Imperial College London

Centre for Transport Studies
Department of Civil and Environmental Engineering

Understanding the costs of urban transportation using causal inference methods

Anupriya

Submitted in part fulfilment of the requirements
for the degree of Doctor of Philosophy at
Imperial College London

March 2021

I hereby declare that this thesis and the work reported herein was composed by and originated entirely from me. Information derived from the published and unpublished work of others has been acknowledged in the text and references are given in the list of sources.

Anupriya (2021)

Abstract

With urbanisation on the rise, the need to transport the population within cities in an efficient, safe and sustainable manner has increased tremendously. In serving the growing demand for urban travel, one of the key policy question for decision makers is whether to invest more in road infrastructure or in public transportation. As both of these solutions require substantial spending of public money, understanding their costs continues to be a major area of research. This thesis aims to improve our understanding of the technology underlying costs of operation of public and private modes of urban travel and provide new empirical insights using large-scale datasets and application of causal econometric modelling techniques. The thesis provides empirical and theoretical contributions to three different strands in the transportation literature.

Firstly, by assessing the relative costs of a group of twenty-four metro systems across the world over the period 2004 to 2016, this thesis models the cost structure of these metros and quantifies the important external sources of cost-efficiency. The main methodological development is to control for confounding from observed and unobserved characteristics of metro operations by application of dynamic panel data methods.

Secondly, the thesis provides a quantification of the travel efficiency arising from increasing the provision of road-based urban travel. A crucial pre-condition of this analysis is a reliable characterisation of the technology describing congestion in a road network. In pursuit of this goal, this study develops novel causal econometric models describing vehicular flow-density relationship, both for a highway section and for an urban network, using large-scale traffic detector data and application of non-parametric instrumental variables estimation. Our model is unique as we control for bias from unobserved confounding, for instance, differences in driving behaviour. As an important intermediate research outcome, this thesis also provides a detailed association of the economic theory underlying the link between the flow-density relationship and the corresponding production function for travel in a highway section and in an urban road network.

Finally, the influence of density economies in metros is investigated further using large-scale smart card and train location data from the Mass Transit Railway network in Hong Kong. This thesis delivers novel station-based causal econometric models to understand

how passenger congestion delays arise in metro networks at higher passenger densities. The model is aimed at providing metro operators with a tool to predict the likely occurrences of a problem in the network well in advance and materialise appropriate control measures to minimise the impact of delays and improve the overall system reliability.

The empirical results from this thesis have important implications for appraisal of transportation investment projects.

Acknowledgements

Throughout my PhD journey, I have received a great deal of support and assistance. I would first like to thank my supervisor, Professor Daniel J. Graham, whose expertise was invaluable in formulating the research topic and methodology in particular. His knowledge into the subject matter steered me through this research. A special thanks to my co-supervisor, Dr. Prateek Bansal, whose insights, encouragement and support allowed my studies to go the extra mile. I would also like to thank my other co-supervisors, Dr. Daniel Hörcher and Dr. Jose M. Carbo, whose guidance and suggestions were a real boost to my work at different stages of my PhD.

Further, I am grateful to the Transport Strategy Centre (TSC) at Imperial College London for the financial support provided during the doctoral studies. Thank you for all of the opportunities I was given to present my research at various conferences and for the excellent cooperation in completion of my research. Particular thanks goes to Mr. Richard Anderson and Mr. Alexander Barron from the TSC for their valuable insights to enhance the practical applicability of my research findings. I am also thankful to the MTR Corporation in Hong Kong and to Dr. Lukas Ambuhl from ETH Zurich for providing the necessary data to carry out this research work. I would also like to express my gratitude to the administrative support provided by Alexandra Williams, Fionnuala Ni Dhonnabhain and Sarah Willis from the Department of Civil and Environmental Engineering, who have helped substantially in the organisation of the studies. I acknowledge my fellow PhD students and the postdocs: Ana, Csaba, Farah, Joris, Keita, Laila, Liang, Nan, Praj, Ramandeep, Saeed, Samira and Shane, and other researchers within the Centre for Transport Studies, who have supported me for the past four years of study.

In addition, I would like to thank my parents, my brother, my sister-in-law and other family members for their wise counsel and sympathetic ear. You are always there for me. I am also grateful to my accommodation managers, Laura and Doreen, for providing me with a conducive environment to continue with my research amid the COVID-19 crisis. Finally, there are my friends: Abhilash, Abhineet, Aditya, Amit, Ankita, Anubhav, Anup, Ashish, Deepa, Harsha, Harshit, Kamna, Prateek, Sharad, Shubhechyya, Soumya, Sourish, Robyn, Rohit, Yaruq, and many others who were of great support in deliberating over my

problems and findings, as well as providing a happy distraction to rest my mind outside of my research.

Dedicated to my parents
Binod Kumar and Shobha Gupta.

Excellence happens not by accident. It is a process.

- APJ Abdul Kalam

Contents

List of Tables

List of Figures

Nomenclature

2SLS Two-stage Least Squares

3SLS Three-stage Least Squares

AR Auto Regressive

Caltrans California Department of Transportation

CEE/FSU Central and Eastern Europe and Former Soviet Union

CEEC Central and Eastern European Countries

CoBA Cost Benefit Analysis

CoMet Community of Metros

CRE Correlated Random Effects

DEA Data Envelopment Analysis

DGP Data Generating Process

DP Dirichlet Process

DPGMM Dynamic Panel Generalised Method of Moments

DPM Dirichlet Process Mixture

EU European Union

FD Fundamental Diagram

FE Fixed Effects

FR Fundamental Relationship

GMM Generalised Method of Moments

HCM Highway Capacity Manual

I- Interstate

IV Instrumental Variables

MCMC Markov Chain Monte Carlo

MFD Macroscopic Fundamental Diagram

ML Maximum Likelihood

MTR Mass Transit Railway

NPIV Non-parametric Instrumental Variables

NP Non-parametric

OLS Ordinary Least Squares

pax passengers

PeMS Performance Measurement System

POLS Pooled Ordinary Least Squares

PPP Purchasing Power Parity

RE Random Effects

RMSE Root Mean Squared Error

RTD Returns to Density

RTS Returns to Scale

SFA Stochastic Frontier Analysis

SR State Route

TfL Transport for London

TFP Total Factor Productivity

TSC Transport Strategy Centre

US United States

VHT Vehicle Kilometres Travelled

VKT Vehicle Hours Travelled

WEC Western European Countries

Chapter 1

Introduction

1.1 Background

By 2050, nearly sixty-eight percent of the world's population will be living in cities, that is an increase of two and a half billion people on the present urban population (UN-DESA 2018). Cities are central to economic development as they concentrate the majority of economic activities and output (Henderson 2010), thus contributing more than 80 percent of the world's GDP (The World Bank 2020). However, as a result of this concentration, they generate an unprecedented amount of urban travel demand.

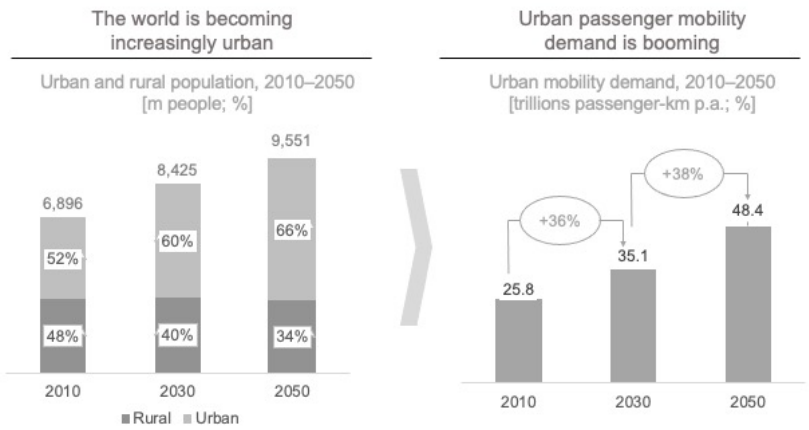
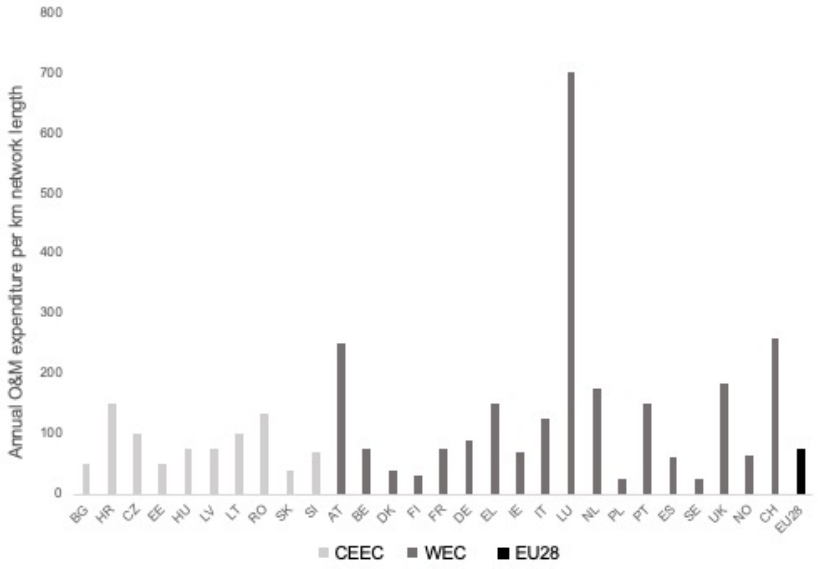


Figure 1.1: Urbanisation and mobility (adapted from Arthur D. Little 2018).

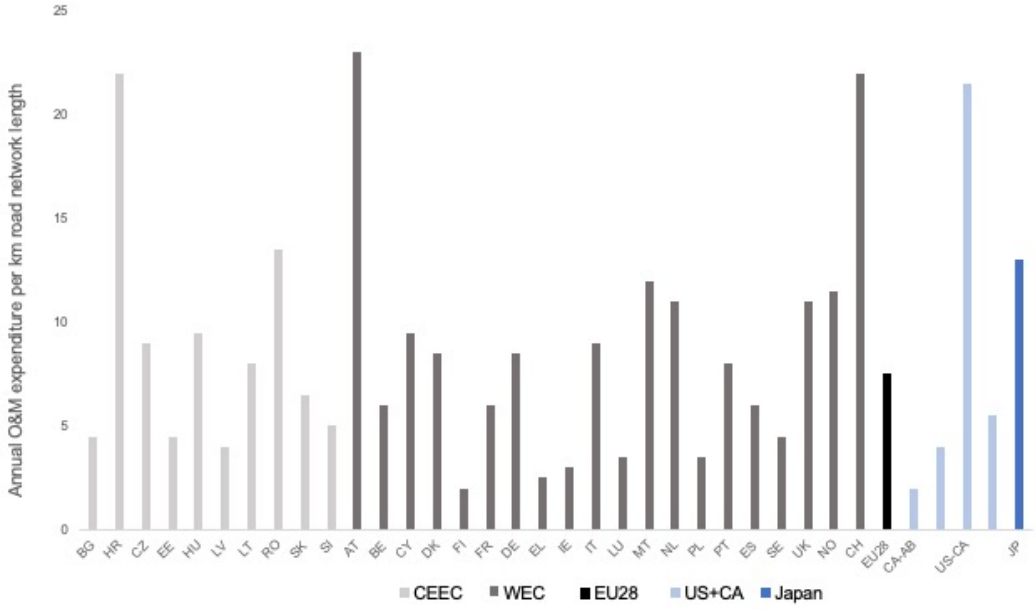
Rapid urbanisation thus calls for transport and mobility solutions that can fulfil the need for transporting the population within cities in an efficient, safe and sustainable manner. Experts suggest that solutions must be based on a balanced mix of public and private transport governed by the characteristics of each city (Litman 2017, ITF 2015, TRB 2017, Hoornweg & Freire 2013).

In the face of growing urban travel demand, a key policy question for decision makers is whether to invest more in increasing the provision of road-based transportation services or rail-based transportation services (Pojani & Stead 2015, Mohan 2008). Not only do these transportation solutions involve huge initial expenditure to start their operations, but their operations and maintenance are also very cost-intensive and require substantial spending of public money every year. For instance, Figure 1.2 illustrates the average annualised operations and maintenance expenditures per unit network length in the period 1995-2016 for rail and road networks in Europe (adapted from European Commission 2019). Thus, understanding the operational cost-efficiency of these modes of urban transportation continues to be a major area of research. Pricing, subsidies and taxation are similarly pressing policy issues where research on operational costs are applied extensively (see, for instance, Hörcher et al. 2020, Small 2013, De Borger et al. 2005, Arnott & Yan 2000, Verhoef et al. 1996).

The research described in this PhD centers around the operational costs of urban transportation: it analyses the causal mechanism driving the costs of operation of rail-based public and road-based private modes of travel in cities using large-scale data, and it quantifies the important determinants of cost-efficiency of these modes. Empirical studies to estimate cost functions for different modes of transportation have been undertaken in the past (refer to Basso et al. 2011, for a review). For instance, Savage (1997) estimates an operational cost function for urban rail transit systems in the US; Akbar & Duranton (2017) estimate a time cost function for road-based vehicular travel in Bogotá. However, most of these studies have been severely limited by the application of simplified statistical



(a) per track-km of rail network length (×1,000 €/km, PPP adjusted)



(b) per kilometre road network length (€/km, PPP adjusted)

Figure 1.2: Average annual operations and maintenance expenditures per unit network length in the period 1995-2016 (adapted from European Commission 2019).

methods. These simple methods when used with observational datasets, fail to control for confounding from various observed and unobserved characteristics of transport operations. This lack of control may result in biased quantification of the sources of cost-efficiency of these modes of travel. Moreover, the findings of such studies are also limited by the scale of data used in the analyses.

The growing availability of large scale datasets and continued advancements in econometric methodologies allows us to advance the existing literature and substantially improve the comprehension of the operational costs of rail-based public and road-based private modes of urban transportation, which is the main focus of this PhD. We use advanced parametric and non-parametric causal econometric techniques that can exploit the large volume of information in various sources of big data, such as large-scale highway and urban road traffic count data from multiple regions and automated fare collection data on urban rail transit systems, to produce detailed findings on the determinants of cost-efficiency of the aforementioned modes of urban transportation.

1.2 Aims and Objectives

This thesis aims to provide new empirical insights on the operational costs of rail-based public and road-based private modes of urban transportation using large-scale datasets and application of causal econometric modelling techniques. This thesis focuses on three broad strands in the transportation literature. The key objectives of the thesis, and associated research questions area as follows:

1. Understand the operational costs of urban rail transport (or metro) systems.
    (a) Why does average cost (that is, cost per unit of output) of operations of metro systems vary considerably across different metro systems?
    (b) What are the main technological determinants of the cost-efficiency of metro operations?

2. Quantify the production of vehicular travel in urban road networks.

   (a) What is the main causal mechanism underlying the production of vehicular travel in a road network? What is the economic theory underlying this production process: how is supply and demand defined?

   (b) What are the returns to scale associated with increasing the provision of vehicular travel in urban road networks?

3. Determine the mechanism driving congestion in near capacity metro operations.

   (a) Does high levels of passenger boardings and alightings (passenger movements, hereafter) at stations give rise to passenger congestion delays in the metro network?

   (b) How do we identify these stations that behave as active bottlenecks in the network? What is the optimum level of passenger movements at the bottleneck station above which such delays arise?

Following the above listed objectives, the research presented in the thesis can be summarised in three parts:

1. Research area 1:

   - We assess the relative short-run operational costs of a group of twenty-four urban metro systems across the world over the period 2004 to 2016 to determine the technology driving differences in unit costs of operations across these systems. Covariates in the model include: output of the metro firm in car-kilometres, labour costs, electricity costs, residual costs and network size.
   - We apply dynamic panel data methods to control for confounding from observed and unobserved time-invariant and time-variant characteristics of metro operations. An example of such a confounder is the managerial efficiency of a metro

firm, which is unobserved but important in determining the cost-efficiency of the firm.

- We deliver estimates of important external sources of cost-efficiency in metro operations, that are,: (i) returns to scale and (ii) returns to density.

2. Research area 2:

   - First of all, we critically review the technology underlying the production of vehicular travel in urban road networks.
     - We explore the link between the economic modelling of production and costs of travel in a road section and the technology used to describe congestion in the road section in economics, also known as the fundamental relationship of traffic flow in the transportation engineering literature.
     - We investigate the economics of demand and supply of travel in a road section.
   - As an important precursor to the empirical analysis of the production technology, we revisit the fundamental diagram of traffic flow for a highway section.
     - We adopt a causal econometric approach to obtain an unbiased estimate of the fundamental flow-density relationship for a highway section using traffic detector data. In particular, we apply a Bayesian non-parametric spline-based regression approach with instrumental variables to control for confounding from omitted variables such as driving behaviour and weather.
     - We validate the proposed approach by estimating the flow-density relationship for three highway bottlenecks in the United States.
     - We emphasise that our causal approach is based on the physical laws that drive the movement of vehicles in a traffic stream as opposed to a demand-supply framework adopted in the economics literature. By doing

so, we also aim to conciliate the engineering and economics approaches to this empirical problem.

  - We re-examine various questions like the existence of hypercongestion, that is, the backward bending part of the fundamental diagram of traffic flow for a highway section, and existence of capacity-drop in a highway bottleneck.

- We extend the above causal framework to estimate the macroscopic fundamental relationship of traffic flow in an urban road network.
  - We use comprehensive data comprising traffic consisting of billions of observations of the three fundamental traffic variables: speed, flow and density, collected from stationary traffic sensors for forty cities around the globe.
  - For each city, we identify approximately homogeneously congested regions (or reservoirs) in the network. We estimate the macroscopic flow-density relationship using the Bayesian non-parametric instrumental variables approach for each reservoir.
- From the estimated macroscopic fundamental relationships, we quantify the scale economies resulting from increasing the provision of vehicular transportation in urban road networks.

3. Research area 3:

- We further investigate a finding from the first study that reveals the density of metro operations as a key influence in determining their cost-efficiency.
  - We quantify how metro performance under high frequency of operations varies with increasing passenger numbers in the network. In particular, we investigate how the vicious circle of passenger congestion and train delays, that can impact the punctuality of high frequency metro operations.

  - To do so, we conduct the first station-level econometric analysis to estimate a causal relationship between boarding-alighting movements and train flow using data from entry/exit gates and train movement data of the Mass Transit Railway, Hong Kong.
  - We thus explore the existence of traffic fundamental diagram like relationships in a metro network.

- We adopt a Bayesian non-parametric spline-based regression approach and apply instrumental variables estimation to control for confounding bias that may occur due to unobserved characteristics of metro operations.
- Based on the estimated relationships, we identify potential bottleneck stations in the network. At these bottlenecks, an increased amount of passenger boardings and alightings may lead to increased and inconsistent dwell times of trains, which may eventually cause disruption in service frequencies due to queuing of trains upstream of these stations.
- We deliver novel estimates of optimal passenger movements at these stations, which when used along with real data on daily demand, could be instrumental for metro operators to develop informed station-level control strategies.

1.3 Contributions

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

1. **Analysis of unique and new sources of data** - For the analysis of metro operational costs, we use a unique and very high quality panel dataset that relates to twenty-four metro operators across the world, collected by the Transport Strategy Centre (TSC) at Imperial College London over the period of 2004-2016. Previous empirical studies on metro costs, although very few in number, have used country specific data. For analysis of production of travel in an urban road network, we

again use a unique and comprehensive dataset consisting of billions of observations on traffic states for forty cities around the globe. To our knowledge, this data has not been used in previous empirical studies and it will enable a deeper level of analysis of the technical efficiency of urban road networks than previously possible.

2. **Application of advanced econometric methods** - Simple statistical models have been used to analyse operational costs of various modes of travel in the past, however, the application of such methods often fail to control for confounding from various key characteristics of transport operations, which may remain unobserved. To adjust for these potential sources of bias, we use advanced econometric methods that have not been previously applied in this field of research. Wherever applicable, we adopt non-parametric specifications to enable the investigation of more complex, non-linear relationships that have been ignored in past studies for analytical simplification.

3. **New research questions** - Our research contributes in developing the understanding of various new research questions such as exploring the existence of traffic fundamental diagram like relationships in a metro network, developing a comprehension of the fundamental diagram of traffic flow in a causal inference framework, and delivering new causal estimates of returns to scale in increasing the provision of vehicular travel in urban road networks.

4. **Integrated research approach** - Although this thesis combines three different areas of transportation research, we ensure comparability between findings from these different areas. For instance, the third research area further investigates a finding from the first study that reveals the density of metro operations to be a key determinant of their cost-efficiency. It extends the idea of traffic fundamental diagrams from the second research area to the context of metro operations and explores the impact of increased passenger density in a metro network on its performance. This integrated research approach enables a more holistic understanding of

operational costs of the two major modes of urban travel that are focused upon in this research.

1.4 Thesis Outline

The remainder of the thesis comprises six chapters. The thesis begins with an overview of the transport cost function literature in Chapter 2 relating to the research areas addressed in Chapters 3, 4, and 5 of the thesis. The research relating to congestion in near capacity transport operations is reviewed separately in Chapter 6, as a systematic exposition of congestion delays in metro networks is presented by reviewing the literature directly prior to the analysis. Chapters 3 to 6 present the main analytical results of the thesis. A detailed outline of the thesis by chapter is summarised in the following paragraphs.

Chapter 2 is organised into three parts, beginning with a general definition of a transport cost function and important concepts related to it. This is followed by a review of the economics literature on cost functions for road-based vehicular travel. The first two parts of this review chapter are supplemented by detailed contextual reviews of the relevant literature in Chapters 3, 4, and 5. The final part includes a methodological overview wherein we first discuss the methodological challenges related to endogeneity. Thereafter, the state-of-the-art econometric approaches approaches to treatment of endogeneity both in parametric and non-parametric models are reviewed.

Chapter 3 focuses on the operational costs of metro systems in the short-run. The chapter first reviews the relevant economics literature on public transit cost functions. An empirical model to describe the short-run variable cost function for metro operations is constructed using a unique and high-quality panel dataset on twenty-four metro systems around the world collected over the period 2004-2016 and applying a dynamic panel generalised method of moments (DPGMM) estimation. The scale economy estimates,(i) returns to network size and (ii) returns to density, are then derived from the estimated cost function and an application of these estimates to understand the variation in unit

operational costs across metro systems is presented.

In Chapter 4, a causal econometric framework to estimate the fundamental relationship of traffic flow for a highway section is proposed and applied on data from three highway bottlenecks in the US. The chapter commences with a detailed review of the relevant engineering on empirical fundamental diagram and identifies the limitations of the existing approach. This review is followed by an econometric analysis driven by traffic physics that underpins the movement of vehicles in a traffic stream. The estimated results are then discussed in detail and compared with relevant findings from engineering and economics literature.

In Chapter 5, an extended application of the causal econometric framework from Chapter 4 is presented. The focus of the analysis is to estimate macroscopic level fundamental flow-density relationships for urban networks to understand the production of travel in these networks. This chapter first reviews the engineering literature on macroscopic fundamental diagrams and a related economics literature on estimation of production and cost functions for vehicular travel in an urban road network. A unique large-scale data consisting of traffic state measurements from forty urban road networks around the globe recorded over multiple days is then used for the econometric analysis. Within each network, homogeneously congested reservoirs are identified and reservoir-level aggregated flow-density relationships are estimated. From the estimated relationships, the returns to scale in the provision of vehicular travel are derived.

In Chapter 6, a study on the mechanism driving congestion and delays in high-frequency metro operations is undertaken. The chapter begins with a review of the relevant literature, followed by a microsimulation study to explain this phenomenon. Thereafter, the idea of traffic fundamental diagram from Chapters 4 and 5 is extended to the context of metro operations and station-level causal relationships between boarding-alighting movements and train flow are estimated using data from entry/exit gates and train movement data of the Mass Transit Railway, Hong Kong. Furthermore, potential applications of the

estimated relationships in devising station-level control measures are discussed.

In the final chapter of the thesis, Chapter 7, a summary of the conclusions from the analysis chapters is presented, along with recommendations for potential future research.

1.5 List of Publications

1. Anupriya, Graham, D.J., Carbo, J.M., Anderson, R.J., & Bansal, P. (2020). Understanding the costs of urban rail transport operations. *Transportation Research Part B: Methodological*, Vol. 138, pp. 292-316.

2. Anupriya, Graham, D.J., Bansal, P., Hörcher, D. & Anderson, R.J. (under revision). Congestion in near capacity metro operations: optimum boardings and alightings at bottleneck stations. *in Transportation Research Part C: Emerging Technologies.*

3. Anupriya, Graham, D.J., Hörcher, D. & Bansal, P. (under review). Revisiting the empirical fundamental relationship of traffic flow for highways using a causal econometric approach.

4. Anupriya, Graham, D.J., & Bansal, P. (under review). Understanding the production of vehicular travel in cities.

5. Anupriya, Graham, D.J., Bansal, P. & Hörcher, D. (under development). A review of engineering and economic approaches to empirical fundamental diagram of traffic flow.

Additional work undertaken during PhD:

1. Anupriya, Graham, D.J., Hörcher, D., Anderson, R.J., & Bansal, P. (2020). Quantifying the ex-post causal impact of differential pricing on commuter trip scheduling in Hong Kong. *Transportation Research Part A: Policy and Practice*, Vol. 141, pp. 16-34.

2. Carbo, J.M., Graham, D.J., Anupriya, Casas, D., & Melo, P.C. (2019). Evaluating the causal economic impacts of transport investments: evidence from the Madrid–Barcelona high speed rail corridor. *Journal of Applied Statistics*, Vol. 46(9), pp. 1714-1723.

3. Gutierrez-Rave, J.P., Graham, D.J., Bansal, P., & Anupriya (under development). Analysis of urban metro demand: evidence from slope heterogeneity models.

4. Mateo, E.M., Graham, D.J., Bansal, P., & Anupriya (under development). Quantification of non-linear effects in agglomeration economies for transport appraisals.

References

Akbar, P. & Duranton, G. (2017), 'Measuring the cost of congestion in highly congested city: Bogotá'.

Arnott, R. & Yan, A. (2000), 'The two-mode problem: Second-best pricing and capacity', *Review of urban & regional development studies* **12**(3), 170–199.

Arthur D. Little (2018), The future of mobility 3.0: Reinventing mobility in the era of disruption and creativity, Technical report, Arthur D. Little and UTIP.
URL: *http://tinyurl.com/2v9w2cst*

Basso, L. J., Jara-Díaz, S. R. & II, W. G. W. (2011), Cost functions for transport firms, *in* A. de Palma, R. Lindsey, E. Quinet & R. Vickerman, eds, 'A Handbook of Transport Economics', Edward Elgar Publishing Ltd., Northampton, chapter 12, pp. 273–297.

De Borger, B., Proost, S. & Van Dender, K. (2005), 'Congestion and tax competition in a parallel network', *European Economic Review* **49**(8), 2013–2040.

European Commission (2019), Overview of transport infrastructure expenditures and costs, Technical report, European Commission.
URL: *https://tinyurl.com/3khosmz6*

Henderson, J. V. (2010), 'Cities and development', *Journal of regional science* **50**(1), 515–540.

Hoornweg, D. & Freire, M. (2013), 'Building sustainability in an urbanizing world: a partnership report', *Urban Development & Resilience Unit, The World Bank Group* .
URL: *http://tinyurl.com/1dnyb6ms*

Hörcher, D., De Borger, B., Seifu, W. & Graham, D. J. (2020), 'Public transport provision under agglomeration economies', *Regional Science and Urban Economics* **81**, 103503.

