102885.

[75] IEA, "World energy outlook 2005", IEA, 2005. [Online]. Available: https://www.iea.org/reports/world-energy-outlook-2005.

[76] A. Bilali, U. Fastenrath, and K. Bogenberger, "Analytical model to estimate ride pooling traffic impacts by using the macroscopic fundamental diagram", *Transportation Research Record*, vol. 2676, pp. 697–709, 4 2022. DOI: 10.1177/03611981211064892.

[77] FHWA, *Ridesharing options analysis and practitioners' toolkit*, 2010. [Online]. Available: https://www.planning.dot.gov/documents/ridesharingoptions_toolkit.pdf.

[78] D. Sperling, S. Pike, and R. Chase, *Will the transportation revolutions improve our lives—or make them worse?*, D. Sperling, Ed., 2018. DOI: 10.5822/978-1-61091-906-7_1. [Online]. Available: https://link.springer.com/chapter/10.5822/978-1-61091-906-7_1.

[79] J. Ke, H. Yang, and Z. Zheng, "On ride-pooling and traffic congestion", *Transportation Research Part B: Methodological*, vol. 142, pp. 213–231, Dec. 2020, ISSN: 0191-2615. DOI: 10.1016/J.TRB.2020.10.003.

[80] W. Li, Y. Li, Z. Pu, L. Cheng, L. Wang, and L. Yang, *Revealing the co2 emission reduction of ridesplitting and its determinants based on real-world data*, Apr. 2022. DOI: 10.48550/arxiv.2204.00777. [Online]. Available: https://arxiv.org/abs/2204.00777v2.

[81] Z. Ma, H. N. Koutsopoulos, and Y. Zheng, "Evaluation of on-demand ridesplitting services", Jan. 2019. [Online]. Available: https://www.researchgate.net/publication/338061174_Evaluation_of_On-Demand_Ridesplitting_Services.

[82] N. Kostorz, E. Fraedrich, and M. Kagerbauer, "Usage and user characteristics—insights from moia, europe's largest ridepooling service", *Sustainability*, vol. 13, p. 958, 2 Jan. 2021, ISSN: 2071-1050. DOI: 10.3390/SU13020958. [Online]. Available: https://www.mdpi.com/2071-1050/13/2/958/.

[83] Z. Wang, X. Chen, and X. M. Chen, "Ridesplitting is shaping young people's travel behavior: Evidence from comparative survey via ride-sourcing

platform", *Transportation Research Part D: Transport and Environment*, vol. 75, pp. 57–71, Oct. 2019, ISSN: 1361-9209. DOI: 10.1016/J.TRD.2019.08.017.

[84] S. R. Gehrke, H. M. P., and T. G. Reardon, "Social and trip-level predictors of pooled ride-hailing service adoption in the greater boston region", *Case Studies on Transport Policy*, vol. 9, pp. 1026–1034, 3 2021. DOI: 10.1016/j.cstp.2021.05.004.

[85] J. Wang, Q. Wu, Z. Chen, Y. Ren, and Y. Gao, "Exploring the factors of intercity ridesplitting based on observed and gis data: A case study in china", *ISPRS International Journal of Geo-Information*, vol. 10, p. 622, 9 Sep. 2021, ISSN: 2220-9964. DOI: 10.3390/IJGI10090622. [Online]. Available: https://www.mdpi.com/2220-9964/10/9/622.

[86] M. Taiebat, E. Amini, and M. Xu, "Sharing behavior in ride-hailing trips: A machine learning inference approach", *Transportation Research Part D: Transport and Environment*, vol. 103, Feb. 2022, ISSN: 13619209. DOI: 10.1016/J.TRD.2021.103166.

[87] W. Li, Z. Pu, Y. Li, and X. Ban, "Characterization of ridesplitting based on observed data: A case study of chengdu, china", *Transportation Research Part C: Emerging Technologies*, vol. 100, pp. 330–353, Mar. 2019, ISSN: 0968-090X. DOI: 10.1016/J.TRC.2019.01.030.