ITF (2015), 'Urban mobility system upgrade', *OECD Publishing* .

URL: *http://dx.doi.org/10.1787/5jlwvzdk29g5-en*

Litman, T. (2017), *Autonomous vehicle implementation predictions*, Victoria Transport Policy Institute.

URL: *http://tinyurl.com/d2b4yvhc*

Mohan, D. (2008), 'Mythologies, metro rail systems and future urban transport', *Economic and Political Weekly* pp. 41–53.

Pojani, D. & Stead, D. (2015), 'Sustainable urban transport in the developing world: beyond megacities', *Sustainability* **7**(6), 7784–7805.

Savage, I. (1997), 'Scale economies in United States rail transit systems', *Transportation Research Part A: Policy and Practice* **31**(6), 459–473.

Small, K. (2013), *Urban transportation economics*, Vol. 4, Taylor & Francis.

The World Bank (2020), Urbanization reviews, Technical report, The World Bank.

URL: *https://www.worldbank.org/en/topic/urbandevelopment/overview*

TRB (2017), 'Strategies to advance automated and connected vehicles', *Transportation Research Board* .

URL: *http://www.nap.edu/download/24873*

UN-DESA (2018), The 2018 revision of world urbanization prospects, Technical report, Department of Economic and Social Affairs, United Nations.

URL: *https://esa.un.org/unpd/wup*

Verhoef, E., Nijkamp, P. & Rietveld, P. (1996), 'Second-best congestion pricing: the case of an untolled alternative', *Journal of Urban Economics* **40**(3), 279–302.

Chapter 2

Literature Review

2.1 An Introduction to Transport Cost Functions

The empirical analysis of transport costs is indispensable both for transport operators as well as policy makers. Cost functions enable performance comparisons across firms over time and across regulatory regimes and facilitate broad characterisation of the industry by determining the extent of scale economies (Small & Verhoef 2007). The knowledge of cost functions is a key to understand the relative efficiency of various modes of transportation and the relative importance of various factors of production such as infrastructure, operator and staff wages, and even externalities (spillovers to non-users). Cost studies thus have enormous applications ranging from input/output analysis and guiding investments to supporting decisions on pricing rules and organisation of market structure (Borts 1960, Viton 1981). The literature concerning the empirical analysis of transport costs is huge and several extensive reviews can be found in Jara-Diaz (1982), Oum & Waters (1996) and Basso et al. (2011). In this chapter, we provide an overview of the important theoretical concepts related a transport cost function. This review is supplemented by in-depth reviews in Chapters 3, 4 and 5.

2.1.1 General definition of costs

Theoretically, a cost function, $C(\mathbf{w}, y)$, is a relationship between the minimum cost of producing an output, y, given a production technology, $F(y, \mathbf{x})$, and a vector of prices, $\mathbf{w}$, of the factors of production, $\mathbf{x}$ (Varian 2014). Here, the production technology, $F(y, \mathbf{x})$, describes the boundaries of technical feasibility of a firm for producing the output y using the combination $\mathbf{x}$ of inputs.

The cost function $C(\mathbf{w}, y)$ is derived as a solution to the following optimisation problem:

$$\begin{aligned} &\min C(\mathbf{w}, y) = \mathbf{w}'\mathbf{x} \\ &\text{subject to } F(y, \mathbf{x}) = 0 \end{aligned} \tag{2.1}$$

The resulting cost function of the firm is dual to its production technology – where production function represents technical efficiency, cost function represents cost efficiency (Varian 2014). Arguments of the cost function, that is, factors prices and output, are assumed to be *exogenous* to the firm. The obtained cost function exhibits the following main properties as explained by McFadden (1978):

1. Technical characteristics of the production process like existence of scale economies can be analysed through the cost function.
   Degree of scale economies $(S) = \frac{\text{average cost}}{\text{marginal cost}} = \frac{C(\mathbf{w},y)/y}{\partial C(\mathbf{w},y)/\partial y}$.

2. Partial derivative of the cost function with respect to the i^{th} factor price, w_i, gives the (conditional) demand for factor i, that is, x_i. This property is commonly known as Shepherd's lemma.
   $x_i = \frac{\partial C(\mathbf{w},y)}{\partial w_i}$

3. $C(\mathbf{w}, y)$ is non-decreasing, homogeneous of degree one and concave in w_i.

If one or more inputs are held fixed over the period of production the resulting cost function a *short run cost function*. Typically, the fixed factor of production, represented by $\bar{x}_i$, is one element of capital stock of the firm. For instance, in a transport cost function,

$\bar{x_n}$ is usually a measure of length of the network operated by the transport firm. The short-run cost function is then $C^s(\mathbf{w}, y, \bar{x_n})$, which always consists of a *fixed cost* C^0, $C^0 = \lim_{y \to 0} C^s(\mathbf{w}, y, \bar{x_n})$. The rest of the short-run cost is known as *operating cost*. It is to note that the operating cost may comprise a fixed component that is independent of the output y, for instance, the cost of maintaining the transport infrastructure.

In Chapter 3, we apply these definitions to specify a short-run cost function for metro operations. We also discuss the failure of exogeneity assumption on covariates (for instance, output and factor prices) and its implications on estimation of the cost function.

2.1.2 Definition of outputs

The output, in case of a transport firm, comprises of a large number of spatially and temporally varying services, the production of which involves decisions on route structure and capacity, design of the network, service frequencies, among other factors (Jara-Diaz 1982). In essence, the process involves operationalisation of input combinations to produce the desired service. Ideally, transport output is characterised by a vector of flows $\mathbf{y} = y_{ijmt}$ between many origin-destination pairs ij served by vehicle m in time period t (Basso et al. 2011). However, for statistical modelling, it is necessary to aggregate this huge vector of flows into a traceable measure representing total output. Aggregated measures of output used in the literature vary from intermediate or supply-oriented measures like vehicle-kilometres or vehicle-hours, to final or demand-oriented measures such as passenger-kilometres (Basso et al. 2011, Small & Verhoef 2007).

Small & Verhoef (2007) suggest that the adopted definition of output should depend on the purpose of analysis. If technical efficiency of transport firms' production process is being studied, then intermediate outputs should be used, while a study of the effectiveness of the firms' service should use final outputs (Small & Verhoef 2007). However, De-Borger et al. (2002) argue that passenger-kilometres or related demand-oriented measures are more relevant for an analysis for transport operations because these measures represent

the economic motive of providing transport services. Nonetheless, Small & Verhoef (2007) argue that supply-oriented indicators are the fundamental decision variables for transport operators and thus, are under their control. As Small & Verhoef (2007) rightly point out, this decision rule is implicit in the definition of a cost function because the cost function is a dual representation of a firm's production function, which indeed is an engineering relationship between physical inputs and physical outputs. As a consequence, use of intermediate-output measures is more reasonable.

Consistent with this discussion, we use supply-oriented measures for output in Chapters 3 and 5. In Chapter 3, we use car-kilometres as the aggregated measure of output in our metro cost function specification. In Chapter 5, we use vehicle kilometres travelled (VKT) as output to comprehend the production of vehicular travel on urban road networks.

2.1.3 Scale economies

The degree of scale economies, S, describes how fast costs increase with respect to output (Varian 2014). As mentioned in Section 2.1.1, S is defined as:

$$S = \frac{\text{average cost}}{\text{marginal cost}} = \frac{C(\mathbf{w}, y)/y}{\partial C(\mathbf{w}, y)/\partial y} \tag{2.2}$$

If a proportional increase in output leads to: (i) a less than proportionate increase in cost, that is, $S > 1$, there exists economies of scale; (ii) a directly proportionate increase in cost, that is, $S = 1$, there are neutral scale economies; and; (iii) to a more than proportionate increase in cost, that is, $S < 1$, there exists dis-economies of scale. These three cases are equivalent to increasing, constant and decreasing returns to scale (RTS), which are properties of the production technology underlying the cost function. RTS describes the output response to a proportionate increase in all inputs.

The transportation literature derives two main descriptors of industry behaviour from cost studies: (a) returns to density (RTD) and (b) returns to network size (RTS). RTD describes the effect of increasing the density of output, that is, operating more vehicle

kilometres on a fixed network, while RTS describes the effect of increasing the spatial scale of output, that is , expanding the network to serve new locations (Graham et al. 2003, Graham 2008).

We further review the literature on scale economies in provision of urban metro services and urban vehicular travel in Chapters 3 and 5 respectively.

2.2 Cost Functions for Road Travel

In this section, we first describe how modelling congestion in a given road facility is an integral part of an economist's definition of costs of travel in the facility. We then discuss why the fundamental relationship of traffic flow is of primary importance to transport economists in the exercise of modelling congestion. Further, we review the important theoretical developments from the transport economics literature that relate to this relationship and highlight how these developments have driven recent empirical analyses in the economics literature that question some well-established findings from the engineering literature.

2.2.1 Congestion modelling and cost function

As mentioned in Section 2.1.2, in absolute terms, transport output is a complex vector of flows. Each element of this vector is associated with a service quality dimension which plays a key role in determining the associated demand. Thus, Small & Verhoef (2007) argue that this dimension must be included when defining the aggregate measure of output. Because defining service quality for each output is cumbersome, Small & Verhoef (2007) adopt an approach similar to Becker (1965)'s theory of household production where consumers are viewed as part of the production process. In this approach, user-supplied inputs, such as time, are treated as factors of production. As a consequence, user inputs are moved from the demand side to the supply side of the analysis and get directly

embedded into cost functions.

The economics literature mostly follows this approach to formulate a cost function for road-based vehicular travel (see, for instance, Couture et al. 2018, Small & Verhoef 2007, Johnson 1964, Walters 1961). Thus, modelling the variation in the quality of service (for instance, expected travel time and reliability) of a given road section, or in other words, the congestion technology of the section, over the intensity of its use, is a rudimentary exercise to model the associated cost function for travel.

2.2.2 Importance of the traffic fundamental relationship (FR)

In the economics literature, the congestion technology of a given road section is represented either via stationary-state (static) or dynamic models of traffic flow (see Small & Verhoef 2007, for a review). The simplest model of congestion considers a uniform road section with no physical bottlenecks within the section. One of the key approaches to model congestion in this section is to estimate its fundamental speed-flow relationship. The fundamental relationship (FR) is a standard engineering relationship between two of the three key traffic variables: (i) vehicular flow q, that is, the number of vehicles passing a given point per unit time, (ii) density k, that is, the number of vehicles per unit distance in the road section, and, (iii) average vehicular speed, v. This relationship is defined based on the assumption that traffic conditions along the section are stationary, which means that q, k and v, are the same at each and every point in the section (Daganzo 1997, May 1990). In Chapter 4, we provide an in-depth review of the engineering literature on empirical estimation of the FR.

Transport economists further use this congestion model to formulate a cost function as mentioned previously. In the most basic set-up where all users are assumed to have identical value of time, α, the average cost c (borne entirely by the user) on a road section is defined as:

$$c = [c_{00} + \alpha.T_f] + [\alpha.(T - T_f) + c_s] \tag{2.3}$$

where c_{00} consists of exogenous monetary costs like fuel consumption, $\alpha.T_f$ is the cost of free-flow travel time, $\alpha.(T - T_f)$ represents the cost of delays, and c_s is the scheduling cost. The first two cost components denote the travel cost in absence of congestion and the latter two capture the congestion-related cost.

Transport economists assume that user time is the main input in the production of travel in a road section (Small & Verhoef 2007, Couture et al. 2018, Akbar & Duranton 2017). Thus, users are suppliers in the sense that they supply time to the travel process. Under stationary state or static conditions, traffic flow, q, equals both the rates of inflow and outflow, or in other words, the rates at which trips are started and ended. Thus, q represents the quantity demanded per unit time by users as well as the quantity supplied per unit time (based on the congestion technology) at a given time cost (Small & Verhoef 2007). This time cost equals the inverse of the average speed in the road section (Small & Verhoef 2007, Couture et al. 2018, Akbar & Duranton 2017) as scheduling costs are zero under static conditions. Assuming the free flow travel time to be zero, equation 2.3 reduces to:

$$c(q) = c_{00} + \alpha.T(q) = c_{00} + \alpha.\frac{L}{v(q)} \tag{2.4}$$

Based on equation 2.4, transport economists view the *supply curve* for travel in a road section as a mathematically scaled version of *speed-flow relationship* (Small & Verhoef 2007, Small & Chu 2003, Johnson 1964, Walters 1961). Since users are assumed as price-takers, $c(q)$ represents both average and marginal private cost.

2.2.3 Theoretical developments related to the FR

Walters (1961) suggested that the interaction of the supply curve $c(q)$ with a standard

demand curve $d(q)$ gives the equilibrium traffic state as shown in Figure 2.1. Furthermore, the interaction of the demand curve with the marginal social cost curve, that is derived from the average cost curve (for details, refer to Small & Verhoef 2007), gives the social optimum flow. The optimum flow according to Walters (1961) can be achieved via a Pigovian charge, widely known as a congestion toll.

However, as Figure 2.1 illustrates, there can be multiple candidate equilibria, namely x, y and z. The uniqueness and stability of these equilibria have led to a major debate in the literature (Verhoef 1999, Small & Chu 2003). Existence of a backward bending supply curve implies that the average cost is not single-valued and the theoretical definition of a cost function (that is, the minimum cost of producing a given level of output) does not apply (Small & Verhoef 2007). Several researchers have even questioned the suitability of *traffic flow* as a measure of the quantity demanded or supplied, and instead proposed the use of *traffic density* (Ohta 2001, Hills & Evans 1993, Carey & Else 1985).

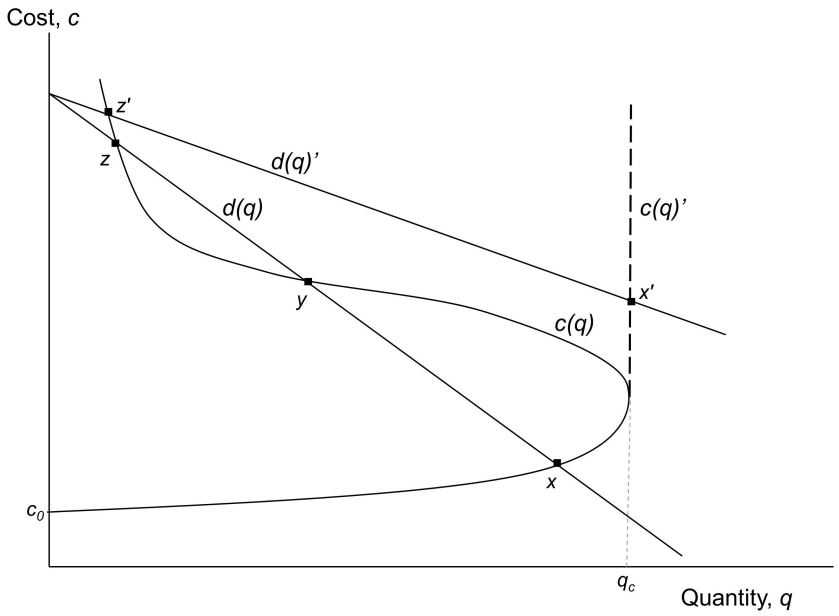


Figure 2.1: The conventional and amended supply curves (adopted from Small & Verhoef 2007).

Conventional stability analysis of the candidate equilibria suggests that x is stable for both price and flow perturbations, whereas y is only stable for flow perturbations and z is

only stable for price perturbations (Newbery 1990). However, a higher demand curve such as $d(q)'$ in Figure 2.1 yields no stable equilibrium. Verhoef (1999, 2001) suggested that the conventional stability analysis considers perturbations in the flow rates both into and along the road section, which is not physically possible. They instead consider perturbations in inflow rate treating flow along the road section as endogenous, a concept which they refer to as dynamic stability analysis. Following this analysis, Verhoef (1999, 2001) find that the entire hypercongested branch of the $c(q)$ curve is dynamically unstable. Thus, they argue that the backward-bending region of the FR is not suitable as a supply curve. They instead propose an amendment based on the car-following theory – when inflow reaches the capacity of the section, the travel supplied by the road section becomes constant instead of decreasing or bending backwards (Verhoef 1999, 2001) (see the amended curve $c(q)'$ in Figure 2.1). A new equilibrium state x' is obtained, which is dynamically stable and involves a maximum flow q_c on the road section and a constant length queue before its entrance (Verhoef 2001). This amendment by Verhoef (2001) suggests that there should be no backward bending or hypercongestion in the speed-flow FR, which is inconsistent with the empirically-established backward bending relationship in the engineering literature.

For a highway section with a physical bottleneck, economists suggest that the existence of hypercongestion only in the queues upstream of the bottleneck (Lindsey & Verhoef 2007, Mun 1999, Small & Chu 2003). These studies argue that any drop in observed capacity within the bottleneck should be related to extraneous influences, not to mechanisms occurring within the bottleneck section itself. However, these studies do not recognise the existence of a *two-capacity* phenomenon, that is, a drop in capacity at the bottleneck upon the onset of queue formation (Cassidy & Bertini 1999, Cassidy & Rudjanakanoknad 2005, Daganzo 2002), a finding that is well-established in the engineering literature. Moreover, a recent study in the economics literature by Anderson & Davis (2020) presents empirical evidence to question the existence of capacity drop.

In an earlier version of their study, Anderson & Davis (2018) derive causal estimates of

changes in outflow from a bottleneck (supply) with changes in length of queue upstream of the bottleneck (demand). Anderson & Davis (2018) argue that the low-speed, low-flow observations that form the backward bending region of the flow-density or flow-speed curve arise due to supply shocks as opposed to excess demand as argued in the engineering literature. This argument is based on the interpretation of the speed-flow FR as a supply curve for travel. Anderson & Davis (2018) suggest that these supply-shocks such as lane closures, accidents, disabled vehicles and weather, among others, represent a *shift* in the supply curve, rather than a movement along the curve. Further, Anderson & Davis (2018) point out that because both demand and supply are shifting, so the observed data on speed and flow represents a locus of all possible equilibria, rather than points on a supply curve. Thus, they use exogenous shifters in demand as an instrument to estimate the underlying supply relationship between outflow and queuing and do not find any evidence of hypercongestion (neither capacity-drop nor backward bending) in highway bottlenecks.

We argue that economists' interpretation of the speed-flow FR as a supply curve for travel is unsuitable for obtaining causal estimates of the FR. This is because the interpretation holds true only under stationary-state traffic conditions which seldom exist particularly under congested conditions.

In Chapters 4 and 5, we develop a causal understanding of the traffic fundamental relationship for a highway bottleneck and for urban road network respectively, within an engineering framework that is consistent with traffic physics.

2.3 Methodological Overview

As mentioned in Chapter 1, the main methodological contribution of this thesis is to develop novel causal models using large-scale data to improve the understanding the operational costs underlying different modes of urban transportation. The empirical estimates in this thesis aim to present credible causal relationships between the covariates and the response variable in the model. This section provides a general overview of the

main methodological challenges, that is, the endogeneity concerns in estimating such causal relationships and reviews the viable methods from the econometrics literature to account for these concerns. The discussion in this section is rather general and brief and is followed by in-depth contextual explanations in the main analytical chapters of this thesis: Chapters 3, 4, 5 and 6.

2.3.1 Methodological challenges

There are two major concerns in relation to endogeneity: (i) omitted variable bias, and, (ii) reverse causality (simultaneity) (Cameron & Trivedi 2005, Wooldridge 2010). Omitted covariates that are correlated with both the dependent variable and the included covariates in a regression may result in inconsistent estimates of model parameters. Reverse causality is a consequence of the existence of a two-way causal relationship or a cause-effect relationship, contrary to the one assumed in the model. The presence of reverse causality may also lead to inconsistent estimates.

In the rest of this sub-section, we further discuss these endogeneity biases in general terms by mathematically demonstrating the two sources of confounding and resulting biases. In Chapters 3, 4, 5 and 6, we undertake a contextual re-discussion of these biases and their potential implications on the estimated model.

Omitted variable bias

To illustrate the endogeneity bias due to omitted covariates, we consider a basic linear regression model with a data generating process given by:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{z}\alpha + \mathbf{u}, \tag{2.5}$$

where $\mathbf{Y}$ is an $N \times 1$ vector of dependent variables, $\mathbf{X}$ and $\mathbf{z}$ are $N \times K$ and $N \times 1$ matrices respectively, β and α are $K \times 1$ and 1×1 vector of parameters and $\mathbf{u}$ is an $N \times 1$ error vector that is assumed to be uncorrelated with $\mathbf{X}$ and $\mathbf{z}$. Application of a standard

regression technique such as an ordinary least squares (OLS) estimation of $\mathbf{Y}$ on $\mathbf{X}$ and $\mathbf{z}$ yields consistent parameter estimates of α and β [1].

Suppose instead that $\mathbf{z}$ is omitted from the equation and $\mathbf{y}$ is regressed on $\mathbf{X}$ alone. Then $\mathbf{z}\alpha$ becomes a part of the error term and the estimated model becomes:

$$\mathbf{Y} = \mathbf{X}\beta + (\mathbf{z}\alpha + \mathbf{u}),$$

where $(\mathbf{z}\alpha + \mathbf{u})$ is the new error term. The OLS estimator of β equals:

$$\begin{aligned}
\beta_{OLS} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{z}\alpha + \mathbf{u}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}\alpha + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\
&= \beta + (N^{-1}\mathbf{X}'\mathbf{X}^{-1})(N^{-1}\mathbf{X}'\mathbf{z})\alpha + (N^{-1}\mathbf{X}'\mathbf{X})^{-1}(N^{-1}\mathbf{X}'\mathbf{u})
\end{aligned}$$

Under the assumption that $\mathbf{X}$ is uncorrelated with $\mathbf{u}$, the final term has probability limit zero. However, because, $\mathbf{X}$ is correlated with $\mathbf{z}$,

$$\text{plim}[\beta_{OLS}] = \beta + \delta\alpha$$

where, $\delta = \text{plim}[(N^{-1}\mathbf{X}'\mathbf{X}^{-1})(N^{-1}\mathbf{X}'\mathbf{z})]$ is the probability limit of the OLS estimator in the regression of the omitted regressor ($\mathbf{z}$) on the included regressors ($\mathbf{X}$). This inconsistency is called omitted variable bias, which exists as long as the omitted variable is correlated with the included regressors, or in other words, $\delta \neq 0$. In general the inconsistency could be positive or negative (Cameron & Trivedi 2005). A positive bias exists if the correlation between $\mathbf{X}$ and $\mathbf{z}$, that is, δ and that between $\mathbf{y}$ and $\mathbf{z}$, that is, α are both either positive or negative, that is, $\alpha\delta > 0$. If the correlations are of opposite sign, that is, $\alpha\delta < 0$, the bias is negative.

[1] Note that an estimator $\hat{\beta}$ is said to be consistent for β if it converges in probability to the true value β, that is, $\text{plim}(\hat{\beta}) \to \beta$.

Reverse Causality

To illustrate bias due to reverse causality, we further simplify the data generating process in equation 2.5 as follows:

$$\mathbf{Y} = \mathbf{X}\beta + \xi, \tag{2.6}$$

To obtain an unbiased estimate of β via OLS, the Gauss Markov condition of zero conditional mean of errors, that is, $E[\xi|X] = 0$, or in other terms, $\text{Cov}[\xi, X] = 0$,must be satisfied. In case of reverse causality, there exists another data generating process given by:

$$\mathbf{X} = \mathbf{Y}\gamma + \psi, \tag{2.7}$$

Consequently, we have,

$$\begin{aligned}
\text{Cov}[\xi, X] &= \text{Cov}[\xi, (Y\gamma + \psi)] \\
&= \gamma\text{Cov}[\xi, Y] \quad \text{assuming that} \quad \xi \perp \psi \\
&= \gamma\text{Cov}[\xi, (X\beta + \xi)] \\
&= \gamma\text{Cov}[\xi, X\beta] + \text{Var}(\xi) \\
&\neq 0
\end{aligned}$$

Thus, the zero conditional mean assumption of errors is violated and OLS may result into a biased estimate of β.

2.3.2 Treating endogeneity

In this sub-section, we briefly review the different econometric approaches for treatment of endogeneity in parametric and non-parametric regression models.

Parametric models

In Chapter 3, we estimate a variant of the following linear unobserved effects model:

$$y_{it} = X_{it}\beta + v_{it}, \quad i = 1, ..., N; \quad t = 1, ..., T. \tag{2.8}$$

where y_{it} represents the response variable for unit i at time t, X_{it} denotes the set of covariates in the model, v_{it} is a composite error term such that $v_{it} = c_i + u_{it}$, in which, c_i is a random variable, widely known as *unobserved effect* or *individual heterogeneity*, which may or may not be correlated with the X_{it} and u_{it} are the idiosyncratic errors.

The econometrics literature commonly uses the following approaches to estimate this equation: (i) pooled ordinary least squares (POLS), (ii) fixed effects (FE), (iii) instrumental variables (IV), and (iv) dynamic panel generalised methods of moments (DPGMM).

In the POLS approach, the observations are pooled across i and t and ordinary least squares (OLS) estimation is applied. Consistency of this estimator requires the contemporaneous exogeneity assumptions: (a) $Cov(X_{it}, c_i) = 0$, and, (b) $Cov(X_{it}, u_{it}) = 0$, for all $t = 1, \ldots, T$ (Wooldridge 2010). As we discuss in Chapter 3, the former assumption can be highly restrictive because the unobserved effect c_i is often correlated with either X_{it} or u_{it}.

The FE approach offers some degree of treatment for unobserved effects or the fixed time-invariant component of unit-specific heterogeneity, c_i. FE estimates are obtained by applying OLS estimation to the time-demeaned form or within transformation of Equation 2.8. The consistency of this estimator, however, requires strict exogeneity assumption, that is, $Cov(X_{is}, u_{it}) = 0$ for all $s, t = 1, 2, \ldots, T$, which rules out lagged dependent variables, that is, excludes situations where shocks today affect future decisions about the covariates (Wooldridge 2010). This assumption is again very restrictive because in a dynamic context, changes in covariates over time might be related to past shocks. We further explain the implication of this restriction in Chapter 3.

The IV approach allows for correlations between X_{it} and v_{it} via the use a vector of

time-varying IVs, given by Z_{it} that are (i) exogenous, that is, uncorrelated with the composite errors v_{it}, and, (ii) relevant, that is, strongly correlated with the covariate vector X_{it}. The most widely used IV-based estimation, known as the two-staged least squares (2SLS) estimation, follows a two-step process. In the first stage, the endogeneous covariate are predicted using the instrument variable, followed by the second stage in which these predictions are used as covariates to estimate a regression model for the response variable. Such an estimator is consistent if, (a) $Cov(Z_{it}, c_i) = 0$, and, (b) $Cov(Z_{it}, u_{it}) = 0$, for all $t = 1, \ldots, T$ (Wooldridge 2010). To eliminate time-invariant heterogeneity, estimates are obtained by first applying first-differencing as in equation 2.9, followed by IV estimation. This again requires strict exogeneity of IVs, that is, $Cov(Z_{is}, u_{it}) = 0$; for all $s, t = 1, 2, \ldots, T$, for consistency.

$$\Delta y_{it} = \Delta X_{it}\beta + \Delta u_{it}, \tag{2.9}$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$ and so on.

In the absence of external IVs, suitable instruments can be derived from the panel nature of the dataset. Lagged levels of endogenous covariates can be used as their instruments for differenced equations. In this case, consistency of the estimator relies on the sequential exogeneity assumption that covariates X_{it} are chosen before anything is known of u_{it}, that is, $Cov(X_{is}, u_{it}) = 0$ for all $s \leq t$. Parameter estimates are obtained via GMM estimation.