[88] M. D. Dean and K. M. Kockelman, "Spatial variation in shared ride-hail trip demand and factors contributing to sharing: Lessons from chicago", *Journal of Transport Geography*, vol. 91, p. 102 944, Feb. 2021, ISSN: 0966-6923. DOI: 10.1016/J.JTRANGEO.2020.102944.

[89] F. Zwick, E. Fraedrich, and K. W. Axhausen, "Ride-pooling in the light of covid-19: Determining spatiotemporal demand characteristics on the example of moia", *IET Intelligent Transport Systems*, pp. 1–16, 2022. DOI: 10.1049/itr2.12293. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/itr2.12293.

[90] Y. Hou, V. Garikapati, D. Weigl, A. Henao, M. Moniot, and J. Sperling, "Factors influencing willingness to pool in ride-hailing trips", *Transportation Research Record*, vol. 2674, pp. 419–429, 5 May 2020, ISSN: 21694052. DOI: 10.1177/0361198120915886. [Online]. Available: https://journals.sagepub.com/doi/10.1177/0361198120915886.

[91] Z. Chen, H. Yu, and R. Liu, "Exploring the spatiotemporal factors of ridesplitting demand based on the geographically and temporally weighted regression", *2021 IEEE Conference on Telecommunications, Optics and Computer Science, TOCS 2021*, pp. 245–250, 2021. DOI: 10.1109/TOCS53301.2021.9688587.

[92] M. Du, L. Cheng, X. Li, Q. Liu, and J. Yang, "Spatial variation of ridesplitting adoption rate in chicago", *Transportation Research Part A: Policy and Practice*, vol. 164, pp. 13–37, Oct. 2022, ISSN: 0965-8564. DOI: 10.1016/J.TRA.2022.07.018.

[93] Z. Chen, Y. Ren, X. Zheng, and X. Fu, "Impact of covid-19 on spatiotemporal factors affecting ridesplitting demand", American Society of Civil Engineers, 2022, pp. 899–908, ISBN: 9780784484265. DOI: 10.1061/9780784484265.084. [Online]. Available: https://ascelibrary.org/doi/10.1061/9780784484265.084.

[94] H. Abkarian, S. Hegde, and H. S. Mahmassani, "Does taxing tnc trips discourage solo riders and increase the demand for ride pooling? a case study of chicago using interrupted time series and bayesian hierarchical modeling", *Transportation Research Record*, vol. 2677, 1 Jul. 2022, ISSN: 0361-1981.

DOI: 10.1177/03611981221098665. [Online]. Available: https://journals.sagepub.com/doi/10.1177/03611981221098665.

[95] Y. Xu, X. Yan, X. Liu, and X. Zhao, "Identifying key factors associated with ridesplitting adoption rate and modeling their nonlinear relationships", *Transportation Research Part A: Policy and Practice*, vol. 144, pp. 170–188, Feb. 2021, ISSN: 0965-8564. DOI: 10.1016/J.TRA.2020.12.005.

[96] L. Romeo, C. Atkinson-Palombo, N. Garrick, and D. Chacón-Hurtado, "Characteristics of pooled trips offered by ridesourcing services in chicago", 2021.

[97] L. Wang, W. Li, J. Weng, D. Zhang, and W. Ma, "Do low-carbon rewards incentivize people to ridesplitting? evidence from structural analysis", *Transportation*, pp. 1–33, Jun. 2022, ISSN: 15729435. DOI: 10.1007/S11116-022-10302-Y/TABLES/18. [Online]. Available: https://link.springer.com/article/10.1007/s11116-022-10302-y.

[98] H. Abkarian, Y. Chen, and H. S. Mahmassani, "Understanding ridesplitting behavior with interpretable machine learning models using chicago transportation network company data", *Transportation Research Record*, vol. 2676, pp. 83–99, 2 Sep. 2021, ISSN: 21694052. DOI: 10.1177/03611981211036363. [Online]. Available: https://journals.sagepub.com/doi/epub/10.1177/03611981211036363.