Time-demeaning and first-differencing operations mentioned previously lead to complete elimination of the cross-sectional variation in time-invariant covariates, resulting into a downward bias in the parameters estimates of such covariates. A dynamic panel model allows us to overcome this problem via inclusion of an auto-regressive (AR) component. This AR component helps us investigate any adjustments in the response variable conditional on the response in the previous time period as shown in equation 2.10.

$$Y_{it} = Y_{i,t-1}\rho + X_{it}\beta + v_{it}, \tag{2.10}$$

Parameter estimates of this model are again derived via GMM estimation. We discuss these estimators in detail in Chapter 3.

Non-parametric models

As mentioned in the previous subsection, to address potential endogeneity biases, we adopt regression estimators with instrumental variables (IV). IV-based estimators such as two-stage least squares (2SLS) are widely adopted in applied econometrics to estimate parametric models that contain endogenous explanatory variables. However, finite-dimensional parametric models, for instance, with a linear or a quadratic specification, may not always be suitable as such functional form restrictions are based on assumptions that are rarely justified by engineering or economic theories. The resulting mis-specification may lead to erroneous estimates of the model. On the other hand, non-parametric methods have the potential to capture the salient features in a data-driven manner without making a priori assumptions on the functional form of the relationship (Horowitz 2011). Therefore, a fairly growing strand in the econometrics literature proposes different approaches for non-parametric instrumental variables (NPIV) regression, but such methods have not been considered in the transportation engineering and economics literature. Extensive reviews of these methods can be found in Newey & Powell (2003) and Horowitz (2011).

The NPIV approaches are either based on regularisation or control function. In Chapters 4, 5 and 6, we adopt a control-function based Bayesian NPIV estimator proposed by Wiesenfarth et al. (2014). In what follows, we start with the general model set-up. Subsequently, we summarise challenges in the regularisation-based approach, followed by discussing the advantages of the adopted control-function-based Bayesian approach.

In Chapters 4, 5 and 6, we have a model with a traditional two-stage IV-based regression set up:

$$y = S(x) + \epsilon_2, \quad x = h(z) + \epsilon_1 \tag{2.11}$$

with response y, endogenous covariate x, IV z for x and idiosyncratic error terms ϵ_1 and ϵ_2 for the first and second stage regressions, respectively. For the notational simplicity, we drop unit-time subscripts. Endogeneity bias arises as $E(\epsilon_2|x) \neq 0$. We assume the following identification restrictions:

$$E(\epsilon_1|z) = 0 \quad and \quad E(\epsilon_2|\epsilon_1, z) = E(\epsilon_2|\epsilon_1), \tag{2.12}$$

which yields

$$\begin{aligned} E(y|x,z) &= S(x) + E(\epsilon_2|\epsilon_1, z) = S(x) + E(\epsilon_2|\epsilon_1) \\ &= S(x) + \nu(\epsilon_1), \end{aligned} \tag{2.13}$$

where $\nu(\epsilon_1)$ is a function of the unobserved error term ϵ_1. This function is known as the control function.

Regularisation-based approaches

Regularisation-based approaches to NPIV regression assume $y = g(x) + \epsilon$ with $E[\epsilon|z] = E[(y - g(x))|z] = 0$. There are three challenges to this approach.

First, within this framework, it becomes difficult to recover non-linearities as the data identifies only reduced form conditional expectations, that is, $E[y|z] = E[g(x)|z] = \int g(x)f(x|z)dx$, leading to an ill-posed inverse problem (for details, see Newey & Powell 2003, Horowitz 2011). An ill-posed problem means that the solution is not continuous in functions $E[y|z]$ and $f(x|z)$, which implies that a consistent estimator of $g(.)$ may not result from plugging in consistent estimators of $E[y|z]$ and $f(x|z)$, and approximately solving this equation. To overcome this problem, several ways to regularise the integral have been proposed in the literature, the most common being a series estimator. A series

estimator is based on approximating the unknown function $g(x)$ by a linear combination of known functions, such as power series or regression splines (for details, see Horowitz 2011, Newey 2013, Chetverikov & Wilhelm 2017).

However, the approximation leads to the second challenge with these amendments not being entirely data-driven and requiring some user-defined parameters, such as, the degree of the power series or spline-basis approximating function (for instance, see Chetverikov & Wilhelm 2017). Subsequently, model selection becomes difficult in such approaches.

Furthermore, the third challenge arises from the difficulty in constructing simultaneous confidence bands (that is, inference) due to the inherent asymptotic bias in non-parametric estimators (for details, see Horowitz 2011). In a very first attempt to solve this problem, Horowitz & Lee (2012) proposed a solution based on bootstrapping where they chose the regularisation parameter smaller than optimal so that the bias term becomes asymptotically negligible. However, the practical applicability of this method is limited as the required degree of under-smoothing is unknown. Moreover, the variability due to estimating the smoothing and regularisation parameters are not taken into account, which may lead to under-coverage in small samples. Newey (2013) proposed a simpler approach to obtain standard errors by treating the series estimator as parametric, which further justifies that regularisation-based approach is not truly non-parametric from an application perspective.

Control function-based approaches

Several control function-based approaches to estimation of equation 2.11 in the literature, adopt a two-stage approach where residuals $\hat{\epsilon_1}$, that is, $x - \hat{h(z)}$ from the first stage are used as additional covariate in the second stage (for details, see Newey & Powell 2003). However, as pointed out by Wiesenfarth et al. (2014), such two-stage approaches have certain limitations. First, the uncertainty introduced by estimating the parameters in the first stage remains unincorporated in the second stage. Second, a precise estimate of $\nu(\epsilon_1)$ to achieve full control for endogeneity is difficult to obtain because the focus is on minimising the error in predicting x in the first stage. Third, a robustness control

is required to account for outliers and extreme observations in ϵ_1 that may affect the endogeneity correction.

Bayesian control-function-based approaches can address these shortcomings of frequentist counterparts and regularisation-based approaches by estimating equation 2.11 as a simultaneous system of equations, allowing for automatic smoothing parameter selection for a precise estimation of the control function and for construction of simultaneous credible bands [2].

However, early Bayesian control-function-based approaches consider a bivariate Gaussian distribution of errors $(\epsilon_1, \epsilon_2) \sim N(0, \Sigma)$ (for instance, see Chib et al. 2009). This assumption leads to linearity of the conditional expectation as, $E(\epsilon_1|\epsilon_2) = \frac{\sigma_{12}}{\sigma_1^2}$, where $\sigma_{12} = cov(\epsilon_1, \epsilon_2)$ and $\sigma_1^2 = var(\epsilon_1)$, restricting the control function to be linear in ϵ_1 (Wiesenfarth et al. 2014, Conley et al. 2008). Since outliers can be a common source of non-linearity in error terms, they can aggravate the robustness issues of such linear specifications. To overcome these limitations, Conley et al. (2008) proposed the application of a Dirichlet process mixture (DPM) prior to obtain a flexible error distribution, but still relied on linear covariate effects. The method proposed by Wiesenfarth et al. (2014) and adopted in this study, extends the approach by Conley et al. (2008) and allows for fully-flexible covariate effects.

Adopted Bayesian NPIV approach (Wiesenfarth et al. 2014)

The Wiesenfarth et al. (2014)'s Bayesian NPIV approach thus allows us to correct for endogeneity bias in regression models where the covariate effects and error distributions are learned in a data-driven manner, obviating the need of a priori assumptions on the functional form. Bias correction relies on a IV-based simultaneous equation specification (see equation 2.11) and the joint error distribution is modelled flexibly via a DPM prior. To account for nonlinear effects of continuous covariates, both the structural and instrumental variable equations (i.e., $S(.)$ and $h(.)$ in equation 2.11) are specified in terms of additive

[2] Credible bands are the Bayesian analogue to confidence bands in the frequentist set up that represent the uncertainty of an estimated curve.

predictors comprising penalised splines. Efficient Markov chain Monte Carlo (MCMC) simulation method is employed for a fully Bayesian inference. The resulting posterior samples allow us to construct simultaneous credible bands for the non-parametric effects.

A contextual discussion of the Wiesenfarth et al. (2014)'s Bayesian NPIV approach is further presented in Chapters 4, 5 and 6.

References

Akbar, P. & Duranton, G. (2017), 'Measuring the cost of congestion in highly congested city: Bogotá'.

Anderson, M. L. & Davis, L. W. (2018), 'Does hypercongestion exist?: New evidence suggests not', *National Bureau of Economic Research* .

Anderson, M. L. & Davis, L. W. (2020), 'An empirical test of hypercongestion in highway bottlenecks', *Journal of Public Economics* **187**, 104197.

Basso, L. J., Jara-Díaz, S. R. & II, W. G. W. (2011), Cost functions for transport firms, *in* A. de Palma, R. Lindsey, E. Quinet & R. Vickerman, eds, 'A Handbook of Transport Economics', Edward Elgar Publishing Ltd., Northampton, chapter 12, pp. 273–297.

Becker, G. S. (1965), 'A theory of the allocation of time', *The economic journal* **75**(299), 493–517.

Borts, G. H. (1960), 'The estimation of rail cost functions', *Econometrica, Journal of the Econometric Society* pp. 108–131.

Cameron, A. C. & Trivedi, P. K. (2005), *Microeconometrics: methods and applications*, Cambridge university press.

Carey, M. & Else, P. K. (1985), 'A reformulation of the theory of optimal congestion taxes', *Journal of transport economics and policy* **19**(1), 91–94.

Cassidy, M. J. & Bertini, R. L. (1999), 'Some traffic features at freeway bottlenecks', *Transportation Research Part B: Methodological* **33**(1), 25–42.

Cassidy, M. J. & Rudjanakanoknad, J. (2005), 'Increasing the capacity of an isolated merge by metering its on-ramp', *Transportation Research Part B: Methodological* **39**(10), 896–913.

Chetverikov, D. & Wilhelm, D. (2017), 'Nonparametric instrumental variables estimation under monotonicity', *Econometrica* **85**(4), 1303–1320.

Chib, S., Greenberg, E. & Jeliazkov, I. (2009), 'Estimation of semiparametric models in the presence of endogeneity and sample selection', *Journal of Computational and Graphical Statistics* **18**(2), 321–348.

Conley, T. G., Hansen, C. B., McCulloch, R. E. & Rossi, P. E. (2008), 'A semi-parametric bayesian approach to the instrumental variable problem', *Journal of Econometrics* **144**(1), 276–305.

Couture, V., Duranton, G. & Turner, M. A. (2018), 'Speed', *Review of Economics and Statistics* **100**(4), 725–739.

Daganzo, C. F. (1997), *Fundamentals of transportation and traffic operations*, Vol. 30, Pergamon Oxford.

Daganzo, C. F. (2002), 'A behavioral theory of multi-lane traffic flow. part ii: Merges and the onset of congestion', *Transportation Research Part B: Methodological* **36**(2), 159–169.

De-Borger, B., Kerstens, K. & Costa, A. (2002), 'Public transit performance: what does one learn from frontier studies?', *Transport reviews* **22**(1), 1–38.

Graham, D. J. (2008), 'Productivity and efficiency in urban railways: Parametric and non-parametric estimates', *Transportation Research Part E: Logistics and Transportation Review* **44**(1), 84–99.

Graham, D. J., Couto, A., Adeney, W. E. & Glaister, S. (2003), 'Economies of scale and density in urban rail transport: effects on productivity', *Transportation Research Part E: Logistics and Transportation Review* **39**(6), 443–458.

Hills, P. & Evans, A. W. (1993), 'Road congestion pricing: When is it a good pol-

icy?(comment and rejoinder)', *Journal of Transport Economics and Policy* **27**(1), 91–105.

Horowitz, J. L. (2011), 'Applied nonparametric instrumental variables estimation', *Econometrica* **79**(2), 347–394.

Horowitz, J. L. & Lee, S. (2012), 'Uniform confidence bands for functions estimated nonparametrically with instrumental variables', *Journal of Econometrics* **168**(2), 175–188.

Jara-Diaz, S. R. (1982), 'The estimation of transport cost functions: a methodological review', *Transport Reviews* **2**(3), 257–278.

Johnson, M. B. (1964), 'On the economics of road congestion', *Econometrica (pre-1986)* **32**(1, 2), 137.

Lindsey, R. & Verhoef, E. (2007), *Congestion modelling*, Handbook of Transport Modelling: 2nd Edition, Emerald Group Publishing Limited, pp. 417–441.

May, A. D. (1990), *Traffic flow fundamentals*, Englewood Cliffs, N.J. : Prentice Hall.

McFadden, D. (1978), Cost, revenue, and profit functions, *in* M. Fuss & D. McFadden, eds, 'Production Economics: A Dual Approach to Theory and Application', Vol. 1, Amsterdam: North Holland.

Mun, S. (1999), 'Peak-load pricing of a bottleneck with traffic jam', *Journal of Urban Economics* **46**(3), 323–349.

Newbery, D. M. (1990), 'Pricing and congestion: economic principles relevant to pricing roads', *Oxford review of economic policy* **6**(2), 22–38.

Newey, W. K. (2013), 'Nonparametric instrumental variables estimation', *American Economic Review* **103**(3), 550–556.

Newey, W. K. & Powell, J. L. (2003), 'Instrumental variables estimation of nonparametric models', *Econometrica* **71**(5), 1565–1578.

Ohta, H. (2001), 'Probing a traffic congestion controversy: density and flow scrutinized', *Journal of Regional Science* **41**(4), 659–680.

Oum, T. H. & Waters, W. G. (1996), 'A survey of recent developments in transportation cost function research', *Logistics and Transportation Review* **32**(4), 423–463.

Small, K. A. & Chu, X. (2003), 'Hypercongestion', *Journal of Transport Economics and Policy (JTEP)* **37**(3), 319–352.

Small, K. A. & Verhoef, E. T. (2007), *The economics of urban transportation*, Routledge, New York.

Varian, H. R. (2014), *Intermediate Microeconomics: A Modern Approach: Ninth International Student Edition*, WW Norton and Company.

Verhoef, E. T. (1999), 'Time, speeds, flows and densities in static models of road traffic congestion and congestion pricing', *Regional Science and Urban Economics* **29**(3), 341–369.

Verhoef, E. T. (2001), 'An integrated dynamic model of road traffic congestion based on simple car-following theory: exploring hypercongestion', *Journal of Urban Economics* **49**(3), 505–542.

Viton, P. A. (1981), 'A translog cost function for urban bus transit', *The Journal of Industrial Economics* pp. 287–304.

Walters, A. A. (1961), 'The theory and measurement of private and social cost of highway congestion', *Econometrica: Journal of the Econometric Society* pp. 676–699.

Wiesenfarth, M., Hisgen, C. M., Kneib, T. & Cadarso-Suarez, C. (2014), 'Bayesian nonparametric instrumental variables regression based on penalized splines and dirichlet process mixtures', *Journal of Business and Economic Statistics* **32**(3), 468–482.

Wooldridge, J. M. (2010), *Econometric analysis of cross section and panel data*, MIT press.

Chapter 3

Understanding the costs of urban rail transport operations

There is considerable variation in the average cost of operations across urban rail transport (or metro) systems. Since metros are typically owned and operated by public authorities, there is a public interest case in understanding the key drivers of their operational costs. This chapter estimates short-run cost functions for metro operations using a unique panel dataset from twenty-four metro systems around the world. We use a flexible translog specification and apply dynamic panel generalised method of moments (DPGMM) estimation to control for confounding from observed and unobserved characteristics of metro operations. Our empirical results show that metro systems with a high density of usage are the most cost-efficient. We also find that operational costs fall as metro size increases. These results have important implications for the economic appraisal of metro systems. The core findings of this chapter have been published as:

> Anupriya, Graham, D.J., Carbo, J.M., Anderson, R.J., & Bansal, P. (2020). Understanding the costs of urban rail transport operations. *Transportation Research Part B: Methodological*, Vol. 138, pp. 292-316.

3.1 Introduction

To cater for the needs of the growing urban travel demand, cities around the world are increasingly investing in high-capacity urban rail transportation systems, also known as metros. According to the International Association of Public Transport, around forty-five new metros have been opened in the last decade and another two-hundred new metro-lines are expected over the next five years[1]. Metros, therefore, are very important in attaining urban mobility requirements. Comparison of average costs (that is, cost per unit car-kilometre) of metro operations reveals considerable variation across systems. Figure 3.1 shows average operational costs in 2015 for a group of thirty-two metro systems, normalised with respect to the mean value of the group. The normalised average cost ranges from 0.5 to 3.0, showing that the order of magnitude of unit operational costs for some metros is six times higher than others. Since metros are typically owned and operated by public authorities, there is a public interest case in understanding the underlying technology that determines the observed differences in unit costs.

There exists a large volume of research dedicated to the empirical analysis of transportation costs. As discussed in Chapter 2, the estimation of cost functions is important to compare the performance of firms over time and across different regulatory regimes. It also facilitates a broad characterisation of the industry by determining the extent of scale economies. While previous studies have mostly analysed cost structures and productivity for main line railways and for the airlines industry, the academic literature on analysis of costs of metro operations is relatively scarce.

One of the main challenges for the existing literature is treatment of endogenous covariates in the estimation of a metro cost function. The assumption that covariates such as output and factor prices are fixed and known to the firm a priori is misleading. As metro firms typically have market power, there are unobserved firm-level sources of efficiency or productivity that play an important role in determining the firm's decision on the

[1] https://www.uitp.org/world-metro-figures-2018

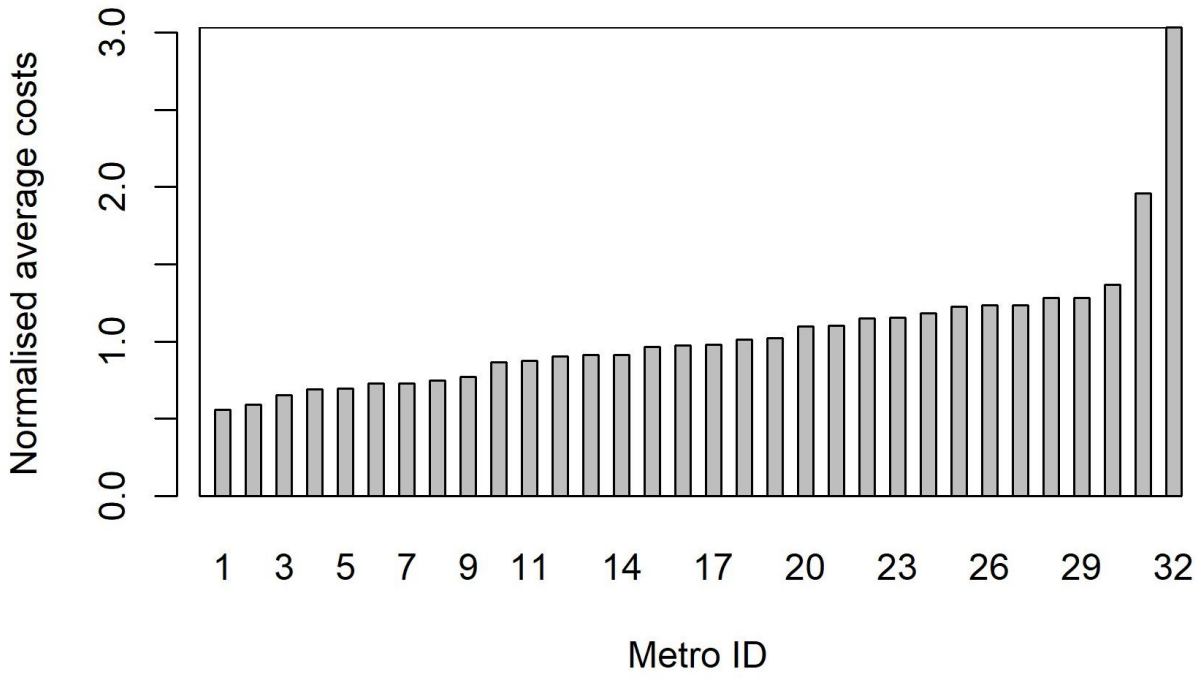


Figure 3.1: Normalised average costs of urban metro operations, 2015.

quantity of its output and its level of input prices. These covariates are thus endogenously determined by the firm. Some previous attempts to address such confounding issues focus only on endogeneity of firm output through the application of instrumental variables (IV) estimation to linear cost models (Savage 1997, Mizutani 2004). However, most of the past studies do not offer adequate control for firm level unobserved sources of confounding. This lack of control on endogenous covariates can result in biased estimates of the scale economies in provision of metro services (Greene 1980). These scale economy estimates, (i) returns to network size and (ii) returns to density, develop our understanding of the technology driving unit-cost differences in the metro industry. We find that there may be inconsistency between the scale economy estimates from the literature and the observed behaviour in the metro industry. While the weight of evidence in the literature supports constant returns to network size, metro firms have been found to expand their network over time (Basso et al. 2011). Some researchers rightly suggest that even without any significant cost savings, metro firms may expand their network due to a host of reasons including serving more of a city in order to justify raising subsidy funds for a citywide tax,

economic development and so on. However, we argue that in addition to these reasons, there may be various other sources that may result into true cost advantages for metro firms when they expand their networks. For instance, as suggested by Graham et al. (2003), in cases where sufficient latent demand exists, a metro firm may try to exploit the density economies by expanding the network. Such cost savings may not be reflected in the estimates of economies of scale from the literature.

In this research, we estimate the causal relationship between short-run metro operational costs and output using a flexible translog specification and dynamic panel generalised method of moments (DPGMM) estimation. The application of DPGMM can effectively deal with endogeneity in various covariates in the cost-output relationship. To our knowledge, this is the first time that a transport cost function has been studied using such econometric methods. The original DPGMM formulation used in this study were introduced by Arellano & Bover (1995) and Blundell & Bond (2000) for the estimation of production functions with unobserved firm level confounding. Here, we apply these developments to the analysis of costs using the fact that a cost function is dual to its production technology.

In this chapter, we make use of a unique panel dataset that has been collected by the Transport Strategy Centre (TSC) at Imperial College London since 1994. The data relate to twenty-four metro systems around the world. In contrast to previous studies that mostly use country specific data on metro systems (for instance, Savage (1997) uses data on US metro systems), we combine data from metros of varied sizes and operational characteristics from all over the world. We thus develop an inclusive understanding of the short-run operational cost structure for the metro industry. We also provide estimates of some important descriptors of the underlying technology that are the nature of scale economies, input factor separability and homotheticity.

The major contributions of this research can be summarised as follows:

1. We contribute with a unique and very high quality panel data to estimate the

technology underlying cost of short-run operations of metro systems.

2. We develop a rigorous understanding of the endogeneity issues in the empirical estimation of the cost function and apply an appropriate econometric framework to address these issues.

3. We provide new and more reliable empirical insights into the external sources of cost-efficiency for metro systems.

This chapter is organised as follows. Section 3.2 reviews the relevant literature. Section 3.3 describes the data and variables. Section 3.4 explains the properties and advantages of DPGMM by comparing it with traditional econometric approaches used in the cost function literature. Section 3.5 presents a simulation to compare parameter estimates from different methodologies. Section 3.6 presents our results. Section 3.7 presents an application of our scale economy estimates to understand the variation in unit operational costs across metro systems. Conclusions and implications for policy are presented in the final section.

3.2 Literature Review

The literature on the estimation of transport cost functions and the nature of scale economies in the provision of transport services is well known and several extensive reviews can be found in Jara-Diaz (1982), Oum & Waters (1996) and Basso et al. (2011). In this section, we discuss scale economies and endogeneity issues in the cost function estimation, and highlight how developments in the production function estimation can be employed to correctly estimate the cost function.

3.2.1 Scale economies

Conventionally, researchers derive two main descriptors of industry behaviour from cost studies: economies of scale and economies of density. Economies of scale or returns to

scale (RTS) describe the relationship between average costs and overall scale of operations varying both output and network size, that is, they reflect returns to firm size. Economies of density or returns to density (RTD) refer to the relationship between average costs and output keeping network size fixed. Jara-Diaz & Cortes (1996) and Basso & Jara-Diaz (2006) suggest the use of economies of spatial scope[2] in place of RTS as they point out that RTS analyses the effect of increasing network size but with same density of traffic movements along each link in the network. RTS estimates, however, are still widely used to explain network expansions (refer the literature review summary Table 3.1) because as Batarce (2016), Batarce & Galilea (2018) point out, the exact definition of economies of spatial scope is difficult to apply using an aggregate measure of output. Therefore, in this analysis we focus on RTS and RTD estimates.

Although the literature on the analysis of cost structures and productivity for mainline railways is huge, studies focusing on the costs of urban rail transport operations is relatively scarce. The cost characteristics of metro systems do not necessarily correspond to those of mainline railways (Graham et al. 2003), however, we review the mainline railway literature to develop an understanding of the determinants of costs in the railway industry in general. The mainline railway literature indicates the presence of increasing RTD over a wide range of output (see Table 3.1 for a summary of results from key literature). Increasing RTD arises due to the existence of a range of fixed and semi-fixed costs in the railway industry that do not vary proportionally with output (Graham et al. 2003).The results on the existence of RTS are inconsistent (refer Table 3.1). The majority support the prevalence of constant RTS implying that railway firms do not have significant cost advantages in expanding their networks. However, railway firms have been found to expand their networks over time and this observed behaviour is inconsistent with the

[2]A transport firm produces movements of different types (passengers or freight) between different origins and destinations (ODs) at different time periods. The final output of a transport firm is actually a vector $Y = y_{ijt}$, where y_{ijkt} represents flow between OD pair ij at time period t. Economies of spatial scope measures the cost advantages (or disadvantages) of jointly producing two sets of outputs (Refer Basso & Jara–Díaz (2005) for an example application).

constant RTS estimates from the literature (Jara-Diaz 1982, Oum & Waters 1996, Basso et al. 2011). A limited few cost studies (Pozdena & Merewitz 1978, Savage 1997) and productivity studies (Graham et al. 2003, Graham 2008) on urban rail transport (see Table 3.1) draw similar conclusions. It is evident that the RTS estimates are unable to explain the network expansions in the industry. This has stimulated the growth of many interesting facets in the literature, from identifying and controlling for endogeneity issues in cost function estimation, to application of new estimation methods, to redefinition of traditional descriptors of the technology in the industry. The following paragraphs summarise the main findings.

3.2.2 Endogeneity challenges in cost function estimation

Early studies on transport cost functions assume that covariates in a cost model are fixed or exogenously determined (see Jara-Diaz (1982), Oum & Waters (1996) for details). To support the assumption on exogeneity of the output variable, these studies argue that since firms are regulated and fares are normally imposed externally, firms cannot influence their level of demand. Thus, the level of output produced by the firm is known to the firm a priori. This assumption can lead to problem of endogeneity in estimation because a firm's decision on the quantity of its output should depend upon its level of productivity that cannot be directly observed. Moreover, Jara-Diaz (1982) rightly adds in criticism of the output exogeneity assumption that, (i) not all firms are regulated, and (ii) demand levels are also influenced by the level of service (crowding, reliability etc.) in addition to fares. A few studies, for instance Savage (1997) consider output as endogenous and use instrument variables influencing demand like population density per unit area. Although exogeneity of these instruments has been argued based on investigation of the data used

Table 3.1: Summary of key literature on the existence of RTS and RTD in transport operations in the short-run.