[99] A. Maria, "Introduction to modeling and simulation", S Andradottir, K. J. Healy, D. H. Withers, and B. L. Nelson, Eds., 1997. DOI: 10.1145/268437.268440.

[100] J. A. Chamizo, "Filosofía de la química: I. sobre el método y los modelos", *Educación Química*, vol. 20, pp. 6–11, 1 Jan. 2009, ISSN: 0187-893X. DOI: 10.1016/S0187-893X(18)30002-8.

[101] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Wiley, Aug. 2013, pp. 1–510, ISBN: 9781118548387. DOI: 10.1002/9781118548387. [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387.

[102] A. Field, *Discovering Statistics Using IBM SPSS Statistics*, 4th ed. Sage Publications, 2013, pp. 293–356, ISBN: 1446249174. [Online]. Available: https://books.google.com/books/about/Discovering_Statistics_Using_IBM_SPSS_St.html?id=srb0a9fmMEoC.

[103] R. I. Jennrich and S. M. Robinson, "A newton-raphson algorithm for maximum likelihood factor analysis", *Psychometrika*, vol. 34, pp. 111–123, 1 Mar. 1969, ISSN: 00333123. DOI: 10.1007/BF02290176/METRICS. [Online]. Available: https://link.springer.com/article/10.1007/BF02290176.

[104] J. Wang, "A bisection algorithm for finding estimates of parameters in the maximum likelihood function", *Annals of the Shanghai Observatory, Academia Sinica*, vol. 18, pp. 45–48, 1997. [Online]. Available: https://ui.adsabs.harvard.edu/abs/1997AnShO..18...45W/abstract.

[105] A. Hyvärinen, "Fixed-point algorithm and maximum likelihood estimation for independent component analysis", *Neural Processing Letters*, vol. 10, pp. 1–5, 1 1999, ISSN: 13704621. DOI: 10.1023/A:1018647011077/METRICS. [Online]. Available: https://link.springer.com/article/10.1023/A:1018647011077.

[106] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma", *Neural Computation*, vol. 4, pp. 1–58, 1 Jan. 1992, ISSN: 0899-7667. DOI: 10.1162/NECO.1992.4.1.1. [Online]. Available: https://direct.mit.edu/neco/article/4/1/1/5624/Neural-Networks-and-the-Bias-Variance-Dilemma.

[107] S. Doroudi, "The bias-variance tradeoff: How data science can inform educational debates", *AERA Open*, vol. 6, p. 233 285 842 097 720, 4 Dec. 2020, ISSN: 2332-8584. DOI: 10.1177/2332858420977208. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/2332858420977208.

[108] H. Akaike, "Information theory and an extension of the maximum likelihood principle", B. Petrov and F. Csaki, Eds., Academiai Kiado, 1973, pp. 267–281.

[109] S. I. Vrieze, "Model selection and psychological theory: A discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic)", *Psychological Methods*, vol. 17, pp. 228–243, 2 Jun. 2012, ISSN: 1082989X. DOI: 10.1037/A0027127. [Online]. Available: /record/2012-03019-001.

[110] C. M. Hurvich and C. L. Tsai, "Regression and time series model selection in small samples", *Biometrika*, vol. 76, pp. 297–307, 2 Jun. 1989, ISSN: 0006-3444. DOI: 10.1093/BIOMET/76.2.297. [Online]. Available: https://academic.oup.com/biomet/article/76/2/297/265326.

[111] G. Schwarz, "Estimating the dimension of a model", *Annals of Statistics*, vol. 6, pp. 461–464, 2 Mar. 1978, ISSN: 0090-5364. DOI: 10.1214/AOS/1176344136. [Online]. Available: https://typeset.io/papers/estimating-the-dimension-of-a-model-4nf5v47tw0.