Author(s)	Country	Data	Dependent Variable	Functional Form	Measure of Output	Estimation Methodology	Returns to Scale	Returns to Density
Studies on Mainline Railway:								
Wills-Johnson (2010)	Australia	1900-1992	Total Cost	Translog	Multi-output	POLS	Increasing	Increasing
Ivaldi & McCullough (2007)	US	1981-2004	Total Cost	Translog	Car-miles	ML	Increasing	
Farsi et al. (2005)	Switzerland	1985-1997	Total Cost	Cobb-Douglas	Multi-output	POLS, RE, FE	Increasing	Increasing
Mizutani (2004)	Japan	1970-2000	Variable Cost	Translog	Vehicle-kilometres	ML	Increasing	Increasing
Bitzan (2003)	US	1983-1997	Variable Cost	Translog	Ton-miles	FE	Increasing	Increasing
Sànchez & Villaroyya (2000)	Europe	1970-1990	Total Cost	Translog	Multi-output	SFA	Constant	
Hensher et al. (1995)	Australia	1971-1992	TFP index	Parametric	Multi-output	POLS		Increasing
McGeehan (1993)	Ireland	1973-1983	Variable Cost	Translog	Multi-output	POLS		Increasing
Filippini & Maggi (1992)	Switzerland	1985-1988	Total Cost	Translog	Wagon-kilometres	POLS	Increasing	Increasing
Caves et al. (1981)	US	1955-1974	Variable Cost	Translog	Multi-output	POLS	Constant	
Keeler (1974)	US	1969-1971	Total Cost	Cobb-Douglas	Ton-miles	POLS	Constant	Increasing
Studies on Urban Rail Transport:								
Graham (2008)	World-wide	1996-2007	Output	Cobb-Douglas	Pax. journeys	POLS	Constant	Increasing
Graham et al. (2003)	World-wide	1996-2002	Output	Translog	Car-kilometres	DEA/TFP	Constant	Increasing
Gagnepain & Ivaldi (2002)	France	1985-1993	Total Cost	Cobb-Douglas	Seat-kilometres	SFA	Increasing	
Savage (1997)	US	1985-1991	Variable cost	Translog	Multi-output	IV: 3SLS	Constant	Increasing
Pozdena & Merewitz (1978)	US	1960-1970	Total cost	Cobb-Douglas	Vehicle-miles	POLS	Increasing	Increasing
Studies on Urban Bus Transport:								
Batarce & Galilea (2018);	Chile	2007-2010	Variable cost	Cobb Douglas	Vehicle-kilometres	POLS	Increasing	Increasing
Karlaftis & McCarthy (2002);	US	1986-1994	Variable cost	Translog	Vehicle-miles	POLS	Increasing	Increasing
Karlaftis et al. (1999)							(small sys.)	(large sys.)
							Decreasing	Decreasing
							(large sys.)	(suburban)
Viton (1981)	US	1975	Variable cost	Translog	Vehicle-miles	POLS	Decreasing	Increasing
							(large systems)	

*POLS: Pooled Ordinary Least Squares; FE: Fixed Effects; RE: Random Effects, IV: Instrumental Variables, 3SLS: Three stage least squares

**SFA: Stochastic Frontier Analysis, TFP: Total Factor Productivity, DEA: Data Envelopment Analysis, ML: Maximum Likelihood, pax.: passengers, sys.:systems

in this study, their generalisation to other studies is questionable. Similar assumptions on exogeneity of factor prices (for example, see Savage (1997)) can aggravate the bias in the estimated parameters. This assumption is again backed by the unrealistic argument that transport firms do not have the ability to control their input prices. Some studies use exogenous factor prices like gross domestic product per capita for labour prices (for instance, see Karlaftis & McCarthy (2002)) as proxies for actual factor prices to overcome this bias in the estimation. However, these proxies do not take into account the actual productivity differences between firms (Borts 1960). We argue that if a firm has buying power, then it is quite feasible for the firm to set its input prices in the short-run in response to the input quantities and its productivity. Input prices are thus endogenous[3]. In order to reduce the correlation between variables of a transport cost function and its error term, researchers have also added hedonic characteristics or *attributes* of output[4] to the cost specification, which control for the differences in operational conditions and network characteristics between firms (Jara-Diaz 1982, Oum & Waters 1996). However, some of these hedonic characteristics are endogenous. We introduce a methodological framework for cost function estimation that can appropriately control for endogeneity in multiple covariates.

3.2.3 Endogeneity due to unobserved inefficiencies

The main assumption behind a cost function is that the firm minimises the expenses incurred in producing a given level of output. A related strand in the literature hypothesises that transport firms may not be cost minimising due to several reasons including (i) threat posed by competitive forces and elastic demand or (ii) technical inefficiencies due to

[3]Refer section 3.3 for a discussion on the various input prices for a metro firm and the type of control the metro firm may have over these prices in the short-run and in the long-run.

[4]Examples of hedonic characteristics include average length of haul that captures the characteristics of the market served by the transport system (Caves et al. 1981) and indicators for peaking such as peak-to-base ratio intended to capture the overall productivity of the factors of production and those for average loads (Savage 1997).

existence of a regulatory framework (as mentioned in Basso et al. (2011))[5]. More recently, techniques like data envelopment analysis have been used to account for these unobserved inefficiencies in the cost model. Gagnepain & Ivaldi (2002) argue that the cost frontiers estimated using these techniques are capable to correct for the problem of endogeneity in traditional regression analysis. However, we argue that even in the presence of these sources of inefficiencies firms may exhibit cost minimising behaviour subject to constraints. Therefore, to obtain valid conclusions from our estimated cost models, it is important to account for these unobservables in the cost function estimation. To support our hypothesis, we discuss some key findings from the complimentary production function literature in the rest of this section.

3.2.4 Production function estimation

Past few decades of research in total factor productivity (TFP) analysis has seen interesting developments particularly on empirical front (refer (Beveren 2012) for a comprehensive review of these developments). The main caveats identified are: (i) TFP analysis using traditional methods like pooled ordinary least squares (POLS) estimation introduces simultaneity or endogeneity bias because a firm's unobserved productivity shocks and its input choices are likely to be correlated (De-Loecker 2011). (ii) Selection bias emerges if no allowance for entry and exit are made, as entry and exit decisions are systematically related to unobserved productivity differences. Input choices of a firm are conditional on its survival and in turn on these unobserved productivity differences (Ackerberg et al. 2007). (iii) In the presence of imperfect competition, using industry-level price indices

[5]The first point relates to issues of cross-efficiency in a market where a transport firm may have a few effective competitors or threat from a potential entrant. Under such a scenario, the operational structure of the firm not only influences its own costs and demand but also the profit of its competitor firms. For instance, the transport firm may adopt a service frequency or a route structure that may not necessarily be cost minimising, however, such a structure allows it to increase its own demand at the expense of its competitor. The second point corresponds to a framework under which a regulatory agent asks a transport firm to produce a given level output and in turn covers the associated costs. However, the regulator has limited knowledge of the firm's production technology. This may impact the input allocation and cost reducing efforts of the firm and thus, the firm may not be cost-minimising at a given level of output.

or scheduled rates as proxies for firm-level prices leads to a biased representation of the firm's productivity level (Katayama et al. 2009). This is because a firm sets its input price in response to its input quantity and productivity.

In essence, unobserved productivity has been recognised to play a key role in the TFP analysis. Not accounting for this unobserved productivity can lead to erroneous estimates of coefficients associated with factor inputs and underestimation of returns to scale. Studies on TFP analysis have evolved progressively to develop estimation methodologies that can appropriately control for the unobserved productivity in the estimation of production function (for example, see Olley & Pakes (1996), Blundell & Bond (2000), Levinsohn & Petrin (2003), Wooldridge (2009) and so on).

Since the cost function is the dual characterisation of the firm's production technology, all relevant covariates present in a firm's production function should also be embodied in its cost function. In this study, we address the deficiencies in the estimation of a metro cost function by applying the developments summarised above in the production function literature.

3.3 Data and Relevant Variables

In this section we describe the variables that are used in this analysis along with their respective sources and hypotheses. We make use of data that has been collected by two consortia of metro system operators, namely the Community of Metros (CoMET) and the Nova Group of Metros (Nova) [6] since 1994, managed by the Transport Strategy Centre (TSC) at Imperial College London. The consortia focus on benchmarking using an extensive dataset comprising of key performance indicators related to 38 metro operations in 36 cities around the world[7]. However, the dataset has several missing values depending upon the extent of information that is reported by the metro operator during each year.

[6] https://cometandnova.org/

[7] These metros are represented in figure A.1 in Appendix A.1

Accordingly, we obtain an unbalanced panel dataset with 165 observations consisting of 24 systems over 14 years, between 2003 and 2016. The high-standard quality of this dataset is worth noting, as during the years of benchmarking work, the dataset has been cleaned systematically using one-to-one verification with operators and validation tests. Due to the sensitive commercial nature of the TSC data, we present our results in an anonymised form.

For the purpose of this study, we use data on operational costs, as well as a set of covariates used in the literature that are fundamental in describing the relationship between transport costs and output, derived on the basis of production-cost duality.

Cost Variable

This analysis deals with short-run variable costs of operation of metro systems. We use the data on total operating costs, that is, the sum of all costs of operations of the system including maintenance of rolling stock and way and structure, however, excluding capital investments, either related to depreciation or asset renewals as these are fixed in the short-run. The main components of the total operating cost variable are service costs[8], administration costs[9] and maintenance costs[10] (see figure A.2 in Appendix A.1).

The cost data have been converted to 2016 international dollar equivalent using a purchasing power parity (PPP) index [11] [12] published by the World Bank. We observe that the literature is inconsistent in terms of defining the short-run variable costs of metro

[8]Service costs comprise the energy and labour costs required to move the train along the network and to operate the stations. Such costs, for instance, include train service costs related to drivers and traction, service costs related to ticketing, revenue control, police and security, among others.

[9]Administrative costs refers to all engineering and project costs and general administration costs for the metro, for instance, costs related to marketing, revenue development, human resource development, information systems and communications technology, expenditure on public or corporate relations, purchasing, contracting, procurement and so on.

[10]Maintenance costs consist of the costs of maintaining all the assets used by the train, that is, track, station and rolling stock. These include costs that cannot be capitalised in the short-run, for instance, costs related to cleaning of rolling stock and stations and train service parts, and do not include any capital investments, for instance, asset renewals.

[11]https://data.worldbank.org/indicator/NY.GDP.MKTP.PP.CD.

[12]Although the use of the World Bank PPP indices for international cost comparisons has its own limitations, however, this is a standard convention adopted in the literature for empirical work of this type.

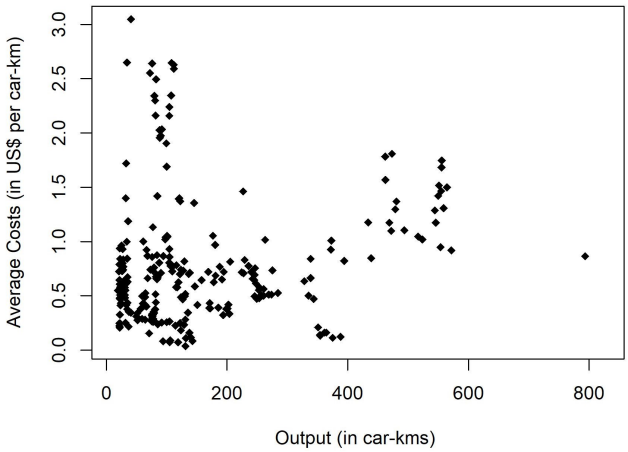
operations. The studies that estimate constant RTS mostly assume that way and structure maintenance costs are fixed in the short-run (refer Table 3.1) and exclude these costs from short-run cost analysis. However, maintenance costs in the short-run comprise both fixed and variable costs. The fixed component of maintenance costs relate to costs that can be capitalised, for instance, costs spent on asset renewals or depreciation costs. However, there exists other rolling stock and infrastructure maintenance costs in addition to these fixed costs. Our TSC dataset suggests that over eighty-percent of these way and structure maintenance cost components are driven by labour and electricity. In the short-run, such maintenance costs can thus be flexibly adjusted in response to changes in planned outputs. Figure 3.2a shows the variation of average way and structure maintenance costs over output as measured in car-kilometres as per the TSC dataset. We also show the variation of the components of way and structure maintenance costs as reported in the TSC dataset (see figure 3.2b and figure 3.2c). Average maintenance costs roughly increase with output indicating that these cost components are variable in the short-run.

Output

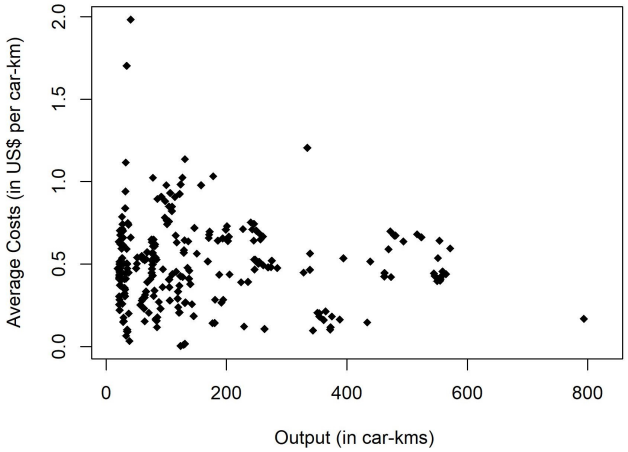
Actual car kilometres operated in revenue service is the primary aggregated measure of output in this analysis. We also use a secondary measure of output that captures terminal expenses driven by passenger usage. This measure is a load factor variable calculated as passenger kilometres divided by revenue car kilometres. Passenger kilometres refers to the sum of distances travelled by all passengers including fare evaders. We control for the endogeneity of both these output variables in our estimation.

Fixed factors

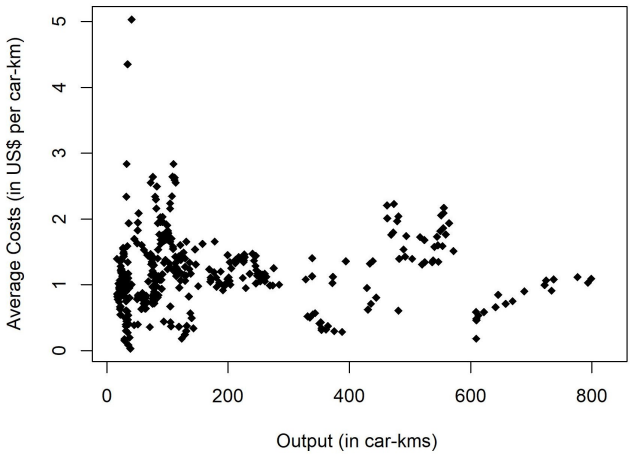
We use network size as the measure of the fixed factor of production, which is measured as the length of the network used by trains operating in service. That is, it refers to the sum of all lines excluding tracks in depots, yards and sidings and those used by trains for turning movements.



(a) Variation of Average Way and Structure Maintenance Costs (in 2016 US$ equivalent per km) over Output (in car-kms in millions)



(b) Variation of Average Infrastructure Maintenance Costs (in 2016 US$ equivalent per km) over Output (in car-kms in millions)



(c) Variation of Average Station Facilities Maintenance Costs (in 2016 US$ equivalent per km) over Output (in car-kms in millions)

Figure 3.2: Variation of Average Maintenance Cost Components over Output

Factor prices

The literature suggests that in the presence of imperfect competition, firms set their input price in response to their respective input requirements and productivity (refer section 3.2.2). Therefore, we treat all factor prices as endogenous[13] and calculate these prices based on the data itself. We include three variable inputs: labour, energy and residuals. The TSC data reports two components of total labour costs,– own labour costs and contracted labour costs and their corresponding labour hours. We calculate unit price for labour by dividing the total labour costs by total labour hours and unit energy (electricity) costs as total energy costs divided by total energy consumption (reported in megawatt hours). We then convert these prices into 2016 international dollar equivalent.

The residual expenses in operations, that is, non-labour and non-energy costs are converted to unit prices by dividing total residual expenses (total operating costs minus total labour costs minus total energy costs) by capacity kilometres. Capacity kilometres represents a standardised measure of capacity that includes both seating and standing capacity, calculated by normalising standing density to four people per square metre. Residual price relates to the price of materials and services, which could not be capitalised, for instance, price of parts, cleaning materials, insurance, telecommunication services and so on[14].

Firstly, we discuss the sources of endogeneity in labour prices. A metro firm usually has to recruit from a large pool of the labour market. Most of the labour skills can be taught or trained, however, even specialist trained workers, for instance, train drivers, usually have the choice to work somewhere else. Therefore, minimum wages are usually subject to the labour market and the metro management has very limited control over

[13]In Appendix A.3, we carry out a robustness test to demonstrate the effect of treatment of factor prices as exogenous on the estimated parameters of our cost model.

[14]The operational cost models for metros in the literature do not include any material price, however we argue that the inclusion of this price is important as these residual expenses comprise approximately 20% of short-run operational costs of metros, which is quite a substantial part. In Appendix A.4, we carry out a robustness test to demonstrate the effect of exclusion of residual prices on the estimated parameters of our cost model.

Table 3.2: Summary statistics for variables used in the analysis.

Variable	Obs.	Min	Max	Median	Mean	Std.Dev
Operational Costs (m) (PPP US$)	165	58.92	3152.31	431.78	695.27	626.48
Car Kilometres (m)	165	19.35	793.60	91.90	124.76	110.40
Load Factor	165	17.38	82.84	49.53	49.01	17.13
Network length (km)	165	35.60	588.00	86.20	132.78	122.61
Labour Price (PPP US$/h)	165	5.09	61.33	27.23	29.74	14.14
Energy Price (PPP US$/MWh)	165	0.025	0.403	0.169	0.172	0.082
Residual Price (PPP US$/cap.km)	165	0.001	0.029	0.005	0.005	0.003

*Legend: Obs.: Observations, Std. Dev.: Standard Deviation, m: millions, cap.: capacity

these prices even if it is a very large employer. Although the management cannot fix wages downwards due to competition in the market, however, several years of benchmarking experience at the TSC suggests that the management often pays more than the market wage. Therefore, we argue that the metro firm may decide its labour price in response to its labour requirements and productivity in the short-run.

In the context of energy prices, the metro firm can leverage buying power because it is a huge buyer of electricity. It is the volume and long-term assurance over the sale of electricity which in general drives down the price of electricity for the metro firm. Thus, the metro firm has the ability and buying power to hedge prices in the short-run. However, long term prices with the supply market will eliminate these short term price fluctuations. On similar lines, we argue that the residual prices may be decided by the metro firm in the short-run.

Table 3.2 provides descriptive statistics for all variables used in the analysis. Appendix A.5 reports the variation of all variables for different metro systems over time.

3.4 Methodology

3.4.1 Theoretical framework

At any output level y, a producer chooses the level of input prices $\mathbf{w}$ that leads to the minimum cost C of producing y. For a three factor Cobb Douglas production technology

with one of the factors being fixed in the short-run, the short-run cost function is the solution to the following optimisation problem:

$$\begin{aligned} C^s(y, \mathbf{w}, \bar{x_3}) &= \min_{x_1, x_2} w_1 x_1 + w_2 x_2 + w_3 \bar{x_3} \\ &\text{such that} \\ y &= e^{\omega} x_1^{\alpha} x_2^{\beta} \bar{x_3}^{\gamma} \end{aligned} \tag{3.1}$$

where, x_1 and x_2 are the variable factors of production and w_1 and w_2 are corresponding factor prices. $\bar{x_3}$ is the fixed factor of production with corresponding factor price w_3. α, β and γ are constants representing the elasticities of output with respect to the associated factor of production. ω stands for the unobserved productivity differences between firms. Our definition of the production technology y is consistent with the recent production function literature (refer section 3.2), which suggests that ω is rudimentary in the definition of y.

The above constrained optimisation problem can be solved via the Lagrangian function:

$$L(y, \mathbf{w}, \bar{x_3}, \lambda) = w_1 x_1 + w_2 x_2 + w_3 \bar{x_3} + \lambda [y - e^{\omega} x_1^{\alpha} x_2^{\beta} \bar{x_3}^{\gamma}]$$

This gives the following first order conditions:

$$\begin{aligned} \frac{\partial L}{\partial x_1} &= w_1 - \lambda e^{\omega} \alpha x_1^{\alpha - 1} x_2^{\beta} \bar{x_3}^{\gamma} = w_1 - \lambda \alpha \frac{y}{x_1} = 0 \\ \frac{\partial L}{\partial x_2} &= w_2 - \lambda e^{\omega} \beta x_2^{\beta - 1} x_1^{\alpha} \bar{x_3}^{\gamma} = w_2 - \lambda \beta \frac{y}{x_2} = 0 \\ \frac{\partial L}{\partial \lambda} &= y - e^{\omega} x_1^{\alpha} x_2^{\beta} \bar{x_3}^{\gamma} = 0 \end{aligned}$$

Solving the above equations, we get $x_1 = \frac{\lambda \alpha y}{w_1}$, $x_2 = \frac{\lambda \beta y}{w_2}$, and $\lambda = \left(\frac{y^{(1-\alpha-\beta)} w_1^{\alpha} w_2^{\beta}}{e^{\omega} \alpha^{\alpha} \beta^{\beta} \bar{x_3}^{\gamma}} \right)^{\frac{1}{(\alpha+\beta)}}$. Substituting these back in the cost equation $C^s(y, \mathbf{w}, \bar{x_3})$ and separating the variable component from the total short-run cost, we get the short-run variable cost function, which is as follows:

$$\log CV^s(y, \mathbf{w}, \omega, \bar{x_3}) = \alpha_0 + \frac{\alpha}{\alpha+\beta}\log w_1 + \frac{\beta}{\alpha+\beta}\log w_2 + \frac{1}{\alpha+\beta}\log y - \frac{\gamma}{\alpha+\beta}\log \bar{x_3} - \frac{\omega}{\alpha+\beta} \tag{3.2}$$

We see that the unobserved productivity term 'ω' ends up in the cost function. Since more productive firms are more likely to produce more output, C and ω are negatively correlated. Thus in absence of ω, the scale economy estimates obtained from empirical cost analysis will have a downward bias as we observe bigger firms to have lower unit costs (Collard-Wexler 2012). Equation 3.1 is illustrative and can be generalised to include multiple factors of production. A second-order Taylor-series expansion of the terms of equation 3.2 yields a translog cost function that is most commonly used in transport cost studies.

3.4.2 The empirical model

The short-run variable cost function for a metro firm $i, i = 1, ..., N$ at time $t, t = 1, ..., T$, $CV^s_{it}(y, \mathbf{w}, N, \omega)$, can be represented by:

$$\begin{aligned}\log CV^s_{it} = {} & \alpha_0 + \alpha_y \log y_{it} + \sum_j \alpha_j \log w(j)_{it} + \alpha_N \log N_{it} + \beta_{yy}(\log y_{it})^2 \\ & + \beta_{NN}(\log N_{it})^2 + \sum_j \sum_k \beta_{jk} \log w(j)_{it} \log w(k)_{it} + \sum_j \beta_{jy} \log w(j)_{it} \log y_{it} \\ & + \sum_j \beta_{jN} \log w(j)_{it} \log N_{it} + \beta_{yN} \log y_{it} \log N_{it} + \delta_t + \omega_{it} + \epsilon_{it}, \quad j, k = 1, ...K\end{aligned} \tag{3.3}$$

where y is a measure of output, $\mathbf{w}$ is a vector of prices for 'K' variable inputs, N is network size that represents capital stock, ω is the unobserved productivity level of the firm, δ_t are year dummies that capture the year-specific effects on productivity, ϵ is a normally distributed idiosyncratic error term or in other words, all random shocks to the short-run

operating costs. We normalise each variable by its mean value to obtain an approximation of a firm's cost structure around mean production level (Friedlaender & Spady 1981).

It is required that the cost function be homogeneous of degree one and concave in variable factor prices (McFadden 1978). The following restrictions are imposed on the parameters to ensure linear homogeneity in factor prices:

$$\begin{aligned} &\sum_{i=1}^{K} \alpha_i = 1, \quad \beta_{ij} = \beta_{ji} \quad \forall i, j \\ &\sum_{i=1}^{K} \beta_{ij} = \sum_{i=1}^{K} \beta_{ji} = \sum_{i=1}^{K} \beta_{iy} = \sum_{i=1}^{K} \beta_{iN} = 0 \end{aligned} \tag{3.4}$$

Concavity in factor prices is ensured if the sub-matrix of the bordered Hessian matrix of the cost function $CV^s(y, \mathbf{w}, N, \omega)$ corresponding to the factor prices has non-positive eigenvalues at each observation.

The cost function is also required to have positive marginal costs, given by,

$$\begin{aligned} \frac{\partial CV^s(y, \mathbf{w}, N, \omega)}{\partial y} &= \frac{\partial \log CV^s(y, \mathbf{w}, N, \omega)}{\partial \log y} \frac{CV^s}{y} \\ &= \alpha_y + 2\beta_{yy}(\log y) + \sum_j \beta_{jy} \log w(j) + \beta_{yN} \log N, \quad j = 1, ...K \end{aligned} \tag{3.5}$$

The conditional factor share equations for input i can be derived directly by partially differentiating the cost function equation 3.3 with respect to the price of input i and using Shepherd's lemma:

$$\begin{aligned}\frac{\partial \log CV^s(y,\mathbf{w},N,\omega)}{\partial \log w_i} &= \frac{\partial CV^s(y,\mathbf{w},N,\omega)}{\partial w_i}\frac{w_i}{CV^s(y,\mathbf{w},N,\omega)} = \frac{w_i x_i}{CV^s(y,\mathbf{w},N,\omega)} \\ &= s_i(y,\mathbf{w},N,\omega) \\ &= \alpha_i + \sum_{j=1}^{K}\beta_{ij}\log w_j + \beta_{iy}\log y + \beta_{iN}\log N + \xi_i; j = 1,...K.\end{aligned} \tag{3.6}$$

We calculate short-run estimates of returns to density (RTD) and returns to scale (RTS) as follows:

$$\begin{aligned}RTD &= \left(\frac{\partial \log CV^s}{\partial \log y}\right)^{-1} \\ &= (\alpha_y + 2\beta_{yy}\log y + \beta_{yN}\log N + \sum_i \beta_{iy}\log w_i)^{-1} \\ RTS &= \left(\frac{\partial \log CV^s}{\partial \log y} + \frac{\partial \log CV^s}{\partial \log N}\right)^{-1} \\ &= (\alpha_y + 2\beta_{yy}\log y + \beta_{yN}\log N + \sum_i \beta_{iy}\log w_i \\ &\quad + \alpha_N + 2\beta_{NN}\log N + \beta_{yN}\log y + \sum_i \beta_{iN}\log w_i)^{-1}\end{aligned} \tag{3.7}$$

RTD and RTS estimates greater than (less than) one implies increasing (decreasing) returns to scale and density respectively.

Using equation 3.6, we also calculate the own price elasticities of demand for each input, represented by e_i, and the Allen-Uzawa partial elasticities of substitution between different inputs, given by σ_{ij}. These elasticities are defined as follows:

$$\begin{aligned}e_i &= \frac{\beta_{ii}}{s_i} + s_i - 1, \quad i = 1,...K. \\ \sigma_{ij} &= \frac{\beta_{ij}}{s_i s_j} + 1, \quad i,j = 1,...K, \quad i \neq j.\end{aligned} \tag{3.8}$$

If σ_{ij} is less than(greater than) zero, factors i and j are complements (substitutes).

3.4.3 Econometric estimation

We apply four commonly used panel methods from most to least restrictive: (i) pooled ordinary least squares (POLS), (ii) fixed effects (FE), (iii) instrumental variables (IV), and (iv) dynamic panel generalised methods of moments (DPGMM).