[112] J. M. Obeso, P. García, B. Martínez, F. N. Arroyo-López, A. Garrido-Fernández, and A. Rodriguez, "Use of logistic regression for prediction of the fate of staphylococcus aureus in pasteurized milk in the presence of two lytic phages", *Applied and Environmental Microbiology*, vol. 76, pp. 6038–6046, 18 Sep. 2010, ISSN: 00992240. DOI: 10.1128/AEM.00613-10/. [Online]. Available: https://journals.asm.org/doi/10.1128/AEM.00613-10.

[113] C. Thrane, "Examining tourists' long-distance transportation mode choices using a multinomial logit regression model", *Tourism Management Perspectives*, vol. 15, pp. 115–121, Jul. 2015, ISSN: 2211-9736. DOI: 10.1016/J.TMP.2014.10.004.

[114] G. A. J. Hemmert, L. M. Schons, J. Wieseke, and H. Schimmelpfennig, "Log-likelihood-based pseudo-r 2 in logistic regression: Deriving sample-sensitive benchmarks", *Sociological Methods Research*, vol. 47, pp. 507–531, 3 2018. DOI: 10.1177/0049124116638107.

[115] D. McFadden, *Conditional logit analysis of qualitative choice behavior*, P Zarembka, Ed., 1974.

[116] G. S. Maddala, *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, 1983, vol. 3, p. 401, ISBN: 0521338255. [Online]. Available: https://books.google.com/books/about/Limited_Dependent_and_Qualitative_Variab.html?id=-Ji1ZaUg7gcC.

[117] D. R. Cox and E. Snell, *Analysis of Binary Data*, 2nd ed. Chapman and Hall, 1990, vol. 32, ISBN: 9780412306204. [Online]. Available: https://www.routledge.com/Analysis-of-Binary-Data/Cox-Snell/p/book/9780412306204.

[118] N. M. Kebonye, "Exploring the novel support points-based split method on a soil dataset", *Measurement*, vol. 186, p. 110 131, Dec. 2021, ISSN: 0263-2241. DOI: 10.1016/J.MEASUREMENT.2021.110131.

[119] A. Rácz, D. Bajusz, and K. Héberger, "Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification", *Molecules 2021, Vol. 26, Page 1111*, vol. 26, p. 1111, 4 Feb. 2021, ISSN: 1420-3049. DOI: 10.3390/MOLECULES26041111. [Online]. Available: https://www.mdpi.com/1420-3049/26/4/1111/.

[120] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification", *Lecture Notes in Computer Science (including subseries Lecture Notes*

in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3056, pp. 22–30, 2004, ISSN: 16113349. DOI: 10.1007/978-3-540-24775-3_5/COVER. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-24775-3_5.

[121] S. Visa, B. Ramsay, A. Ralescu, and E. van der Knaap, "Confusion matrix-based feature selection", S. Visa, A. Inoue, and A. Ralescu, Eds., Omnipress, Apr. 2011.

[122] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms", *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, pp. 299–310, 3 2005.

[123] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, T. Münkemüller, C. Mcclean, P. E. Osborne, B. Reineking, B. Schröder, A. K. Skidmore, D. Zurell, and S. Lautenbach, "Collinearity: A review of methods to deal with it and a simulation study evaluating their performance", *Ecography*, vol. 36, pp. 27–46, 1 Jan. 2013, ISSN: 1600-0587. DOI: 10.1111/J.1600-0587.2012.07348.X. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1600-0587.2012.07348.x.

[124] J. C. de Winter, S. D. Gosling, and J. Potter, "Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data", *Psychological Methods*, vol. 21, pp. 273–290, 3 Sep. 2016, ISSN: 1082989X. DOI: 10.1037/MET0000079.

[125] J. Fox and G. Monette, "Generalized collinearity diagnostics", *Journal of the American Statistical Association*, vol. 87, pp. 178–183, 417 1992, ISSN: 1537274X. DOI: 10.1080/01621459.1992.10475190.