Pooled Ordinary Least Squares (POLS) Estimation

The traditional POLS estimation assumes constant coefficients and uses the Gauss-Markov conditions. In this approach, ω_{it} is left in the error term. The observations are pooled across i and t and OLS estimation is applied.

$$CV_{it}^{s} = \gamma X_{it}' + \delta_t + \nu_{it}, \quad where \quad \nu_{it} = \omega_{it} + \epsilon_{it}. \tag{3.9}$$

where X_{it} denotes the set of covariates in the short-run cost model.

Consistency of this estimator requires the contemporaneous exogeneity assumptions (a) $Cov(X_{it}, \omega_{it}) = 0$ and (b) $Cov(X_{it}, \epsilon_{it}) = 0$ for all $t = 1, \ldots, T$. The former assumption is highly restrictive as it requires that the level of input prices and output for any firm are independent of its productivity (refer section 3.2).

Fixed Effects (FE) Estimation

This approach offers the treatment of firm level time-invariant heterogeneity by inclusion of a fixed firm-specific anticipated component of productivity as $\omega_{it} = \omega_i$. FE estimates are obtained by applying OLS estimation to the time-demeaned form or within transformation of Equation 3.14.

$$\ddot{CV}_{it}^{s} = \gamma \ddot{X}_{it}' + \delta_t + \ddot{\epsilon}_{it}, \tag{3.10}$$

where $\ddot{CV}_{it}^{s} = CV_{it}^{s} - \bar{CV}_{i}^{s}$ and so on.

For consistency of this estimator, strict exogeneity assumption on the covariates $\{X_{it}\}$ is required, that is, $Cov(X_{is}, \epsilon_{it}) = 0$ for all $s, t = 1, 2, \ldots, T$. So situations where shocks today affect future decisions about the covariates are ruled out. This assumption is unrealistic as it is quite feasible for metro operators to set their input prices in response to their input requirements and productivity and future decisions on the share of factor inputs employed in the production of a certain level of output are determined by present shocks to productivity. For instance, these shocks to productivity may include any technological innovation in the past year that reduces labour share in the production process (Rios-Rull & Santaeulalia-Llopis 2010). Moreover, as noted by Caves et al. (1987), time-demeaning of covariates yields unreasonably low coefficients of the fixed factor of production and thus biases the estimates of scale economies.

Instrumental Variables

To allow for correlations between the covariates of our cost model and shocks to the short-run operating cost, we use a vector of time-varying instrumental variables (IVs), given by Z_{it} that are uncorrelated with the idiosyncratic errors ϵ_{it} and strongly correlated with the covariate vector X_{it}. Estimates are obtained by first applying first-differencing to eliminate time-invariant heterogeneity as in equation 3.11, followed by IV estimation. This again requires strict exogeneity of IVs, that is, $Cov(Z_{is}, \epsilon_{it}) = 0$; for all $s, t = 1, 2, \ldots, T$, for consistency.

$$\Delta CV_{it}^{s} = \gamma \Delta X_{it}' + \delta_t + \Delta \epsilon_{it}, \tag{3.11}$$

where $\Delta CV_{it}^{s} = CV_{it}^{s} - CV_{i,t-1}^{s}$ and so on.

In the case of a metro cost function, one may think of potential external instruments such as demographic patterns at the city level for metro output and economic growth the national level for input prices. There are previous studies in the literature, for instance, Savage (1997), that use external instruments like population density for output. Demographic patterns in a city, for instance, population density, does determine the

demand for metro services and in turn its output. However, the fact that these patterns also determine the operational costs of metros due to associated presence of economies of density in metro operations, is something that has at large received undue attention. Therefore, in our opinion, the exogeneity of such external instruments is questionable.

Similarly, instruments for input prices such as economic growth at the national level are again not truly exogenous. Higher economic growth implies increased number of economic opportunities, which perhaps may result in increased demand for metro services and consequently affect the metro output. Again, any change in metro output will possibly affect the unit operational costs of metros due to associated presence of strong economies of density and scale.

In the absence of suitable IVs, suitable instruments can be derived from the panel nature of the dataset. Lagged levels of endogenous covariates can be used as their instruments for differenced equations. In this case, consistency of the estimator relies on the sequential exogeneity assumption that input prices and outputs are chosen before anything is known of ϵ_{it}, that is, $Cov(X_{is}, \epsilon_{it}) = 0$ for all $s \leq t$.[15] Under sequential exogeneity, $\Delta\epsilon_{it} = \epsilon_{it} - \epsilon_{i,t-1}$ is uncorrelated with the past history of the covariates, $X^o_{i,t-1} \equiv (X_{i1}, X_{i2}, ..., X_{i,t-1})$. This generates the moment conditions:

$$E(X^o_{i,t-1}\Delta\epsilon_{it}) = 0, \quad t = 2, ..., T. \tag{3.12}$$

Difference Generalised Method of Moments (GMM) estimation is applied to equation 3.11 using the above moment conditions (Arellano & Bond 1991). To assure that we have strong instruments, it is important that our covariates are not highly persistent over time. System GMM provides an augmented approach to overcome any weak instrument problem. As suggested by Arellano & Bover (1995), additional moment conditions are

[15] One may argue that a metro firm is highly likely to predict the future factors that may impact its productivity, such as future demand and future prices of inputs. However, these predictions correspond to future values of covariates in the cost model. All random shocks to operating costs, represented by ϵ, in the future years cannot be predicted by the firm and are thus unobserved. Therefore, the sequential exogeneity assumption is sufficient and lagged levels of endogenous covariates are consistent to achieve identification.

generated for estimation by adding lagged first differences of covariates as instruments in the levels equation (3.9):

$$E[\Delta X'_{it}(CV^s_{it} - \bar{\omega_i} - \gamma X'_{it}] = 0, \quad t = 2, ..., T. \tag{3.13}$$

where $\bar{\omega_i} = E(\omega_i)$.

Consistency of the GMM estimators described above depends on two crucial assumptions: (i) there should be no first order serial-autocorrelation in the error term of the levels equation (3.9), and (ii) the instruments should be exogenous. Two tests are available to evaluate these assumptions. The Arellano and Bond test (Arellano & Bond 1991) evaluates the hypothesis that there is no second-order serial correlation in the error term of the first differenced equation (3.11). This implies that the errors from the levels equation (3.9) are serially uncorrelated. Deeper time lags are tested when serial correlation exists. Validity of instruments is tested using the Sargan/Hansen test of over-identifying restrictions.

It is worth emphasising here that we apply GMM estimation to a translog cost model. For endogenous covariates, the second order interaction terms will also be endogenous. So we need to instrument all endogenous covariates and their second order interactions with other covariates. This will result in a large set of instruments for a translog cost model with endogenous covariates.

Dynamic Panel Generalised Methods of Moments (DPGMM)

Time-demeaning and first-differences operations mentioned previously lead to complete elimination of the cross-sectional variation in time-invariant covariates, resulting into a downward bias in the estimated parameter for the fixed factor of production. To overcome this problem, we use a dynamic panel model for metro costs that allows us to investigate firm dynamics and adjustments in behaviour conditional on costs in the previous time period.

$$CV^s_{it} = \rho CV^s_{i,t-1} + \gamma X'_{it} + \delta_t + \omega_{it} + \epsilon_{it}, \tag{3.14}$$

This model is in accordance with Blundell & Bond (2000) where the time invariant nature of productivity is relaxed by allowing the firm to react to previous shocks to its productivity, that is, by decomposing the productivity term into a fixed effect and an auto-regressive AR(1) component: $\omega_{it} = \rho\omega_{i,t-1} + \xi_{it}$. From equation 3.14, we have the following levels and first-differenced equations:

$$\begin{aligned} CV^s_{it} &= \rho CV^s_{i,t-1} + \gamma X'_{it} + \delta_t + \omega_{it} + \epsilon_{it}, \quad t = 1, ..., T. \\ \Delta CV^s_{it} &= \rho \Delta CV^s_{i,t-1} + \gamma \Delta X'_{it} + \delta_t + \Delta\epsilon_{it}, \quad t = 2, ..., T. \end{aligned} \tag{3.15}$$

The minimal assumptions imposed are that the dynamics are first order:

$$E(CV^s_{is}\epsilon_{it}) = 0; \quad s = 0, 1, \ldots, t-1; \quad t = 1, 2, \ldots, T. \tag{3.16}$$

Parameter estimates are derived by applying the difference GMM or system GMM estimation based on whether our covariates are highly persistent over time or not.

3.5 Simulations

The properties of panel estimators are well understood. In this brief section we demonstrate the problem of endogeneity and the potential of different panel estimators in the specific setting of cost function estimation with endogenous productivity and the resulting estimates of economies of scale and density. For demonstration, we use a Cobb-Douglas cost function but our conclusions are equally applicable for a more flexible specification like translog.

Our simulations are conducted on samples of 1000 observations comprising of 100 firms each observed over 10 years. We index firms by i, $i = (1, ..., N)$ and time points by t, $t = (1, ..., n_i)$ giving a total of $n = \sum_{i=1}^{N} n_i$ observations. The model set-up follows a Cobb-Douglas AR(1) cost function structure with the covariates mentioned in Table 3.2:

$$logCV_{it} = \alpha_y logy_{it} + \alpha_l logl_{it} \sum_j \alpha_j logw_{j,it} + \alpha_N logN_{it} + \omega_{it} + \mu_{it}, \quad j = 1, ..., 4 \quad (3.17)$$

The variables y_{it}, l_{it}, N_{it}, $\mathbf{w_{it}}$, ω_{it} and μ_{it} represent the primary output variable, load factor, network length, vector of factor prices, unobserved productivity and random shocks to productivity respectively for firm i at time t. We introduce three possible sources of confounding into the model: a relationship between the lagged dependent variables and the current independent variable, serial correlation in error and positive correlations between the unobserved productivity ω_{it} and the endogenous independent variables. We model the development of all the independent variables in their logarithmic form as given in equation 3.18. Serial correlation is introduced by including an auto-correlated shock μ_{it} that is independent but exhibits the same variance across the sample as given in equation 3.19. The parameter ω_{it} represents the unobserved-time variant effect that is positively correlated with all the endogenous regressors. We assume that the analyst is ignorant of its presence in the true data generating process for the dependent variable.

$$logX_{it} = \rho logX_{i,t-1} + \gamma logY_{i,t-1} + \epsilon_{it} \quad (3.18)$$

where X_{it} denotes the set of covariates in equation 3.17.

$$log\omega_{it} = \rho log\omega_{i,t-1} + \eta_{it} \quad (3.19)$$

with $\epsilon_{it} \sim N(0,1)$, $\eta_{it} \sim N(0,1)$ and $\omega_{i,1} \sim N(0,1)$.

The parameters ρ and γ are set to 0.7 and 0.07 respectively. The chosen value of α parameters of Equation 3.17 are listed in Table 3.3 under the column α_{true}. We generate a panel with a length of 15 observations for each unit, and subsequently ignore the first 5 observations for the estimation of parameters. We apply the different methodologies detailed in Section 3.4.3 to estimate the coefficients of Equation 3.17.

Table 3.3 reports the estimated coefficients, their associated standard errors and root mean squared errors (RMSE). The results from static panel methods exhibit an upward bias in the estimated coefficients of the output and network length variables. Thus the resulting estimates of RTD and RTS from these models have a downward bias. This confirms our hypothesis that with the erroneous omission of the unobserved productivity confounder ω_{it}, the resulting estimates of scale economies from a cost model have a downward bias. The inclusion of fixed effects in a cost model fails to adjust for this source of confounding, since by construction, the fixed effects are independent of the covariates and therefore only capture time-invariant unobserved variables. Dynamic panel methods deliver parameter and RTS and RTD estimates that indicate relatively small biases and root mean squared errors. This is because they can control for both the endogeneity of covariates resulting from the erroneous omission of ω_{it} and the inherent dynamics of production process as explained in Section 3.4.3. Thus they offer enormous flexibility in approximating the dynamics of the production process while providing adequate adjustment for confounding covariates.

3.6 Results and Discussion

3.6.1 Estimation results of the cost model

We apply the panel data estimators discussed in section 3.4.3 to estimate equation 3.3. The full table of results, Table A.1, is attached in Appendix A.2. Using DPGMM estimation, we find both increasing returns to scale and density as compared to POLS estimates that reveal constant returns to scale (refer to Table A.2 in Appendix A.2). It is evident that there is a downward bias in RTD and RTS estimates if the dynamics of firm-level productivity and the endogeneity in covariates remain unaccounted for. These results are consistent with our conclusions from the simulation analysis presented in Section 3.5.

Table 3.3: Simulation Results for a Cobb-Douglas cost function using different estimation methodologies.

(a) Simulation Results from Static Panel Methods.

		POLS			FE			IV: Diff GMM			IV: Sys GMM		
Coef.	α_{true}	$\bar{\alpha}$	σ	RMSE	$\bar{\alpha}$	σ	RMSE	$\bar{\alpha}$	σ	RMSE	$\bar{\alpha}$	σ	RMSE
α_y	0.640	1.212	0.043	0.574	0.858	0.040	0.222	0.713	0.085	0.112	1.113	0.081	0.480
α_N	0.177	0.674	0.045	0.499	0.398	0.040	0.224	0.277	0.071	0.122	0.500	0.074	0.331
α_l	0.298	0.827	0.042	0.531	0.580	0.038	0.284	0.455	0.078	0.176	0.845	0.070	0.551
α_1	0.485	0.507	0.038	0.044	0.537	0.032	0.061	0.526	0.062	0.074	0.528	0.061	0.075
α_2	0.208	0.201	0.039	0.040	0.204	0.033	0.033	0.257	0.058	0.076	0.208	0.057	0.057
α_3	0.307	0.292	0.039	0.042	0.259	0.033	0.058	0.217	0.067	0.112	0.265	0.066	0.078
ρ	0.700												
RTD	1.563	0.825	0.029	0.738	1.166	0.054	0.401	1.402	0.166	0.231	0.898	0.065	0.667
RTS	1.224	0.530	0.015	0.694	0.796	0.032	0.429	1.010	0.118	0.244	0.620	0.043	0.606

(b) Simulation Results from Dynamic Panel Methods.

		AR(1): Diff GMM			AR(1): Sys GMM		
Coef.	α_{true}	$\bar{\alpha}$	σ	RMSE	$\bar{\alpha}$	σ	RMSE
α_y	0.640	0.485	0.065	0.168	0.648	0.056	0.057
α_N	0.177	0.138	0.057	0.069	0.241	0.054	0.083
α_l	0.298	0.235	0.057	0.085	0.395	0.050	0.109
α_1	0.485	0.476	0.039	0.040	0.436	0.039	0.062
α_2	0.208	0.195	0.043	0.045	0.116	0.040	0.100
α_3	0.307	0.328	0.045	0.050	0.448	0.044	0.147
ρ	0.700	0.353	0.033	0.348	0.609	0.024	0.094
RTD	1.563	2.062	0.276	0.571	1.543	0.134	0.135
RTS	1.224	1.606	0.224	0.443	1.125	0.088	0.132

We discuss our results from the DPGMM estimation (see Table 3.4) in this section. The reduced form regressions ΔX_{it} on the past history of covariates $X^{o}_{i,t-1} \equiv (X_{i,1}, X_{i,2}, ...X_{i,t-1})$ show low levels of correlation and suggests that there is a weak instrument problem. So, we use system GMM estimation. Results of Arellano-Bond tests for AR(1) and AR(2) and Sargan test of over-identifying restriction are also reported in Table 3.4. The Sargan test confirms that the use of GMM estimation is consistent and the Arellano-Bond tests confirm that the instruments are relevant. We end up with a large number of instruments (as reported in Table 3.4) because most of the covariates and their second order interactions in our translog specification are endogenous. The signs of estimated coefficients are consistent with economic theory, although low significance levels can be attributed to fewer observations or increase in the error variance in the IV estimation.

The estimated cost function has positive marginal costs for all observations. We also find that the estimated function is concave in factor prices for 108 out of 165 observations[16] [17]. The exceptions relate to observations where eigenvalues for labour price are significantly greater than zero at the 95 percent confidence level. One way to deal with non-concavity is to impose local concavity using the method suggested in Ryana & Wales (2000). However, Ogawa (2011) shows that imposing any concavity condition mis-specifies the cost model when firms are incapable of minimising their production costs due to extraneous circumstances. For instance, if a firm has many quasi-fixed inputs in the short-run, then it incurs additional costs such as those related to readjustment of

[16]To check whether the estimated cost function is concave in factor prices, we perform a test of the hypothesis that the sub-matrix of the bordered Hessian matrix corresponding to factor prices has non-positive eigenvalues. We obtain the eigenvalues corresponding to each factor and their associated standard errors via bootstrapping.

[17]Savage (1997) and Mizutani (2004) adopt a translog specification of the cost function and the estimated cost function is concave in factor prices at 90 percent of the observations in their studies. However, their specifications do not consider any second order interactions of factor prices with output, that is, they assume that the underlying production technology is homothetic, which is a major limitation of their work. When Ogawa (2011) and Karlaftis and McCarthy (2002) used a fully flexible translog specification, the estimated cost functions were concave at around 40 percent and 70 percent observations respectively.

their organisational structure and the relocation of employees (Pindyck & Rotemberg 1983). Therefore, imposing concavity yields inconsistent estimates of the cost function parameters and leads to biased conclusions about the underlying production technology of the firm. Pindyck & Rotemberg (1983) and Ogawa (2011) suggest not to impose concavity, rather missing variables representing these additional costs that may be correlated with the factor prices should be controlled to obtain unbiased parameter estimates of a cost function. We believe that with the treatment of factor prices as endogenous covariates and the use of suitable instruments, we adjust for potential biases in the cost function estimation that may occur due to such missing variables.

3.6.2 Properties of the underlying production technology

Using the estimated cost function, we test the hypothesis of linear input factor separability and homotheticity of the underlying production technology. Factor separability implies that the marginal rate of technical substitution between two inputs is invariant with the prices of other inputs. This requires that the second order interactions of factor prices are zero, that is, $\beta_{ij} = 0$ for all i, j. At the 95% confidence level, factor separability was rejected.

Homotheticity implies that the marginal rate of technical substitution between factor inputs is homogeneous of degree zero, that is, the factor shares are invariant with the firm size. A necessary and sufficient condition for homotheticity of the underlying production technology is linear homogeneity in factor prices (equation 3.4) combined with a restriction that the second order interactions between factor price and output variables are zero, that is, $\beta_{iy} = 0$ for all i. At the 95% confidence level, some second order interaction terms between the primary and secondary measures of output load factor and factor prices are significantly different from zero. Non-homotheticity implies that changes in factor price will affect both the cost elasticity and corresponding factor demand. Therefore, the scale economies are not independent of the input prices used in the production process.

Table 3.4: Estimates of the cost function parameters and associated robust standard errors.

Explanatory variable	Estimate	Std. Error
Car kms	0.640***	0.116
Network length	0.177*	0.097
Load Factor	0.298*	0.160
Labour Price	0.485***	0.048
Energy Price	0.208***	0.046
Residual Price	0.307***	0.056
Car kms ^{2}	0.390	0.299
Network length ^{2}	0.322	0.288
Load Factor^2	0.376	0.278
Labour Price ^{2}	0.140*	0.083
Energy Price ^{2}	0.162***	0.060
Residual Price ^{2}	0.011	0.032
Car kms x Network length	-0.640	0.579
Car kms x Load Factor	0.244	0.410
Car kms x Labour Price	0.028	0.159
Car kms x Energy Price	-0.200	0.158
Car kms x Residual Price	0.172**	0.076
Network length x Load Factor	-0.628	0.448
Network length x Labour Price	-0.210	0.157
Network length x Energy Price	0.340**	0.147
Network length x Residual Price	-0.130	0.103
Load Factor x Labour Price	0.506***	0.188
Load Factor x Energy Price	-0.459***	0.156
Load Factor x Residual Price	-0.047	0.092
Labour Price x Energy Price	-0.145**	0.066
Labour Price x Residual Price	0.005	0.049
Energy Price x Residual Price	-0.016	0.046
Lag (Dependent Variable)	0.196***	0.064
Year Effects Included	YES	
No. of Observations	119	
No. of Instruments	139	
Arellano-Bond test for AR(1)	z = -1.86	Pr > z = 0.052
Arellano-Bond test for AR(2)	z = 1.21	Pr > z = 0.226
Sargan Test of over-identifying restrictions:	$\chi^2(103) = 412.14$	Pr > χ^2 = 0.000

Notes:

(1) All explanatory variables are in their logarithmic form except for dummy variables.

(2) Significance: (***) 99 percent, (**) 95 percent, (*) 90 percent.

(3) Standard Errors reported are robust standard errors.

Table 3.5 reports the own price elasticities of demand for factor inputs and the partial elasticities of substitution between different inputs (refer equation 3.8). Due to low significance levels of estimated elasticities, it is difficult to make definite conclusions regarding firm behaviour.

Table 3.5: Price elasticities of factor demand and elasticities of substitution

Elasticity	Estimate	Standard Error	95% Confidence Interval	
e_1	-0.226	0.176	-0.570	0.119
e_2	-0.014	0.295	-0.593	0.564
e_3	-5.170***	1.303	-7.724	-2.617
σ_{12}	-0.442	0.664	-1.743	0.858
σ_{13}	1.035***	0.332	0.384	1.686
σ_{23}	0.746	0.708	-0.642	2.135

(i) e: own-price elasticity of demand, σ: Allen partial elasticity of substitution.

(ii) notations: 1 = labour, 2 = energy, 3 = residual.

(iii) Significance: (***) 99 percent, (**) 95 percent, (*) 90 percent.

3.6.3 Economies of density and scale

Table 3.6 reports the estimates of returns to density (RTD) and scale (RTS) calculated using equation 3.7.

At the sample mean, the estimated value of RTD is 1.562 that is statistically greater than unity and implies substantial economies of density. We find that, on an average, if the use of factors associated with density increases by 10%, the average cost of metro operations increases by 6.4% only. This implies that there are significant cost advantages associated with a more intense use of a fixed network achieved by increasing operated car kilometres. Thus systems with higher density of operations face lower unit costs. There are a number of fixed and semi-fixed costs in urban rail operations, for instance, station staffing, which result into economies of density. Our finding on existence of RTD in urban rail transport operations is consistent with the literature. The magnitude of

RTD estimates from previous studies vary between 1.2 to 1.5, with the average value being around 1.4. Our POLS estimate of 1.227 (refer Table A.1 in Appendix A.2) lies in this range. Thus, the estimates in the literature clearly have a downward bias due to the reasons mentioned previously.

Table 3.6: Summary of RTD and RTS estimates.

	Estimate	Std. Error	95% Confidence Interval	
RTD	1.562	0.283	1.006	2.117
RTS	1.223	0.081	1.065	1.381

However, as demand is usually inelastic in car-kilometres, an increase in car-kilometres will depress the load factor variable and this will consequently lead to an increase in the magnitude of the economies of density (Jara-Diaz & Cortes 1996). At mean values of all variables, our estimated cost model indicates that the elasticity of operational costs with respect to load factor is 0.298. Using the value of average elasticity of transit use with respect to transit service frequency, also known as headway elasticity as reported in Litman (2004), which equals 0.5, the adjusted values of economies of density will be::

$$\left(\frac{1}{1.562} - \frac{0.5}{2} * 0.298\right)^{-1} = 1.768$$

We also find that there are economies of scale in urban rail transport operations. Our estimate of RTS is 1.223, which is statistically greater than one. Thus, on an average, if a metro firm expands its network size and output proportionally by 10% , the average costs of operations only increase by 8.18%. Therefore, there are cost savings associated with expansion of a metro service. The literature mostly suggests existence of constant returns to scale in provision of metro services. Our POLS estimate of RTS is 1.045 (refer Table A.1 in Appendix A.2) that is not statistically different from unity. Downward bias in the RTS estimates from the literature due to reasons previously discussed is once again evident. Moreover, our analysis includes track maintenance costs as a component of

variable costs in the short-run. Scale economies in provision of metro services may result from the presence cost complementarities between operational costs and way and track maintenance cost components, as in case of mainline railways (Bitzan 2000, 2003, Ivaldi & McCullough 2007).

While past researchers rightly suggest that metro firms may expand their networks due to extraneous reasons such as to serve more areas in a city in response to rising travel demand, to justify raising subsidy funds for a citywide tax and economic development, among others, our estimate suggests that there may be other factors in addition to the above that may result into actual cost advantages for metros when they expand their networks. For instance, as Graham et al. (2003) suggest, in cases where sufficient latent demand exists, a metro firm may try to exploit the density economies by expanding the network. In addition, as Wei & Hansen (2003) note, when traffic increases significantly, crowding may drive costs up if frequency is increased. In such cases, network expansions may result into network wide savings in costs for a metro firm.

Figure 3.3 shows the variation of returns to scale estimates over a range of output (in car kilometres). We see that RTS estimates decrease approximately over output, where very large output levels approximately correspond to constant RTS, that is, equal to one. This decrease could be because very large metro systems may become less efficient in dealing with their time-varying fleet and staff requirements.

3.7 Understanding the Variation in Unit Costs across Metro Systems

Figure 3.1, shows that unit costs of metro operations $\left(\frac{C}{y}\right)_i$ vary substantially across metro systems. This variation results from two main sources: (1) intrinsic performance of the system, P_i , and (2) exogenous variation, $\sum_j \theta_j.S_i(j)$ from sources $S(j)s$. The adjusted costs are given by the following equation, where the left hand side represents the adjusted

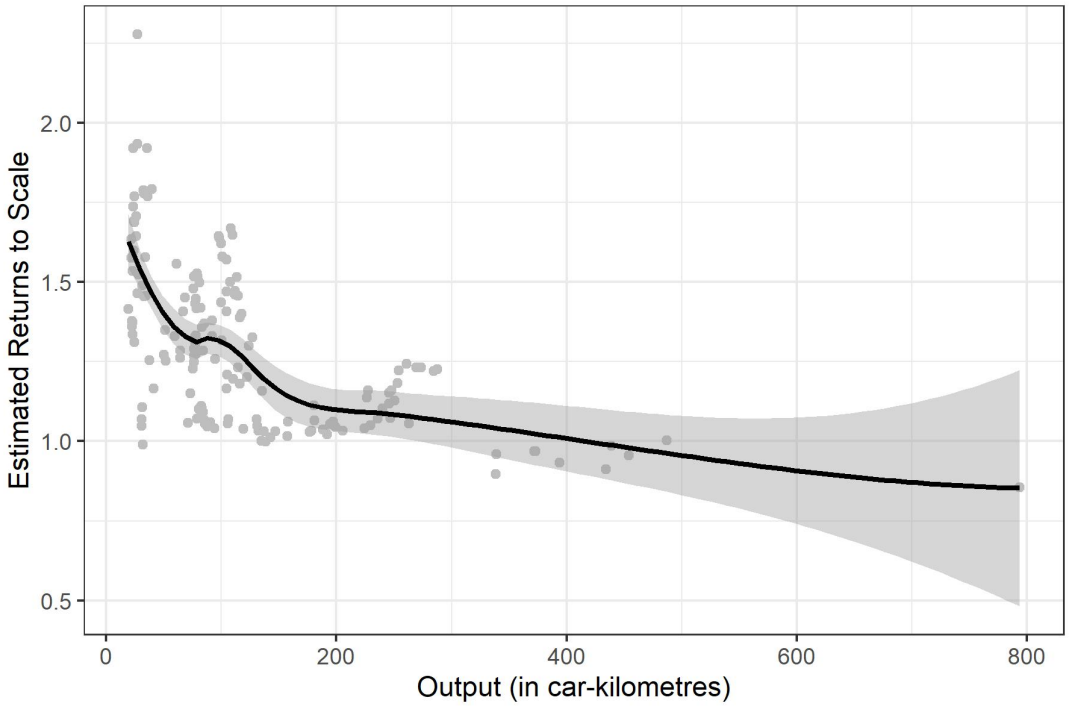


Figure 3.3: Variation of Returns to Scale Estimates over Output.

costs:

$$\left(\frac{C}{y}\right)_i = P_i + \sum_j \theta_j . S_i(j) + e_i \tag{3.20}$$

To understand the variation in costs resulting from intrinsic differences in performance across metro operations, it is important to adjust for any external variation. The exogenous variation can be absorbed by adjusting for differences between $S_i(j)s$ across various systems and constraining them to their mean values $\overline{S_i(j)}$ to ensure comparability, as given by the following equation:

$$\left(\frac{C}{y}\right)_i + \sum_j \theta_j . (\overline{S_i(j)} - S_i(j)) = P_i + \sum_j \theta_j . \overline{S_i(j)} + e_i \tag{3.21}$$

For metro operations, increasing returns to density and network size in metro operations are external benefits. In this brief section, we adjust the unit costs in figure 3.1 for exogenous variation resulting from differences in density and network size. We use both

our scale economy estimates and those from the literature. This adjustment helps us to assess whether scale economy estimates from this analysis are able to explain a greater extent of exogenous influence in unit costs as compared to the estimates from the literature.