[126] M. Cassotti and F. Grisoni, "Variable selection methods: An introduction", Milano Chemometrics and QSAR Research Group, Sep. 2012. DOI: 10.13140/RG.2.1.1009.5129.

[127] D. Wang, W. Zhang, and A. Bakhai, "Comparison of bayesian model averaging and stepwise methods for model selection in logistic regression", *Statistics in Medicine*, vol. 23, pp. 3451–3467, 22 Nov. 2004, ISSN: 1097-0258. DOI: 10.1002/SIM.1930. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.1930.

[128] G. P. Nyaupane, J. T. McCabe, and K. L. Andereck, "Seniors' travel constraints: Stepwise logistic regression analysis", *Tourism Analysis*, vol. 13, pp. 341–354, 4 2008.

[129] Y. H. Chou, C. M. Tiu, G. S. Hung, S. C. Wu, T. Y. Chang, and H. K. Chiang, "Stepwise logistic regression analysis of tumor contour features for breast ultrasound diagnosis", *Ultrasound in Medicine Biology*, vol. 27, pp. 1493–1498, 11 Nov. 2001, ISSN: 0301-5629. DOI: 10.1016/S0301-5629(01)00466-5.

[130] H.-J. Hwang, M. Hahne, K.-R. Müller, T. Obuchi, and Y. Kabashima, "Cross validation in lasso and its acceleration", *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2016, p. 053 304, 5 May 2016, ISSN: 1742-5468. DOI: 10.1088/1742-5468/2016/05/053304. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-5468/2016/05/053304.

[131] "Multi-parametric mri-based radiomics signature for discriminating between clinically significant and insignificant prostate cancer: Cross-validation of a machine learning method", *European Journal of Radiology*, vol. 115, pp. 16–21, Jun. 2019, ISSN: 0720-048X. DOI: 10.1016/J.EJRAD.2019.03.010.

[132] L. E. Melkumova and S. Y. Shatskikh, "Comparing ridge and lasso estimators for data analysis", *Procedia Engineering*, vol. 201, pp. 746–755, Jan. 2017, ISSN: 1877-7058. DOI: 10.1016/J.PROENG.2017.09.615.

[133] J. M. Pereira, M. Basto, and A. F. da Silva, "The logistic lasso and ridge regression in predicting corporate failure", *Procedia Economics and Finance*, vol. 39, pp. 634–641, Jan. 2016, ISSN: 2212-5671. DOI: 10.1016/S2212-5671(16)30310-0.

[134] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor, "Exact post-selection inference, with application to the lasso", *The Annals of Statistics*, vol. 44, pp. 907–927, 3 2016. DOI: 10.1214/15-AOS1371.

[135] S. M. Kim, Y. Kim, K. Jeong, H. Jeong, and J. Kim, "Logistic lasso regression for the diagnosis of breast cancer using clinical demographic data and the bi-rads lexicon for ultrasonography", *Ultrasonography*, vol. 37, p. 36, 1 Jan. 2018, ISSN: 22885943. DOI: 10.14366/USG.16045. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5769953/.

[136] A. J. McEligot, V. Poynor, R. Sharma, and A. Panangadan, "Logistic lasso regression for dietary intakes and breast cancer", *Nutrients*, vol. 12, p. 2652, 9 Aug. 2020, ISSN: 2072-6643. DOI: 10.3390/NU12092652. [Online]. Available: https://www.mdpi.com/2072-6643/12/9/2652.

[137] U. C. Bureau, *American community survey 5-year estimates 2017-2021*, 2022. [Online]. Available: https://data.census.gov/table?q=acs&y=2021&tid=ACSST5Y2021.S0101.

[138] R. K. Guha, "Driving to work: The chicago region and its transport based emissions", The University of Chicago, Apr. 2021.

[139] City of Chicago, *Transportation network providers - trips*, Dec. 2022. [Online]. Available: https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p/.

[140] ——, *Ground transportation tax (7595, 7595ez, 7595us)*, 2022. [Online]. Available: https://www.chicago.gov/city/en/depts/fin/supp_info/revenue/tax_list/ground_transportationtax.html.

[141] ——, *City of chicago congestion pricing*, 2022. [Online]. Available: https://www.chicago.gov/city/en/depts/bacp/supp_info/city_of_chicago_congestion_pricing.html.

[142] ——, *Covid-19 daily rolling average cases, deaths, and hospitalizations*, Dec. 2022. [Online]. Available: https://data.cityofchicago.org/Health-Human-Services/COVID-19-Daily-Rolling-Average-Case-Death-and-Hosp/e68t-c7fv.

[143] ——, *Covid-19 orders*, 2022. [Online]. Available: https://www.chicago.gov/city/en/sites/covid-19/home/health-orders.html.

[144] State of Illinois, *Executive orders related to covid-19*, 2022. [Online]. Available: https://coronavirus.illinois.gov/resources/executive-orders.html.

[145] A. Lee, *Uber and lyft suspend pool rides in us and canada*, Mar. 2020. [Online]. Available: https://edition.cnn.com/2020/03/17/business/coronavirus-uber-pool-trnd/index.html.

[146] Lyft, *Lyft revamps shared rides with new experience for riders and drivers*, Jul. 2021. [Online]. Available: https://www.lyft.com/blog/posts/lyft-revamps-shared-rides-with-new-experience-for-riders-and-drivers.

[147] B. Schulz, *Is uber share available? ride sharing option returns in select cities*, Jun. 2022. [Online]. Available: https://eu.usatoday.com/story/money/2022/06/21/uber-share-returns-save-costs/7689979001/.

[148] City of Chicago, *Boundaries - zoning districts (current)*, 2022. [Online]. Available: https://data.cityofchicago.org/Community-Economic-Development/Boundaries-Zoning-Districts-current-/7cve-jgbp.

[149] ——, *Chicago zoning ordinance*, 2007. [Online]. Available: https://codelibrary.amlegal.com/codes/chicago/latest/chicagozoning_il/0-0-0-48006.

[150] ——, *Boundaries - census tracts - 2010*, 2010. [Online]. Available: https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Census-Tracts-2010/5jrd-6zik.

[151] ——, *Cta - system information - list of 'l' stops - map*, 2022. [Online]. Available: https://data.cityofchicago.org/Transportation/CTA-System-Information-List-of-L-Stops-Map/zbnc-zirh.

[152] ——, *Cta bus stops*, 2022. [Online]. Available: https://data.cityofchicago.org/Transportation/CTA-Bus-Stops/hvnx-qtky.

[153] ——, *Metra stations*, 2022. [Online]. Available: https://data.cityofchicago.org/Transportation/Metra-Stations/nqm8-q2ym.

[154] ——, *Crimes - 2001 to present*, 2022. [Online]. Available: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2.

[155] Weather Underground, *Chicago, il weather history*, 2022. [Online]. Available: https://www.wunderground.com/history/monthly/us/il/chicago/KMDW/date/2019-12.

[156] City of Chicago, *City holidays (offices closed)*, 2022. [Online]. Available: https://www.chicago.gov/city/en/narr/misc/city-holidays.html.

[157] Holiday Calendar, *Holidays and school vacation*, 2022. [Online]. Available: https://holidaycalendar.com/en/month/November/2022/United+States/Illinois/Chicago+public+schools.

[158] City of Chicago, *Covid-19 daily vaccinations - chicago residents - cumulative doses by day*, 2022. [Online]. Available: https://data.cityofchicago.org/Health-Human-Services/COVID-19-Daily-Vaccinations-Chicago-Residents-Cumu/rna5-2pgy.