Based on equation 3.3, we can represent the operational costs of metro i, C_i as a function of its network size, N_i, its output (car-kilometres) , Y_i, and a function $g(x)_i$ that includes other covariates like factor prices, load factor and so on, as given in equation 3.22. The elasticities of cost with respect to network size and output are θ_1 and θ_2.

$$C_i = g(x)_i N_i^{\theta_1} Y_i^{\theta_2} \tag{3.22}$$

From equation 3.22, unit costs of metro operations $\left(\frac{C}{Y}\right)_i$ can be given by:

$$\begin{aligned}
&\left(\frac{C}{Y}\right)_i = g(x)_i N_i^{\theta_1} Y_i^{(\theta_2-1)} \\
\implies &\left(\frac{C}{Y}\right)_i = g(x)_i N_i^{(\theta_1+\theta_2-1)} \left(\frac{Y}{N}\right)_i^{(\theta_2-1)} \\
\implies &\left(\frac{C}{Y}\right)_i = g(x)_i N_i^{\left(\frac{1}{RTS}-1\right)} \left(\frac{Y}{N}\right)_i^{\left(\frac{1}{RTD}-1\right)} \\
\implies &\log\left(\frac{C}{Y}\right)_i = \log g(x)_i + \left(\frac{1}{RTS}-1\right)\log N_i + \left(\frac{1}{RTD}-1\right)\log\left(\frac{Y}{N}\right)_i
\end{aligned} \tag{3.23}$$

where,
$RTD = \left(\frac{\partial \log C_i}{\partial \log Y_i}\right)^{-1} = \frac{1}{\theta 2}$ and $RTS = \left(\frac{\partial \log C_i}{\partial \log Y_i} + \frac{\partial \log C_i}{\partial \log N_i}\right)^{-1} = \frac{1}{\theta 1+\theta 2}$.

We convert the unit cost equation to its equivalent logarithmic form because our θ_1 and θ_2 values represent elasticities rather than the absolute effect.

At this point, it is worth re-emphasising that the technology underlying production of metro output is non-homothetic and the input factors are not linearly separable (refer Section 3.6.2. So, our scale estimates also capture the part of variation in input factor prices and other covariates that results from differences in network size and density of operations.

Adjusting for differences in network size and density of operations in equation 3.23, we have:

$$\log\left(\frac{C}{Y}\right)_i + \left(\frac{1}{RTS} - 1\right)(\overline{\log N_i} - \log N_i) + \left(\frac{1}{RTD} - 1\right)\left(\overline{\log\left(\frac{Y}{N}\right)_i} - \log\left(\frac{Y}{N}\right)_i\right) = \log g(x)_i + \left(\frac{1}{RTS} - 1\right)\overline{\log N_i} + \left(\frac{1}{RTD} - 1\right)\overline{\log\left(\frac{Y}{N}\right)_i} + e_i \tag{3.24}$$

Figure 3.4 shows the variation in unit operational costs in 2015 for the same group of thirty-two metro systems as in figure 3.1. The figure also shows the variation in unit operational costs adjusted using our scale economy estimates, that is, RTS = 1.223 and RTD = 1.562, represented by adjusted unit costs (1) and those adjusted using the scale economy estimates from the literature, that is, RTS = 1.000 and RTD = 1.400, represented by adjusted unit costs (2). Table 3.7 presents the mean, variance and coefficient of variation of unadjusted and adjusted unit operational costs.

Table 3.7: Mean and Standard Deviations of Unadjusted and Adjusted Unit Operational Costs.

	Mean	Std. Error	Coefficient of Variation
Unadjusted Unit Operational Costs (in US$)	5.765	6.206	0.432
Unit Operational Costs adjusted using estimates from this analysis (in US$)	5.800	4.615	0.370
Unit Operational Costs adjusted using estimates from the literature (in US$)	5.774	5.311	0.399

From table 3.7 and figure 3.4, we find that our scale economy estimates are able to explain a greater deal of variation in unit costs of metro operations as compared to the estimates from the literature. Thus, the methodological improvement in the estimation of a metro cost function demonstrated in this chapter is worthwhile and the resulting estimates from this analysis are more representative in explaining the industry-wide exogenous variation in unit costs of operations resulting from differences in network size and density of operations across metro systems.

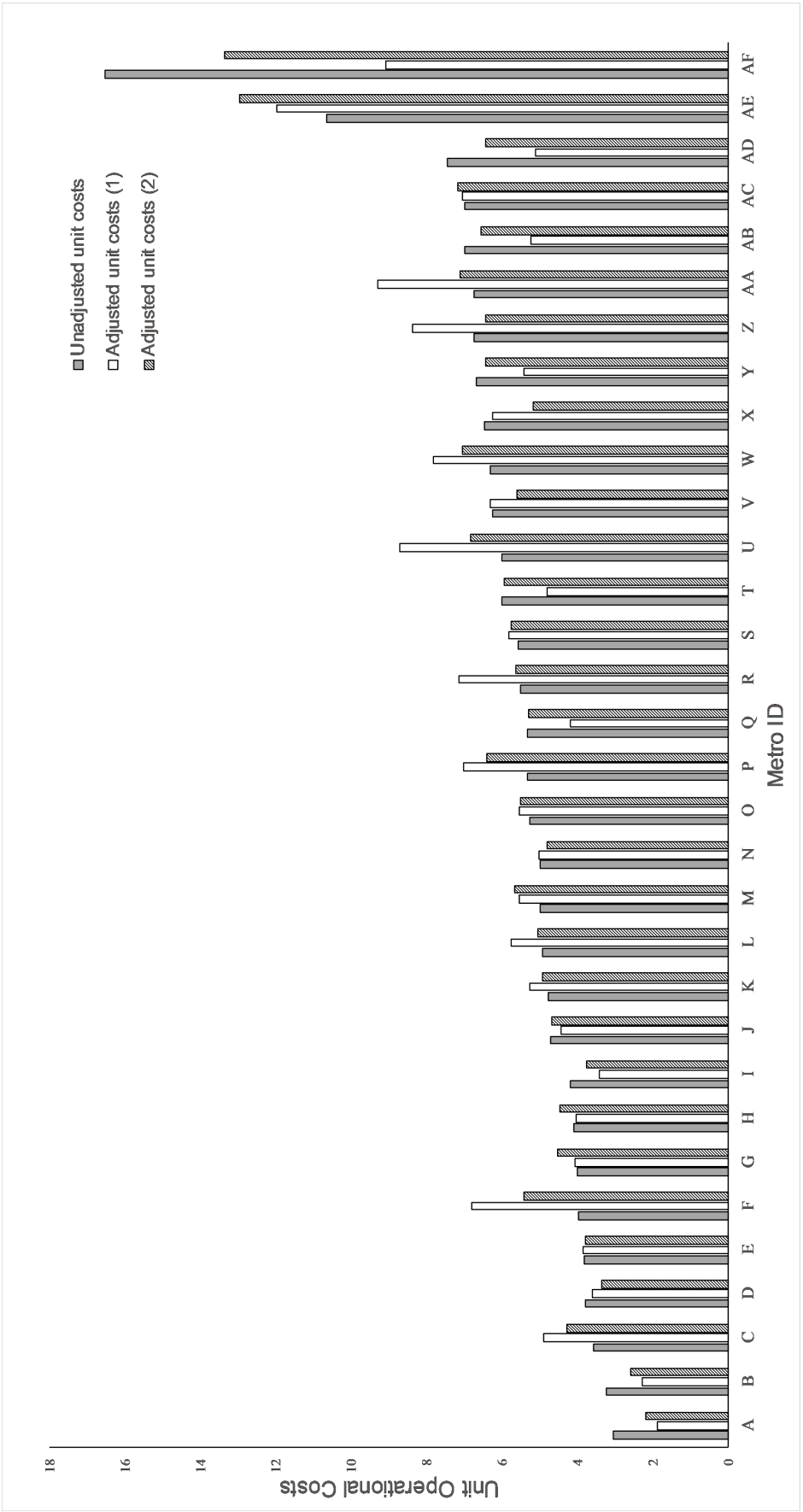


Figure 3.4: Variation of unit operational costs across various metro systems.

3.8 Conclusions and Policy Implications

This chapter has developed a comprehensive understanding of operational costs of urban rail transport and determined the important aspects of the technology that drives unit cost differences between metro firms. We use dynamic panel generalised method of moments (DPGMM) with a very high quality panel dataset on urban rail transport operations to estimate the underlying cost function. This chapter offers the key methodological insight on controlling for observed and unobserved time-invariant and time-variant firm level sources of confounding in the estimation of a transport cost function. We illustrate that DPGMM is attractive for the cost function estimation because it permits flexible representation of unobserved productivity level differences between firms and offers better remedies for endogenous covariates. The key results that emerge from our analysis are as follows:

1. A comparison of our DPGMM results with the traditional estimation methods like pooled ordinary least squares estimation confirms that failure to account for unobserved productivity differences between firms in empirical cost analysis creates a downward bias in the estimates of RTS and RTD.

2. Our estimate of RTD is 1.562, which is statistically greater than one. We thus find evidence of increasing RTD that is consistent with literature on urban rail transport costs, although the average RTD estimate from the literature is around 1.40. Increasing RTD results from the existence a range of fixed and semi-fixed costs are prevalent in the urban rail transport industry that do not vary proportionally with output.

3. We find evidence of increasing RTS, which justifies the presence of large size firms in urban rail transport industry. Our RTS estimate is 1.223, which is again statistically greater than one. The weight of evidence in the urban rail transport literature indicates that the industry is characterised by constant RTS. However, we find that

controlling for endogeneity in empirical cost analysis and accounting for dynamics in firm-level productivity gives RTS estimates that is consistent with the observed industry behaviour.

4. The TSC dataset indicates that around eighty-percent of way and structure maintenance costs comprise of labour and electricity costs, which can be varied in the short-run. We, therefore, include infrastructure maintenance costs as a component of variable costs in our short-run operational cost analysis. Increasing returns to scale may have resulted from the presence of cost complementarities between operational and way and track cost components as found in case of mainline railways.

5. We also study other aspects of the underlying production technology. We find that the marginal rate of technical substitution between any two inputs for production of metro output depends on the prices of other inputs, that is, the underlying technology shows non-separability of input factors. Our results also show non-homotheticity implying that changes in factor prices affects both cost elasticity and corresponding factor demand. Therefore, scale economies in provision of urban rail transport services are not independent of input prices.

Thus, by controlling for various sources of endogeneity in the estimation of a short-run variable cost function for urban rail transport industry, we are able to provide more reliable estimates of industry indices for transport investment appraisal and guiding decisions on pricing rules. Our proposed methodology provides a general specification that could be useful in cost analysis in other modes of transportation, whether be mainline railways, bus or airline operations.

We find that metro systems with high density of usage are highly cost efficient. This could be taken as an evidence in support of providing metro services in city centres where high frequency services are required to serve the travel demand. In addition, the presence of network size economies may be relevant from a policy point of view, particularly for

the economic appraisal of large infrastructure projects that lead to network expansion. Returns to network size implies that such investments may generate external benefits in the form of a network-wide reduction in operational costs. It would be interesting to quantify this external benefit and assess whether it could have significant impact on the outcome of traditional cost-benefit analyses. We aim to undertake this analysis in future.

References

Ackerberg, D., Benkard, C. L., Berry, S. & Pakes, A. (2007), Econometric tools for analyzing market outcomes, *in* J. J. Heckman & E. E. Leamer, eds, 'Handbook of econometrics', Vol. 6A, Amsterdam and Boston: Elsevier, North-Holland, pp. 4171–4276.

Arellano, M. & Bond, S. (1991), 'Some tests of specification for panel data: Monte carlo evidence and an application to employment equations', *The review of economic studies* **58**(2), 277–297.

Arellano, M. & Bover, O. (1995), 'Another look at the instrumental variable estimation of error-components models', *Journal of Econometrics* **68**(1), 29–51.

Basso, L. J. & Jara-Diaz, S. R. (2006), 'Distinguishing multiproduct economies of scale from economies of density on a fixed-size transport network', *Networks and Spatial Economics* **6**(2), 149–162.

Basso, L. J., Jara-Díaz, S. R. & II, W. G. W. (2011), Cost functions for transport firms, *in* A. de Palma, R. Lindsey, E. Quinet & R. Vickerman, eds, 'A Handbook of Transport Economics', Edward Elgar Publishing Ltd., Northampton, chapter 12, pp. 273–297.

Basso, L. J. & Jara–Díaz, S. R. (2005), 'Calculation of economies of spatial scope from transport cost functions with aggregate output with an application to the airline industry', *Journal of Transport Economics and Policy (JTEP)* **39**(1), 25–52.

Batarce, M. (2016), 'Estimation of urban bus transit marginal cost without cost data', *Transportation Research Part B: Methodological* **90**, 241–262. ID: 271728.
URL: *http://www.sciencedirect.com/science/article/pii/S0191261516302661*

Batarce, M. & Galilea, P. (2018), 'Cost and fare estimation for the bus transit system of Santiago', *Transport Policy* **64**, 92–101.

Beveren, I. V. (2012), 'Total factor productivity estimation: A practical review', *Journal of economic surveys* **26**(1), 98–128.

Bitzan, J. D. (2000), Railroad cost conditions: implications for policy, report for the Federal Railroad Administration, Technical report, Upper Great Plains Transportation Institute, North Dakota State University.

Bitzan, J. D. (2003), 'Railroad costs and competition: The implications of introducing competition to railroad networks', *Journal of Transport Economics and Policy (JTEP)* **37**(2), 201–225.

Blundell, R. & Bond, S. (2000), 'GMM estimation with persistent panel data: an application to production functions', *Econometric reviews* **19**(3), 321–340.

Borts, G. H. (1960), 'The estimation of rail cost functions', *Econometrica, Journal of the Econometric Society* pp. 108–131.

Caves, D. W., Christensen, L. R. & Swanson, J. A. (1981), 'Productivity growth, scale economies, and capacity utilization in US railroads, 1955-74', *The American Economic Review* **71**(5), 994–1002.

Caves, D. W., Christensen, L. R., Tretheway, M. W. & Windle, R. J. (1987), An assessment of the efficiency effects of US airline deregulation via an international comparison, *in* E. E. Bailey, ed., 'Public Regulation: New Perspectives on Institutions and Policies', MIT Press, Cambridge, pp. 285–320.

Collard-Wexler, A. (2012), Production and cost functions. Unpublished.
URL: *tinyurl. com/ y3rgn5p9*

De-Loecker, J. (2011), 'Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity.', *Econometrica* **79**(5), 1407–1451.

Farsi, M., Filippini, M. & Greene, W. (2005), 'Efficiency measurement in network industries: application to the Swiss railway companies', *Journal of Regulatory Economics* **28**(1), 69–90.

Filippini, M. & Maggi, R. (1992), 'The cost structure of the Swiss private railways', *International Journal of Transport Economics/Rivista internazionale di economia dei trasporti* **19**(3), 307–327.

Friedlaender, A. F. & Spady, R. H. (1981), *Freight transport regulation: Equity, efficiency and competition in the rail and trucking industries*, MIT Press, Cambridge.

Gagnepain, P. & Ivaldi, M. (2002), 'Incentive regulatory policies: The case of public transit systems in France', *The Rand journal of economics* **33**(4), 605–629.

Graham, D. J. (2008), 'Productivity and efficiency in urban railways: Parametric and non-parametric estimates', *Transportation Research Part E: Logistics and Transportation Review* **44**(1), 84–99.

Graham, D. J., Couto, A., Adeney, W. E. & Glaister, S. (2003), 'Economies of scale and density in urban rail transport: effects on productivity', *Transportation Research Part E: Logistics and Transportation Review* **39**(6), 443–458.

Greene, W. H. (1980), 'Maximum likelihood estimation of econometric frontier functions', *Journal of Econometrics* **13**(1), 27–56.

Hensher, D. A., Daniels, R. & Demellow, I. (1995), 'A comparative assessment of the productivity of Australia's public rail systems 1971/72–1991/92', *Journal of Productivity Analysis* **6**(3), 201–223.

Ivaldi, M. & McCullough, G. (2007), 'Railroad pricing and revenue-to-cost margins in the post-staggers era', *Research in Transportation Economics* **20**, 153–178.

Jara-Diaz, S. R. (1982), 'The estimation of transport cost functions: a methodological review', *Transport Reviews* **2**(3), 257–278.

Jara-Diaz, S. R. & Cortes, C. E. (1996), 'On the calculation of scale economies from transport cost functions', *Journal of Transport Economics and Policy* pp. 157–170.

Karlaftis, M. G. & McCarthy, P. (2002), 'Cost structures of public transit systems: a panel data analysis', *Transportation Research Part E: Logistics and Transportation Review* **38**(1), 1–18.

Karlaftis, M. G., McCarthy, P. S. & Sinha, K. C. (1999), 'System size and cost structure of transit industry', *Journal of Transportation Engineering* **125**(3), 208–215.

Katayama, H., Lu, S. & Tybout, J. R. (2009), 'Firm-level productivity studies: illusions and a solution', *International Journal of Industrial Organization* **27**(3), 403–413.

Keeler, T. E. (1974), 'Railroad costs, returns to scale, and excess capacity', *The review of economics and statistics* **56**(2), 201–208.

Levinsohn, J. & Petrin, A. (2003), 'Estimating production functions using inputs to control for unobservables', *The Review of Economic Studies* **70**(2), 317–341.

Litman, T. (2004), 'Transit price elasticities and cross-elasticities', *Journal of Public Transportation* **7**(2), 3.

McFadden, D. (1978), Cost, revenue, and profit functions, *in* M. Fuss & D. McFadden, eds, 'Production Economics: A Dual Approach to Theory and Application', Vol. 1, Amsterdam: North Holland.

McGeehan, H. (1993), 'Railway costs and productivity growth: The case of the Republic of Ireland, 1973-1983', *Journal of Transport Economics and Policy* pp. 19–32.

Mizutani, F. (2004), 'Privately owned railways' cost function, organization size and ownership', *Journal of Regulatory Economics* **25**(3), 297–322.

Ogawa, K. (2011), 'Why are concavity conditions not satisfied in the cost function? the case of Japanese manufacturing firms during the bubble period', *Oxford Bulletin of Economics and Statistics* **73**(4), 556–580.

Olley, G. S. & Pakes, A. (1996), 'The dynamics of productivity in the telecommunications equipment industry', *Econometrica* **64**(6), 1263–1297.
URL: *http://www.jstor.org/stable/2171831*

Oum, T. H. & Waters, W. G. (1996), 'A survey of recent developments in transportation cost function research', *Logistics and Transportation Review* **32**(4), 423–463.

Pindyck, R. S. & Rotemberg, J. J. (1983), 'Dynamic factor demands and the effects of energy price shocks', *The American Economic Review* **73**(5), 1066–1079.

Pozdena, R. J. & Merewitz, L. (1978), 'Estimating cost functions for rail rapid transit properties', *Transportation Research* **12**(2), 73–78.

Rios-Rull, J.-V. & Santaeulalia-Llopis, R. (2010), 'Redistributive shocks and productivity shocks', *Journal of Monetary Economics* **57**(8), 931–948.

Ryana, D. L. & Wales, T. J. (2000), 'Imposing local concavity in the translog and generalized Leontief cost functions', *Economics Letters* **67**(3), 253–260.

Sànchez, P. C. & Villaroyya, J. M. (2000), 'Efficiency, technical change and productivity in the European rail sector: a stochastic frontier approach', *International Journal of Transport Economics/Rivista internazionale di economia dei trasporti* **27**(1), 55–76.

Savage, I. (1997), 'Scale economies in United States rail transit systems', *Transportation Research Part A: Policy and Practice* **31**(6), 459–473.

Viton, P. A. (1981), 'A translog cost function for urban bus transit', *The Journal of Industrial Economics* pp. 287–304.

Wei, W. & Hansen, M. (2003), 'Cost economics of aircraft size', *Journal of Transport Economics and Policy (JTEP)* **37**(2), 279–296.

Wills-Johnson, N. (2010), 'Cost functions for Australia's railways', *Journal of Infrastructure Systems* **17**(1), 1–14.

Wooldridge, J. M. (2009), 'On estimating firm-level production functions using proxy variables to control for unobservables', *Economics Letters* **104**(3), 112–114.

Chapter 4

Revisiting the empirical fundamental relationship of traffic flow for highways using a causal approach

The fundamental relationship of traffic flow is empirically estimated by fitting a regression curve to a cloud of observations of traffic variables. Such estimates, however, may suffer from the confounding/endogeneity bias due to omitted variables such as driving behaviour and weather. To this end, this paper adopts a causal approach to obtain the unbiased estimate of the fundamental flow-density relationship using traffic detector data. In particular, we apply a Bayesian non-parametric spline-based regression approach with instrumental variables to adjust for the aforementioned confounding bias. The proposed approach is benchmarked against standard curve-fitting methods in estimating the flow-density relationship for three highway bottlenecks in the United States. Our empirical results suggest that the saturated (or hypercongested) regime of the estimated flow-density relationship using correlational curve fitting methods may be severely biased, which in turn leads to biased estimates of important traffic control inputs such as capacity and capacity-drop. We emphasise that our causal approach is based on the physical laws of

vehicle movement in a traffic stream as opposed to a demand-supply framework adopted in the economics literature. By doing so, we also aim to conciliate the engineering and economics approaches to this empirical problem. Our results, thus, have important implications both for traffic engineers and transport economists. The core findings of this chapter are under review as:

> Anupriya, Graham, D.J., Hörcher, D. & Bansal, P. (under review). Revisiting the empirical fundamental relationship of traffic flow for highways using a causal econometric approach. in *Transportation Research Part B: Methodological*

4.1 Introduction

The standard engineering relationship between vehicular flow q, that is, the number of vehicles passing a given point per unit time, and density k, that is, the number of vehicles per unit distance in a highway section, as shown in quadrant (c) of Figure 4.1, is commonly known as the fundamental relationship of traffic flow. This relationship is defined based on the assumption that traffic conditions along the section are stationary, which means that the three key traffic variables, q, k and average vehicular speed, v, are the same at each and every point in the highway section (Daganzo 1997, May 1990). Consequently, the relationship is basically estimated empirically by pooling observations from different cross-sections along the highway across different time-periods and fitting a regression curve to the point cloud. The estimation of such a curve follows from the engineers' interest in a general relationship to characterise the flow of traffic in a given facility. The fundamental relationship can be equivalently expressed as speed-density or flow-speed relationship, as shown in quadrants (a) and (b) of Figure 4.1, respectively.

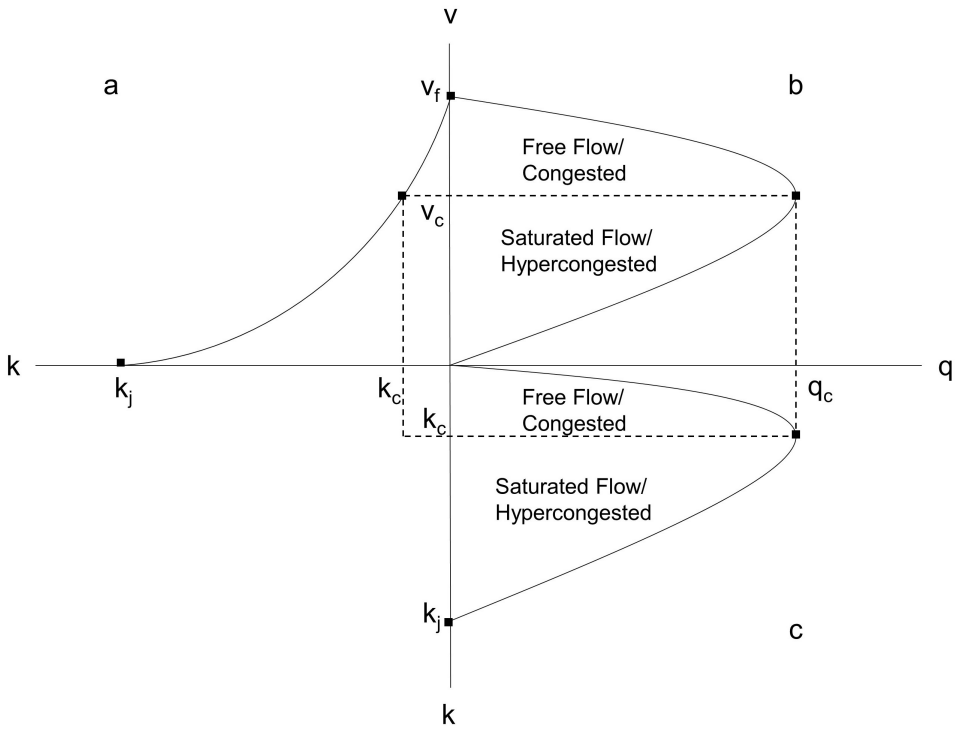


Figure 4.1: The fundamental diagram of traffic flow (adapted from Small & Verhoef 2007)

Engineers assert that the estimated relationship is a property of the road section, the environment, and the population of travellers, because on an average, drivers show the same behaviour (Daganzo 1997). We argue that this estimated relationship, however, is at best only *associational and possibly spurious* due to several possible sources of endogeneity/confounding biases. For instance, there are many external observed and unobserved factors such as driver behaviour, heterogeneous vehicles, weather and demand, that are correlated with the observed traffic variables (Mahnke & Kaupužs 1999, Qu et al. 2017), but are often omitted in the estimation of the fundamental relationship. Fitting a pooled ordinary least square regression curve to the observed scatter plot of traffic variables fails to adjust for the above-mentioned sources of confounding, which may bias the estimated relationship (Wooldridge 2010, Cameron & Trivedi 2005). The parametric limitations on functional form in regression further augments the bias in the estimated relationship.

To address these shortcomings of the traditional approach, in this chapter, we estimate

the fundamental relationship between traffic flow and traffic density using a flexible causal statistical framework. In particular, we adopt a Bayesian non-parametric instrumental variables (NPIV) estimator (Wiesenfarth et al. 2014) that allows us to capture non-linearities in the relationship with a non-parametric specification without presuming the functional form and also adjust for any confounding bias via the use of instrumental variables (IVs). We validate this approach using traffic detector data from three highway bottlenecks located in California, USA[1]. Thus, the main contribution of this research lies in determining a novel causal (unbiased) relationship between traffic flow and traffic density for a highway bottleneck.

To the best of our knowledge, this study presents the first application of *causal inference* in empirical estimation of the fundamental relationship from an engineering perspective that is based on the physics of movement of vehicles. We emphasise that some economists have also adopted a causal framework for this problem in the past (see, for instance, Couture et al. 2018). This framework is based on the interpretation of the speed-flow fundamental relationship as the supply curve for travel in a road section under stationary state traffic conditions (Small & Verhoef 2007, Walters 1961). We argue that in developing a causal understanding of the fundamental relationship, the economic representation of this model as a supply curve can lead to ambiguity. The economic interpretation seeks stationary state traffic conditions, which seldom exist, particularly under congested conditions.

Based on this type of economic representation, a recent empirical study by Anderson & Davis (2020, 2018) discards the existence of the hyper-congested[2] part of the fundamental diagram (see Figure 4.1 to locate the *hypercongested* part). Note that for a highway bottleneck, there are two components of the hypercongested regime of the fundamental

[1]We choose highway bottlenecks over uniform highway sections to demonstrate that our approach not only delivers an unbiased fundamental relationship for a highway section but also correctly estimates capacity-drop, an important feature of the bottleneck section.

[2]Whereas the engineering terminology for the backward bending region of the fundamental diagram is 'saturated flow', economists call it 'hypercongested' (Small & Chu 2003).

diagram: (i) the region representing capacity drop, that is, a sudden reduction in capacity of the bottleneck at the onset of upstream queuing, and, (ii) the region following the capacity drop where the flow-density or flow-speed relationship is backward bending as per the engineering literature. The absence of empirical evidence on the existence of both components of hypercongestion (that is, reduction in traffic flow with increasing traffic density or demand, Anderson & Davis (2020, 2018)) questions the relevance of hypercongestion as a motivating factor for traffic controls measures and congestion pricing strategies.[3]

Through our proposed causal framework, we also aim to conciliate this recently diverging strand from the economics literature with the well-established engineering knowledge on the existence of hypercongestion. Specifically, we contribute to the re-initiated debate on the existence of hypercongestion in highways and deliver novel causal estimates of capacity reduction in various highway bottlenecks. We emphasise that our estimates of the capacity reduction are derived from the estimated fundamental relationship itself, as opposed to the previous literature that uses different methodologies (e.g., change in cumulative vehicle count) to quantify the phenomenon (see, for instance, Cassidy 1998, Oh & Yeo 2012, Srivastava & Geroliminis 2013). Thus, our proposed approach provides a one-stop solution to estimate an unbiased fundamental relationship as well as its important features such as capacity and capacity-drop. As an important intermediate research outcome of this study, we also undertake a critical evaluation of the assumptions underlying the economists' treatment of the fundamental relationship as a supply curve for travel, which may lead to inconclusive empirical evidence on the existence of hypercongestion.

The rest of this chapter is organised as follows. Section 4.2 reviews the relevant engineering literature, critically examining the theoretical foundations underlying the

[3]A few early studies in the engineering literature also report no capacity reductions (Hall & Hall 1990, Persaud 1987, Newman 1961), however, their results have been found to be inconclusive owing to the methods adopted in these studies (Cassidy & Bertini 1999).

empirical estimation of the fundamental relationship between traffic variables. Additionally, we also review the literature on capacity-drop in highway bottlenecks. Section 4.3 describes the chosen study sites and the corresponding traffic detector data and variables. Section 4.4 details the model specification and the adopted methodology explaining how it addresses endogeneity bias in the context of the fundamental relationship. Section 4.5 presents our results and benchmarks them against those derived using a standard non-parametric estimator without instrumental variables. Furthermore, we compare our findings with the relevant engineering and economics literature. Conclusions and implications are presented in the final section.

4.2 Literature Review

In this section, we discuss the theoretical foundations underlying the engineering approach to estimate the fundamental relationship and the key shortcomings of this approach. We also highlight how a causal econometric framework can be employed to obtain a more robust characterisation of the fundamental relationship.

4.2.1 The empirical fundamental relationship

As discussed in the introduction, the fundamental relationship is empirically estimated using observations on traffic state variables that are averaged over time and/or space. The averaging of observations requires strict assumptions like *stationary traffic* and *homogeneous vehicles*. Note that empirical studies use occupancy, o, as a proxy for traffic density because traffic density cannot be measured directly (Daganzo 1997, May 1990)[4]. For a contextual discussion, Figure 4.2 shows a scattered plot of the measured flow versus occupancy from a traffic loop detector (aggregated over 5 minutes) located upstream of the Caldecott Tunnel in the SR24-W, California on several workdays.

[4]Occupancy is defined as the percentage of the sampling period for which vehicles occupy the detector

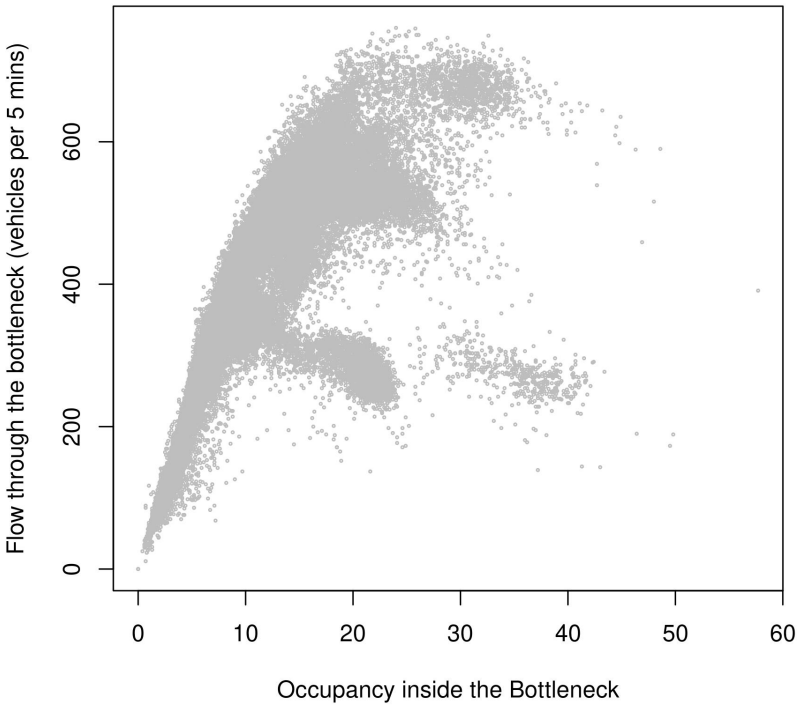


Figure 4.2: Conventional flow versus occupancy plot using detector data aggregated over 5 minutes.

The conventional fundamental relationship is obtained by fitting a predefined curve to the point cloud of aggregated data. While numerous functional forms have been proposed (see Hall et al. 1993, for a review), most researchers agree that the flow-density relationship should either be triangular (Newell 1993, Hall et al. 1986) or parabolic (Greenshields et al. 1935, HCM 2016). Proposed relations also include discontinuous models whereby the functions describing unsaturated and saturated traffic regimes do not come together to form a continuous curve (Payne 1977, Ceder & May 1976, Drake & Schofer 1966).

As noted in previous studies, the data shows considerable scatter particularly in the saturated regime, which has led some researchers to question the existence of reproducible relationships in this regime (refer to May 1990, for details). The challenges posed by this scatter in obtaining an accurate fundamental relationship has stimulated the growth of many interesting strands in the engineering literature: from understanding and accounting

for the various sources of scatter (see, for instance, Cassidy 1998) to introduction of stochastic fundamental relationships (Wang et al. 2013) and fitting multi-regime models (see, for instance, Kidando et al. 2020). The following subsections summarise the main findings from our review.

Characterising the sources of scatter

Cassidy (1998) argued that data from brief periods of near-stationarity (that is, unsustained periods of time over which the traffic stream is marked by nearly constant average vehicle speeds) or non-stationary transitions do not always conform to near-stationary relations because such data points arise due to random fluctuations in the traffic variables. Further, Cassidy (1998) showed that only data points from sustained periods of near stationary traffic condition conform to well-defined reproducible bivariate relations among traffic variables. Cassidy (1998) and Daganzo (1997) attribute such well-defined relationships to the same average behaviour of drivers when confronted with the same average traffic conditions.

However, Coifman (2014*a*) found that even under strict stationary state traffic conditions, aggregated measurements of flow, density (or occupancy) and speed may exhibit large scatter in the queued regime arising from: (i) erroneous measurements of flow due to non-integer number of vehicle headways in a given sampling period, (ii) averaging over a small number of vehicles during low flow periods, (iii) measurement errors due to detectors, (iv) the mixing of inhomogeneous vehicles (for instance, trucks and cars), and, (v) the mixing of inhomogeneous driver behaviours. Consequently, Coifman (2014*b*) relaxed the requirement to seek out strict stationary state conditions by measuring the traffic state (flow and occupancy) for individual vehicles, followed by grouping of these measurements by similar lengths and speeds. For each group, Coifman (2014*b*) derived the flow-occupancy relationship by connecting points corresponding to the median flow and median occupancy. Coifman (2014*b*) argues that the use of median instead of the

conventional use of mean controls for outliers arising from detector errors or uncommon driver behaviour.

We note that there are two main drawbacks of this approach. First, the method is highly data-intensive and requires microscopic-level measurements on individual vehicles. Second, although the method helps in estimating well-defined relationships for homogeneous vehicle classes, a traffic stream seldom consists of homogeneous vehicles only. This deficiency becomes a critical concern because the method does not clearly suggest how to obtain an aggregate relationship for a mix of vehicles from class-specific relationships, which is of general interest to devise traffic control measures and congestion pricing strategies.

Stochastic fundamental relationships

Consistent with Cassidy (1998), a series of other recent studies have attributed the observed scatter to random characteristics of traffic behaviour (Qu et al. 2015, Chen et al. 2015, Mahnke & Kaupužs 1999, Qu et al. 2017, Muralidharan et al. 2011, Jabari et al. 2014, Sopasakis 2004). These studies suggest that the scatter arises due to various external factors such as heterogeneous vehicles, driver behaviour, weather conditions, and the random characteristics of demand. Previous empirical studies also demonstrate how failure to adjust for the stochastic characteristic of traffic flow variables in calibration of the fundamental relationship result into highly inaccurate models (Ni 2015, Wang et al. 2011, Li et al. 2012).

To account for these random characteristics, Wang et al. (2013) introduced a stochastic fundamental relationship in place of traditional deterministic models, in which they assume speed to be a random process of density and a random variable. Subsequently, Hadiuzzaman et al. (2018) and Kidando, Moses & Sando (2019) have used Adaptive Neuro Fuzzy Inference System and Markov Chain Monte Carlo (MCMC) simulations respectively to capture the uncertainty in parameter estimates of the fundamental relationship arising due from the stochastic behaviour of traffic.

We note that the baseline estimates of the flow-density or the speed-density curve derived within the stochastic framework are based on a pooled ordinary least square estimator. We argue that such estimates of the baseline curve may be confounded by the extraneous factors discussed in the literature. We explain this confounding bias in detail in Section 4.4.2.

Multi-regime models

Traditional single-regime models (for instance, Greenberg 1959, Pipes 1966, Munjal & Pipes 1971) that assume a single, pre-defined shaped curve for the entire domain of the fundamental relationship have been found inaccurate because free-flow and congestion-flow regimes have different flow characteristics (Ni 2015, May 1990, Hall et al. 1993). As a consequence, multi-regime models have been introduced in the literature as a flexible alternative to increase the calibration accuracy of the fundamental relationship. Multi-regime models fit different regimes of the fundamental relationship with different pre-defined functional forms where regimes are separated by breakpoints or thresholds (Edie 1961, Drake 1967, Sun & Zhou 2005). Whereas two-regime models comprise of free-flow and congested-flow regimes, three-phase models additionally include a transitional regime between free-flow to congestion, which is consistent with Kerner's three-phase traffic theory (Kerner 2009).

In most studies, modellers pre-define the locations of breakpoints based on the subjective judgement, which may significantly affect the accuracy of the estimated multi-regime models (Wang et al. 2011, Sun & Zhou 2005, Liu et al. 2019). Studies such as Kockelman (2001), Sun & Zhou (2005) and Kidando, Kitali, Lyimo, Sando, Moses, Kwigizile & Chimba (2019) have thus focused on the estimation of these breakpoints based on a user-driven choice of the number of regimes as input. Kidando et al. (2020) proposed a fully data-driven Bayesian approach to estimate the breakpoints of multi-regime models, but did not account for the potential confounding biases. Our flexible non-parametric

approach does not require any user inputs regarding the shape of fundamental relationship, automatically identifies such change-points in a data-driven manner and also accounts for the possible confounding biases.

Research gaps and contributions

Our review suggests that traffic engineers are interested in deriving a well-defined reproducible relationships between traffic variables that can be attributed to the same average behaviour of drivers under the same average traffic conditions. The developments in the engineering literature serve as an excellent starting point to understand the sources of variation in traffic state measurements that leads to large scatter, particularly in the congested regime of the fundamental diagram. While the previous studies on the stochastic fundamental diagram rightly argue that the scatter arises due to various external factors, they do not appropriately adjust for the confounding bias that may occur from these factors when estimating the aggregate (baseline) fundamental relationship. We argue that these factors are likely to be highly correlated with the observed traffic variables, thus, an ordinary least squares based estimation of the fundamental relationship may be biased (see Section 4.4.2 for details). Moreover, most of the multi-regime models require user inputs for calibration and ignoring the endogeneity bias remains the concern.

To fill these gaps in the literature, we introduce a methodological framework to estimate the fundamental relationship that can effectively control for confounding from the external sources identified in the literature, alongside adjusting for other inherent randomness in the data generating process, and produce a more general characterisation of traffic flow for a highway section under an average mix of traffic. The adopted fully flexible non-parametric specification for the fundamental relationship produces a multi-regime fundamental relationship without any prior assumptions on either the shape of the curve or the location of breakpoints. Moreover, as a by product of Bayesian estimation, we also quantify the uncertainty in the estimated relationship with credible intervals. Furthermore,

important traffic control inputs for highway bottlenecks such as capacity and capacity drop are also obtained as a by-product of the estimation. In the rest of this section, we review the relevant literature on highway capacity and capacity drop to illustrate the importance of quantifying these parameters from the estimated fundamental diagram.

4.2.2 Highway Capacity and Capacity Drop

Understanding the capacity of a highway section is critical in modelling of traffic flow in highways (Srivastava & Geroliminis 2013, Siebel et al. 2009, Laval & Daganzo 2006)[5], particularly those with bottlenecks (such as lane drops and merges, among others). This is because, the highway capacity at the bottleneck location may be insufficient for traffic demand during peak hours and hence, traffic jams may occur. Capacity drop is thus defined as the drop in discharge flow through a bottleneck, when it is activated with an increase in demand. The activation of a bottleneck is marked by onset of queuing upstream of the bottleneck Yuan et al. (2015), Oh & Yeo (2012), Chung et al. (2007), Cassidy & Bertini (1999). The literature also acknowledges capacity drop as a *two-capacity* phenomenon of active bottlenecks and relate it to the discontinuity observed at capacity flows near saturation point in the flow-density or flow-speed fundamental diagram. Several empirical observations of capacity drop ranging between 2 percent -25 percent are found in the literature. Table 4.1 presents a summary of the capacity drop reported in the literature for different highway sections with varying bottleneck type. Based on the behavioural theory of traffic flow, Daganzo (2002) attributes the capacity drop to a loss of *motivation* among drivers, that is, these drivers presumably loose their willingness to drive at high speeds with small headways. In addition, Laval & Daganzo (2006) and Leclercq et al. (2011) suggest that capacity drop occurs due to voids in the traffic caused by lane changing.

[5]Note that there are many potential definitions of capacity in the literature (see Kondyli et al. 2017, for a brief review). For instance, Cassidy & Rudjanakanoknad (2005) define capacity as the sustained flow discharged from all highway exits that are unblocked by spillover queues from downstream while the highway entrances are queued. Oh & Yeo (2012) define capacity as the maximum discharge flow of vehicles that persist for 5 minutes in a free-flow state.

Past researchers also relate the differences in capacity drop values with the number of lanes, severity of stop-and-go waves and speeds in congestion Oh & Yeo (2012, 2015), Yuan et al. (2015).

Table 4.1: Summary of key literature on the existence capacity drop in highways.

Study	Location	Type	Capacity Drop (%)
Banks (1990)	I-8, San Diego	on-ramp merge	-0.42 to 1.11
Hall & Agyemang-Duah (1991)	Queen Elizabeth Way, Toronto	on-ramp merge	-7.76 to 10.36
Banks (1991)	Multiple Sites, San Diego	merge/ lane drop/ weave	1.8 to 15.4
Persaud et al. (1998)	Multiple Sites, Toronto	on-ramp merge	10.6-15.3
Cassidy & Bertini (1999)	Multiple Sites, Toronto	on-ramp merge	4 to 10
Bertini & Malik (2004)	US-169, Minneapolis	on-ramp merge	2 to 5
Bertini & Leal (2005)	M4, London	merge	6.7 to 10.7
Cassidy & Rudjanakanoknad (2005)	I-805, San Diego	on-ramp merge	8.3 to 17.3
Chung et al. (2007)	I-805, San Diego	on-ramp merge	5 to 18
	SR-24, California	on-ramp merge, lane reduction	5.1 to 8.5
	Gardiner Expressway, Toronto	on-ramp merge, horizontal curve	3 to 12
Leclercq et al. (2011)	M6, UK	merge	25
Oh & Yeo (2012)	Multiple Sites, California	on-ramp merge	8 to 16.5
Srivastava & Geroliminis (2013)	US-169, Minneapolis	on-ramp merge	8 to 15
Jin et al. (2015)	I-405, California	merge	10.5
Anderson & Davis (2020)	Multiple Sites, California	lane reduction	0

*This table has been adapted from Oh & Yeo (2012).

The existence of capacity-drop in a highway section is well-recognised in the transportation literature and has been a long-standing rationale for application of traffic controls, such as, ramp metering (Cassidy & Rudjanakanoknad 2005, Smaragdis et al. 2004, Diakaki et al. 2000), and highway pricing and tolls (Hall 2018*b*,*a*, Fosgerau & Small 2013, Newbery 1989, Boardman & Lave 1977, Walters 1961) to regulate demand. We note that the methods adopted in the literature to quantify capacity drop differ substantially from each other. For instance, Banks (1990), Persaud et al. (1998), Zhang & Levinson (2004) and many more study minute-to-minute variability in traffic flows to infer decrease in high traffic flow levels. Studies like Bertini & Leal (2005) and Cassidy & Rudjanakanoknad (2005) use cumulative vehicle counts to infer the reduction in flow at downstream detectors relative

to upstream detectors. Srivastava & Geroliminis (2013) study changes in bottleneck flows with respect to upstream density.

However, the capacity-drop phenomenon has recently been called into question in the urban economics literature. Anderson & Davis (2020, 2018) study the changes in capacity of a highway section with a bottleneck during periods of high demand for three bottlenecks in California and conclude that there is no evidence of capacity drop, hence, hypercongestion, in the absence of any infrastructure-related shocks (for instance, lane closures, traffic incidents and so on) and weather-based shocks. Consequently, they question the relevance of hypercongestion in the design of traffic controls and congestion pricing.

The capacity drop estimates in engineering studies are based on only a certain few days of observations. Anderson & Davis (2020) instead use data from several hundred days and adopt an event-study design to measure changes in highway capacity before and after of queue formation averaged over all days. It is important to note that Anderson & Davis (2020) select a speed threshold of 30 mph at the upstream detector closest to the bottleneck to detect the onset of upstream queuing.

We argue that inferring the actual moment of queue formation using a speed threshold for upstream detectors may lead to ambiguity. We instead reevaluate the capacity drop phenomenon by deriving estimates of capacity drop from the causal estimate of the flow-occupancy relationship at the bottleneck location. However, as rightly suggested by Anderson & Davis (2020), we use several months of observations to separate capacity drop from minute-to-minute fluctuations in traffic flow.

4.3 Data and Relevant Variables

We make use of traffic data from three standard highway sections with distinct geometry, each having a single and clearly identified active bottleneck located at its downstream end. At all of the chosen sites, slowing down of traffic and queuing is observed at the

bottleneck location. The associated high-quality data is collected via a series of loop detectors installed at various locations along the highway, which measure traffic flow and occupancy averaged over every 5-minute duration. The data is maintained by the California Department of Transportation (Caltrans) and made publicly available through their Performance Measurement System (PeMS) website[6].

4.3.1 Study Sites

Site 1

The first bottleneck that we study is located in the westbound direction of the California State Route 24 (SR-24) at the Caldecott Tunnel in Oakland, California. The SR-24 connects suburban Contra Costa County in the East Bay region of the San Francisco Bay Area with the cities of Oakland and San Francisco in the west. During the period 2005-2010, the Caldecott tunnel was operated with two reversible lanes carrying the westbound traffic in the morning and the eastbound traffic in the afternoon and evening. Thus, for afternoon and evening hours of the above period, the location features an active bottleneck in the westbound direction with the number of lanes decreasing from *four to two*. As the traffic approaches the tunnel, traffic delays being quite common at this location (previously studied by Chin & May 1991, Chung et al. 2007, Anderson & Davis 2020, among others). We use observations on the westbound traffic in the time period 12:00-24:00 hours on weekdays in the months of June-August during 2005-2010. This highway section is well-isolated, that is, located well away from any major downstream intersection. Consequently, we assume that this section allows us to study traffic dynamics arising solely from the presence of the bottleneck, without being affected by any downstream influences.

[6]Performance Measurement System (PeMS) website: http://pems.dot.ca.gov/

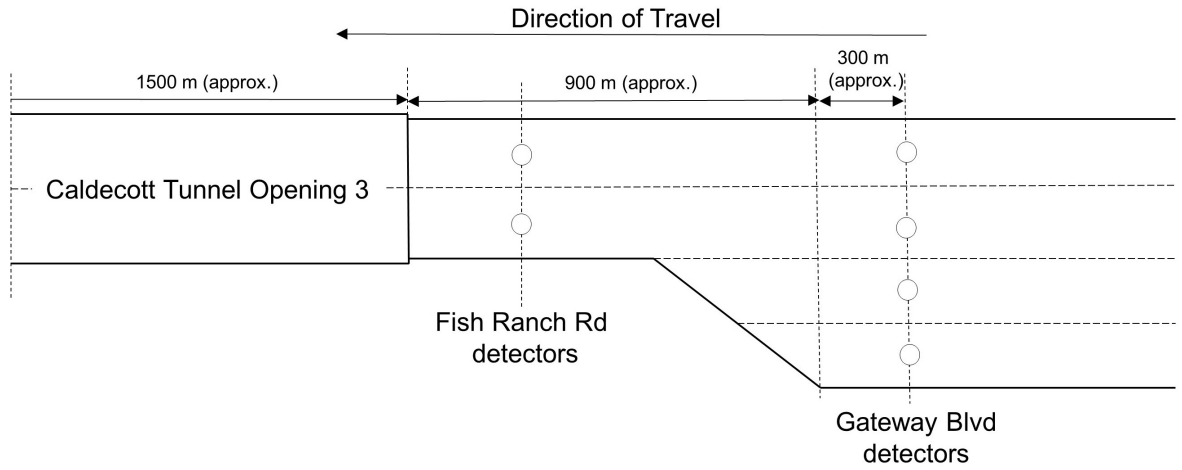
Site 2

Our second study site is located in the eastbound direction of the California State Route 91 (SR-91). SR-91 connects several regions of the Greater Los Angeles urban area in the west with the Orange and Riverside Counties in the east. At the location where two-lane traffic from the Central Avenue- Magnolia Centre in the Riverside Country merges with its three-lane eastbound traffic, it features an active merge bottleneck (previously studied by Oh & Yeo 2012). This bottleneck appears as one of the top 100 bottlenecks in California enlisted on the PEMS website with queuing and delays being quite common at this location during morning and evening hours. We use observations on the eastbound traffic in the time period 06:00-12:00 hours on weekdays in the months of June-August during 2009-2014. This highway section is also well-isolated, thus, the traffic dynamics arising within the section arise solely from the presence of the bottleneck.

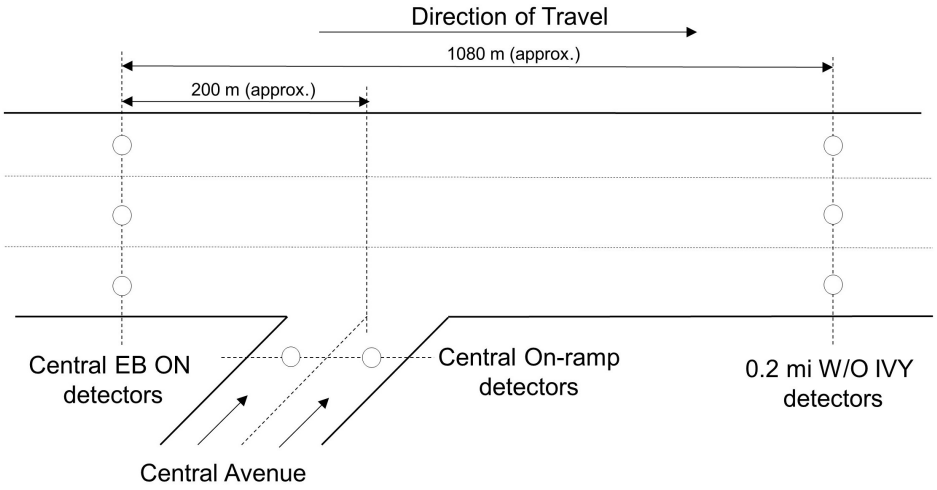
Site 3

The third bottleneck that we study is located in the eastbound direction of California State Route 12 (SR-12). SR-12 connects the Sonoma, Napa, and Solano Counties, following which it merges with Interstate 80 (I-80) which continues north towards Sacramento. At a location just west of I-80, the number of lanes in the highway drops from two lanes to one. This site has been previously studied by Anderson & Davis (2020) who suggest that the lane-drop results in the formation of an active bottleneck with queues that are often very long. We use observations on the eastbound traffic in the time period 12:00-24:00 hours on weekdays in the months of May-August during 2018-2019.

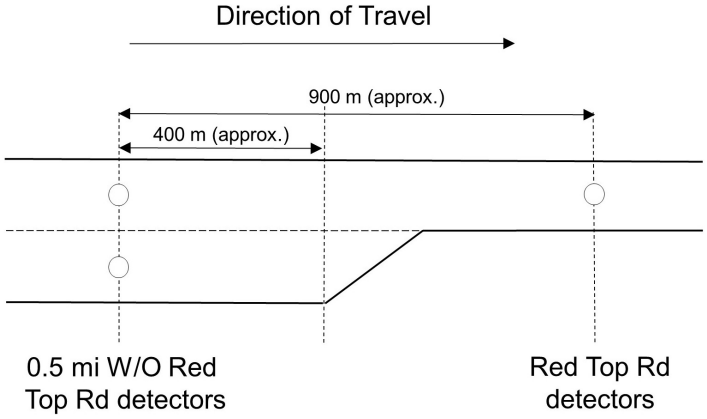
Further downstream of the lane drop, the SR-12 merges with I-80. However, Anderson & Davis (2020) note that the dynamics within this section are not affected by this merge. We adopt this section for further analysis as it offers an interesting avenue to verify the effect of downstream influences on the estimated fundamental relationship within the bottleneck.



(a) A lane-drop bottleneck in the westbound SR-24 in Oakland, California.



(b) A merge bottleneck in the eastbound SR-91 in Riverside, California.



(c) A lane-drop bottleneck in the eastbound SR-12 in Solano, California.