[159] J. Faber and L. M. Fonseca, "How sample size influences research outcomes", *Dental Press Journal of Orthodontics*, vol. 19, p. 27, 4 Jul. 2014, ISSN: 21776709. DOI: 10.1590/2176-9451.19.4.027-029.EBO. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4296634/.

[160] N. Mantel, "Why stepdown procedures in variable selection", *Technometrics*, vol. 12, pp. 621–625, 3 1970, ISSN: 15372723. DOI: 10.1080/00401706.1970.10488701.

[161] F. E. Harrell, *Regression Modeling Strategies*. Springer New York, 2001, ISBN: 978-1-4419-2918-1. DOI: 10.1007/978-1-4757-3462-1. [Online]. Available: http://link.springer.com/10.1007/978-1-4757-3462-1.

[162] J. Zhao, W. Deng, and Y. Song, "Ridership and effectiveness of bikesharing: The effects of urban features and system characteristics on daily use and turnover rate of public bikes in china", *Transport Policy*, vol. 35, pp. 253–264, Sep. 2014, ISSN: 0967-070X. DOI: 10.1016/J.TRANPOL.2014.06.008.

[163] D. D. Rodas, "Identification of spatio-temporal factors affecting arrivals and departures of shared vehicles", Technical University of Munich, Oct. 2017.

[164] S. Menard, *Applied Logistic Regression Analysis*, 2nd ed. SAGE Publications, Incorporated, 2001, vol. 106, ISBN: 9781544332581. [Online]. Available: https://us.sagepub.com/en-us/nam/applied-logistic-regression-analysis/book11277.

[165] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, 1st ed. Springer New York, 2013, ISBN: 978-1-4614-7138-7. DOI: 10.1007/978-1-4614-7138-7. [Online]. Available: https://link.springer.com/book/10.1007/978-1-4614-7138-7.

[166] J. I. Daoud, "Multicollinearity and regression analysis", *Journal of Physics: Conference Series*, vol. 949, 1 Dec. 2017, ISSN: 1742-6596. DOI: 10.1088/1742-6596/949/1/012009. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/949/1/012009.

[167] D. McFadden, *Quantitative methods for analyzing travel behaviour on individuals: Some recent developments*, D. A. Hensher and P. R. Stopher, Eds., 1978.

[168] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment", *Journal of Thoracic Oncology*, vol. 5, pp. 1315–1316, 9 Sep. 2010, ISSN: 1556-0864. DOI: 10.1097/JTO.0B013E3181EC173D.

[169] G. Choueiry, *Understand forward and backward stepwise regression – quantifying health*, 2023. [Online]. Available: https://quantifyinghealth.com/stepwise-selection/.

[170] G. Smith, "Step away from stepwise", *Journal of Big Data*, vol. 5, pp. 1–12, 1 Dec. 2018, ISSN: 21961115. DOI: 10.1186/S40537-018-0143-6/FIGURES/1. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0143-6.

[171] S. I. Altelbany, "Evaluation of ridge, elastic net and lasso regression methods in precedence of multicollinearity problem: A simulation study", *Journal of Applied Economics and Business Studies*, vol. 5, pp. 131–142, 1 Mar. 2021, ISSN: 2523-2614. DOI: 10.34260/jaebs.517. [Online]. Available: https://pepri.edu.pk/jaebs/index.php/files/article/view/232.

[172] N. Herawati, K. Nisa, and Nusyirwan, "Selecting the method to overcome partial and full multicollinearity in binary logistic model", *International Journal of Statistics and Applications*, vol. 10, pp. 55–59, 3 2020. DOI: 10.5923/j.statistics.20201003.01.

[173] K. Lounici, M. Pontil, S. V. D. Geer, and A. B. Tsybakov, "Oracle inequalities and optimal inference under group sparsity", *The Annals of Statistics*, vol. 39, pp. 2164–2204, 4 Aug. 2011, ISSN: 0090-5364. DOI: 10.1214/11-AOS896.