Figure 4.3: Schematic representation of the study sites.

Table 4.2: Summary statistics for variables used in this analysis.

(a) Site 1: SR-24 westbound.

Variable	Detectors	Obs.	Min	Max	Mean	Std.Dev
Traffic Flow (veh/5mins)	Gateway Blvd	54432	0.00	595.00	249.58	77.57
Occupancy	Gateway Blvd	54432	0.00	73.60	18.82	18.16
Traffic Flow (veh/5mins)	Fish Ranch Rd	54432	0.00	760.00	383.95	149.57
Occupancy	Fish Ranch Rd	54432	0.00	57.70	12.29	7.19

(b) Site 2: SR-91 eastbound.

Variable	Detectors	Obs.	Min	Max	Mean	Std.Dev
Traffic Flow (veh/5mins)	Central EB ON	27936	102.00	593.00	372.21	53.03
Occupancy	Central EB ON	27936	3.60	57.90	11.20	5.08
Traffic Flow (veh/5mins)	W/O IVY	27936	152.00	628.00	393.55	69.77
Occupancy	W/O IVY	27936	3.60	63.20	10.89	3.09
Traffic Flow (veh/5mins)	Central On-ramp	27936	0.00	131.00	35.99	33.33

(c) Site 1: SR-12 eastbound.

Variable	Detectors	Obs.	Min	Max	Mean	Std.Dev
Traffic Flow (veh/5mins)	W/O Red Top Rd	24908	0.00	210.00	102.13	43.08
Occupancy	W/O Red Top Rd	24908	0.00	74.90	18.64	19.44
Traffic Flow (veh/5mins)	Red Top Rd	24908	0.00	173.00	99.68	43.71
Occupancy	Red Top Rd	24908	0.00	69.10	11.41	6.84

4.3.2 Relevant Variables

A schematic representation of the three bottlenecks, along with the location of detectors that we use to obtain the relevant data, is shown in Figure 4.3.

For the first site, we observe a set of two detectors downstream of the lane-drop (that is, within the bottleneck) and four detectors upstream to it. For the second site, we observe a set of three detectors downstream of the merge (that is, within the bottleneck) and three detectors upstream to it. For the third site, we observe one detector downstream of the lane-drop (that is, within the bottleneck) and two detectors upstream to it.

It is also worth emphasising that for all the three bottlenecks, there are no reasonable alternative routes to the highway section for the analysed traffic. We can thus assume that, on an average, driver population using the section during the study period does not differ substantially on weekdays. Table 4.2 summarises the relevant variables from the three sites that are used in this study.

4.4 Methodology

This section is divided into four subsections. In the first subsection, we discuss the model specification. In the second subsection, we explain potential endogeneity bias in estimation of the fundamental relationship. In the penultimate subsection, we briefly review NPIV methods in the literature and describe the Bayesian NPIV method in the context of this study. In the concluding section, we benchmark the performance of the Bayes NPIV estimator against state-of-the-art estimators in a Monte Carlo study and illustrate its ability to adjust for endogeneity bias and recover complex functional forms.

4.4.1 Model Specification

We estimate a causal relationship between occupancy inside the bottleneck, o^b_{it}, in the five-minutes interval i, $i = 1, .., N$, on a particular day t, $t = 1, ..., T$, and the flow through the bottleneck, q^b_{it}. We consider q^b_{it} to be a function of o^b_{it}, conditional on the properties of the infrastructure, the environmental conditions and the average behaviour of drivers and vehicles.

$$q^b_{it} = S^b(o^b_{it}) + \delta^b_{it} + \xi^b_{it} \tag{4.1}$$

where δ^b_{it} includes the unobserved (to researchers) traffic-specific behavioural component common to all drivers, traffic-specific vehicular attribute common to all vehicles, weather-specific component affecting the entire traffic stream, and demand-related characteristic. ξ^b_{it} represents an idiosyncratic error term representing all random shocks to the dependent variable. Since the exact structural form of how o^b_{it} enters into equation is unknown, we adopt a non-parametric specification of $S^b(.)$. δ^b_{it} is expected to be correlated with o^b_{it}. We explain the implications of this correlation on the estimated relationship in the next subsection (Section 4.4.2).

As a by-product of this estimation, we quantify the activation of the bottleneck as

follows. Consistent with the engineering literature, we consider that the flow through the bottleneck drops following the activation of the bottleneck. Thus, we infer the critical value of o_c^b at which we observe a significant backward bending in q^b from the estimated relationship $S^b(.)$. We also note that when the occupancy inside the bottleneck remains at and above o_c^b, the bottleneck remains activated.

Through the estimated relationship, we quantify the capacity of the bottleneck q_c^b, that, is flow through the bottleneck corresponding to o_c^b and examine the existence of capacity drop or two-capacity phenomenon following the activation of the bottleneck.

4.4.2 Bias due to Endogeneity

Building a credible causal relationship between traffic variables requires the understanding of potential endogeneity biases. There are two major concerns in relation to endogeneity: omitted variable bias and reverse causality (simultaneity). Omitted covariates that are correlated with both the dependent variable and the included covariates in a regression may result in inconsistent estimates of model parameters. For instance, in equation 4.1, omission of covariates representing driving behaviour of users due to unavailability of a comprehensive aggregate level measure may lead to confounding bias in the estimated relationship. This bias occurs because driving behaviour may be correlated with both occupancy and flow. Reverse causality is a consequence of the existence of a two-way causal relationship or a cause-effect relationship, contrary to the one assumed in the model. For instance, in equation 4.1, we assume the flow through the bottleneck, q_{it}^b to be a function of the occupancy inside the bottleneck, o_{it}^b. However, there may be reverse causality where q_{it}^b affects o_{it}^b in certain traffic situations. The presence of reverse causality may also lead to inconsistent estimates. We further mathematically demonstrate the two sources of confounding and resulting biases in Appendix A.

4.4.3 Bayesian Nonparametric Instrumental Variable Approach

To address both endogeneity biases, we adopt regression estimators with instrumental variables (IV). IV-based estimators such as two-stage least squares (2SLS) are widely adopted in applied econometrics to estimate parametric models that contain endogenous explanatory variables (see, for example Wooldridge 2010). However, finite-dimensional parametric models for the fundamental relationship of traffic flow such as a linear speed-density or a quadratic flow-speed model, are based on assumptions that are rarely justified by engineering or economic theories. The resulting model mis-specification may lead to erroneous estimates of attributes characterising the fundamental relationship (for instance, capacity or capacity drop). On the other hand, non-parametric methods have the potential to capture the salient features in a data-driven manner without making a priori assumptions on the functional form of the relationship (Horowitz 2011). Therefore, a fairly growing strand in the econometrics literature proposes different approaches for non-parametric instrumental variables (NPIV) regression, but such methods have not been considered in the estimation of fundamental diagram. Extensive reviews can be found in Newey & Powell (2003) and Horowitz (2011). The NPIV approaches are either based on regularisation or control function. In this study, we adopt a control-function based Bayesian NPIV estimator (Wiesenfarth et al. 2014). In what follows, we start with the general model set-up. Subsequently, we discuss the advantages of the adopted control-function-based Bayesian approach. Additionally, in Appendix B, we summarise the challenges associated with regularisation based approaches that are more commonly adopted in the empirical economics literature.

We first rewrite equation 4.1 in a traditional two-stage IV-based regression set up:

$$q = S(o) + \epsilon_2, \quad o = h(z) + \epsilon_1 \tag{4.2}$$

with response q, endogenous covariate o, IV z for o and idiosyncratic error terms ϵ_1 and ϵ_2 for the first and second stage regressions, respectively. For the notational simplicity, we

drop time-day subscripts and superscripts. Note that δ in equation 4.1 are encapsulated in ϵ_2. Endogeneity bias arises as $E(\epsilon_2|o) \neq 0$. We assume the following identification restrictions:

$$E(\epsilon_1|z) = 0 \quad and \quad E(\epsilon_2|\epsilon_1, z) = E(\epsilon_2|\epsilon_1), \tag{4.3}$$

which yields

$$\begin{aligned} E(q|o, z) &= S(o) + E(\epsilon_2|\epsilon_1, z) = S(o) + E(\epsilon_2|\epsilon_1) \\ &= S(o) + \nu(\epsilon_1), \end{aligned} \tag{4.4}$$

where $\nu(\epsilon_1)$ is a function of the unobserved error term ϵ_1. This function is known as the control function.

Control function-based approaches

Several control function-based approaches to estimation of equation 4.2 in the literature, adopt a two-stage approach where residuals $\hat{\epsilon_1}$, that is, $o - \hat{h(z)}$ from the first stage are used as additional covariate in the second stage (for details, see Newey & Powell 2003). However, as pointed out by Wiesenfarth et al. (2014), such two-stage approaches have certain limitations. First, the uncertainty introduced by estimating the parameters in the first stage remains unincorporated in the second stage. Second, a precise estimate of $\nu(\epsilon_1)$ to achieve full control for endogeneity is difficult to obtain because the focus is on minimising the error in predicting o in the first stage. Third, a robustness control is required to account for outliers and extreme observations in ϵ_1 that may affect the endogeneity correction.

Bayesian control-function-based approaches can address these shortcomings of frequentist counterparts and regularisation-based approaches by estimating equation 4.2 as a simultaneous system of equations, allowing for automatic smoothing parameter selection

for a precise estimation of the control function and for construction of simultaneous credible bands [7].

However, early Bayesian control-function-based approaches consider a bivariate Gaussian distribution of errors $(\epsilon_1, \epsilon_2) \sim N(0, \Sigma)$ (for instance, see Chib et al. 2009). This assumption leads to linearity of the conditional expectation as, $E(\epsilon_1|\epsilon_2) = \frac{\sigma_{12}}{\sigma_1^2}$, where $\sigma_{12} = cov(\epsilon_1, \epsilon_2)$ and $\sigma_1^2 = var(\epsilon_1)$, restricting the control function to be linear in ϵ_1 (Wiesenfarth et al. 2014, Conley et al. 2008). Since outliers can be a common source of non-linearity in error terms, they can aggravate the robustness issues of such linear specifications. To overcome these limitations, Conley et al. (2008) proposed the application of a Dirichlet process mixture (DPM) prior to obtain a flexible error distribution, but still relied on linear covariate effects. The method proposed by Wiesenfarth et al. (2014) and adopted in this study, extends the approach by Conley et al. (2008) and allows for fully-flexible covariate effects.

Adopted Bayesian NPIV approach (Wiesenfarth et al. 2014)

The Wiesenfarth et al. (2014)'s Bayesian NPIV approach thus allows us to correct for endogeneity bias in regression models where the covariate effects and error distributions are learned in a data-driven manner, obviating the need of a priori assumptions on the functional form.

To satisfy the identification restrictions presented in equation 4.3, we need an instrumental variable (IV) z. The IV should be (i) exogenous, that is, uncorrelated with ϵ_2; (ii) relevant, that is, correlated with the endogenous covariate o, conditional on other covariates in the model. Due to the absence of suitable external instruments, we use an aggregate lagged level of the endogenous covariate (occupancy) as an instrument. Specifically, for occupancy observed in the five-minutes interval i on day t, we consider the average of observations on the covariate from the interval $i-15$ to $i+15$ from the previous

[7] Credible bands are the Bayesian analogue to confidence bands in the frequentist set up that represent the uncertainty of an estimated curve.

workday $t-1$ as its instrument. We argue that the occupancy o^b_{it} in the five-minutes interval i on day t is correlated with the occupancy $o^b_{[i-15,i+15],t-1}$ in the thirty-minutes interval surrounding i on the previous day $t-1$. This correlation follows from the influence of time-of-the-day on demand and from the fact that there are no reasonable alternative routes to the highway sections being studied, so the population of drivers using the section in the duration i over different workdays may not differ substantially. Moreover, as the highway infrastructure remains unaltered during the study period, we expect the average driving behaviour and thus traffic density to not differ substantially over different days as the drivers are already conversant with the route. However, these lagged occupancy values $o^b_{[i-15,i+15],t-1}$ are exogenous because they do not directly determine the response variable q^b_{it} in equation 4.1 and would never feature in the model for that response. To justify the relevance of the considered instrument, we present the estimated $h(.)$ in equation 4.2 and complimentary results from Stock & Yogo (2005) weak instrument F-tests in the Results and Discussion Section (Section 4.5.2).

Conditional on the availability of an instrument, the Bayesian NPIV estimator can correct for the confounding bias. To account for nonlinear effects of continuous covariates, both $S(.)$ and $h(.)$ (refer equation 4.2) are specified in terms of additive predictors comprising penalised splines. Each of the functions $S()$ and $h(.)$ is approximated by a linear combination of suitable B-spline basis functions. The penalised spline approach uses a large enough number of equidistant knots in combination with a penalty to avoid over-fitting. Moreover, the joint distribution of ϵ_1 and ϵ_2 is specified using nonparametric Gaussian DPM, which ensures robustness of the model relative to extreme observations. Efficient Markov chain Monte Carlo (MCMC) simulation technique is employed for a fully Bayesian inference. The resulting posterior samples allow us to construct simultaneous credible bands for the non-parametric effects (i.e., $S(.)$ and $h(.)$). Thereby, the possibility of non-normal error distribution is considered and the complete variability is represented by Bayesian NPIV. We now succinctly discuss specifications of the kernel error distribution

and computation of credible bands in Bayesian NPIV.

To allow for a flexible distribution of error terms, the model considers a Gaussian DPM with infinite mixture components, c, in the following hierarchy:

$$\begin{aligned}
(\epsilon_{1i}, \epsilon_{2i}) &\sim \sum_{c=1}^{\infty} \pi_c \mathrm{N}(\mu_c, \Sigma_c) \\
(\mu_c, \Sigma_c) &\sim G_0 = \mathrm{N}(\mu|\mu_0, \tau_\Sigma^{-1}\Sigma)\,\mathrm{IW}(\Sigma|s_\Sigma, S_\Sigma) \\
\pi_c = \upsilon_c \left(1 - \sum_{j=1}^{c-1}(1-\pi_j)\right) &= \upsilon_c \prod_{j=1}^{c-1}(1-\upsilon_j), \\
c &= 1, 2, ... \\
\upsilon_c &\sim \mathrm{Be}(1, \zeta).
\end{aligned} \tag{4.5}$$

where μ_c, Σ_c and π_c denote the component-specific means, variances and mixing proportions. The mixture components are assumed to be independent and identically distributed with the base distribution G_0 of the Dirichlet process (DP), where G_0 is given by a normal-inverse-Wishart distribution. The mixture weights are generated in a stick-breaking manner based on a Beta distribution with concentration parameter $\zeta > 0$ of the DP. The concentration parameter ζ determines the strength of belief in the base distribution G_0.

Estimation Practicalities

We exclude discussion of the Gibbs sampler of Bayesian NPIV for brevity, and focus mainly on implementation details and posterior analysis. Interested readers can refer to Wiesenfarth et al. (2014) for derivation of conditional posterior updates.

We use the *BayesIV* and *DPpackage* in R to estimate the Bayesian NPIV. We consider 50,000 posterior draws in the estimation, exclude the first 15,000 burn-in draws and keep every 10^{th} draw from the remaining draws for the posterior analysis. The point-wise posterior mean is computed by taking the average of 3,500 posterior draws. Bayesian simultaneous credible bands are obtained using quantiles of the posterior draws. A simultaneous credible band is defined as the region I_δ such that $P_{S|data}(S \in I_\delta) = 1 - \delta$,

that is, the posterior probability that the entire true function $S(.)$ is inside the region given the data equals to $1-\delta$. The Bayesian simultaneous credible bands are constructed using the point-wise credible intervals derived from the $\delta/2$ and $1-\delta/2$ quantiles of the posterior samples of $S(.)$ from the MCMC output such that $(1-\delta)100\%$ of the sampled curves are contained in the credible band. Similar process is used to obtain the credible intervals of $h(.)$.

4.4.4 Monte Carlo Simulations

We succinctly demonstrate the ability of the adopted Bayesian NPIV approach in addressing challenges of functional form mis-specification and endogeneity in an instance of a Monte Carlo study. We benchmark the Bayesian NPIV method against state-of-the-art estimators to show how this method is robust to both issues. In the data generating process (DGP), we consider a concave regression function, that is, a fourth degree polynomial specification but our conclusions are applicable for a more complex specification.

We use a sample of 10000 observations, with the following DGP:

$$\begin{aligned} y &= -40x^4 + 40x^3 + 30w^4 + \epsilon_2 \\ x &= 3.5z + 2.1w + \epsilon_1 \end{aligned} \tag{4.6}$$

where z and w are independent and uniformly distributed on [0,1]. ϵ_1 and ϵ_2 are independent and identically distributed draws from the N [0,0.5] distribution. The variables y represents the primary response variable, x denotes the endogenous covariate and z represents the instrumental variable. The variable w captures the unobserved effects in the model, that is, we assume that the analyst is ignorant of its presence in the true data generating process for the dependent variable. We thus introduce one possible source of confounding into the model: a positive correlation between the unobserved effect w and the endogenous covariate x.

We note that the model set-up is similar in structure to equation 4.2, that is:

$$\begin{aligned} y &= s(x) + \epsilon_2 \\ x &= h(z) + \epsilon_1 \end{aligned} \tag{4.7}$$

We apply four different estimators to estimate the curve $s(x)$:

1. A two-stage least square (2SLS) estimator with a quadratic specification for $s(x)$.[8]
2. A two-stage least square (2SLS) estimator with the true specification for $s(x)$.
3. A Bayesian non-parametric estimator without instrumental variables (Bayes NP).
4. A Bayesian non-parametric estimator with instrumental variables (Bayes NPIV).

In the latter two approaches, we take 40000 Posterior draws to ensure stationarity of Markov chains. For the posterior analysis, the initial 10000 draws were discarded for burn-in and every 40th draw of the subsequent 30,000 draws was used for posterior inference. Figure 4.4 overlays the estimated $s(x)$ from the four approaches and true $s(x)$.

We note that a 2SLS estimator with the true specification for $s(x)$ is able adjust for the endogeneity bias and could produce an unbiased estimate of $s(x)$. However, in practice, it is infeasible for the analyst to know the correct functional form specification a priori. A functional form mis-specification can produce a highly biased estimate of $s(x)$, as shown by the estimated $s(x)$ using the 2SLS estimator with a quadratic specification for $s(x)$. This exercise thus illustrates the importance of adopting a fully flexible non-parametric specification for $s(x)$ in a relationship.

However, in the presence of endogeneity, a traditional non-parametric estimator may fail to produce an unbiased estimate of $s(x)$. From Figure 4.4, we note that the curve produced by the Bayes NP is highly biased. Adopting an estimator such as the Bayes NPIV allows to adjusts for the endogeneity bias and produce an unbiased estimate of the curve $s(x)$.

[8]Instead of a traditionally-used linear specification, we choose quadratic specification in 2SLS because the scatter plot of the data would intuitively suggest the analyst to use such functional form of $s(x)$.

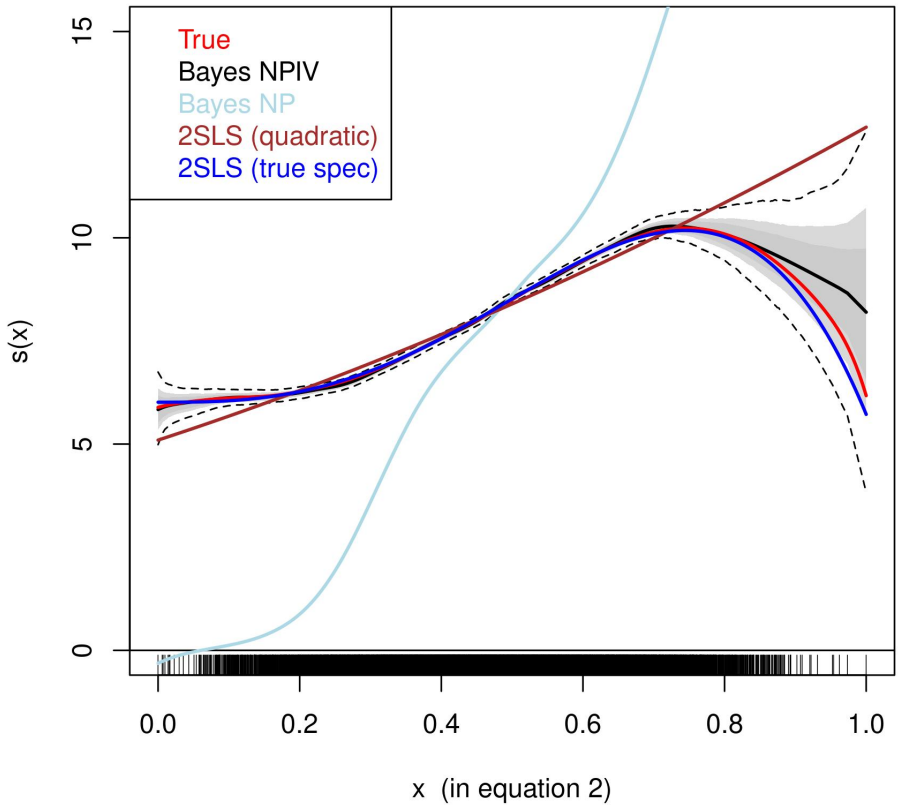


Figure 4.4: Comparison of different estimators in the Monte Carlo study.

In summary, this Monte Carlo exercise shows that the Bayes NPIV estimator, the one adopted in this study, outperforms other parametric and non-parametric approaches as it is allows for a fully flexible functional form specification and controls for any potential confounding bias.

4.5 Results and Discussion

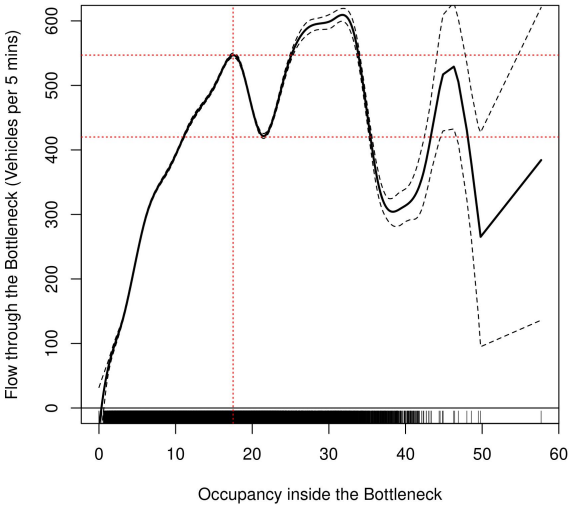
This section is divided into two subsections. In the first subsection, we compare results of the adopted Bayesian NPIV estimator with those of a Bayesian NP estimator and a pooled ordinary least squares (POLS) estimator with a quadratic specification. The Bayesian NP estimator is a counterpart of the Bayesian NPIV, which does not address confounding bias

(that is, $z = x; \epsilon_1 = 0; h(.)$: identity function in Equation 4.2). Furthermore, we discuss the estimates of the capacity and capacity-drop in detail and compare these values with those reported in the literature. In the next subsection, we present the estimated kernel error distributions to illustrate the importance of the non-parametric DPM specification. The relevance of our instruments is also demonstrated in this subsection.

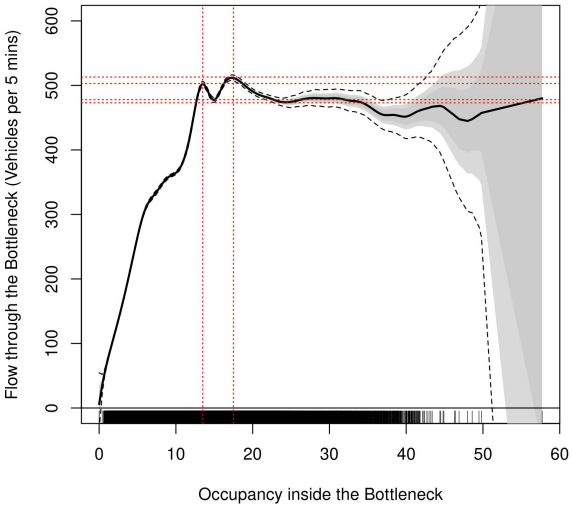
4.5.1 Comparison of Bayesian NPIV and non-IV-based estimators

We present the estimates of $S(.)$ (see equation 4.2, second-stage) using Bayesian NPIV, Bayesian NP, and POLS in Figures 4.5, 4.6 and 4.7 for the three highway sections. POLS results are mainly presented to illustrate how commonly-used parametric non-IV-based specifications can result biased results, but most discussion would revolve around comparing results of Bayesian NPIV and its non-IV counter part (that is, Bayesian NP) .

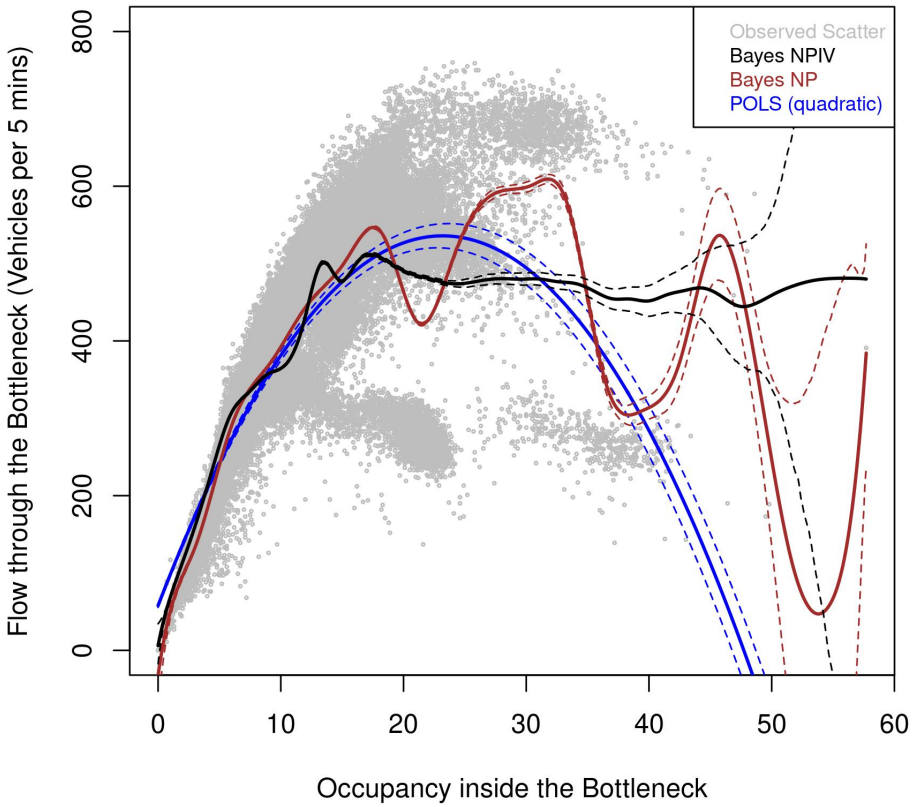
From each of these figures, we do not observe any notable differences between the Bayesian NPIV and Bayesian NP estimate of the free-flow regime of the flow-occupancy curve. In this regime, the Bayesian NPIV estimate of $S(.)$ is as efficient as its Bayesian NP counterpart, as evidenced by tight credible bands in the domain of occupancy where we have sufficient number of observations (note that the density of the tick marks on the X-axis represents the number of observations). However, we observe substantial differences near the saturation (capacity) point and in the congested (or hypercongested as per the economics literature) regime of the estimate curve (see Figures 4.5(c), 4.6(c) and 4.7(c)). We further discuss these differences in detail in next sub-sections.



(a) Non-parametric non-IV estimator.

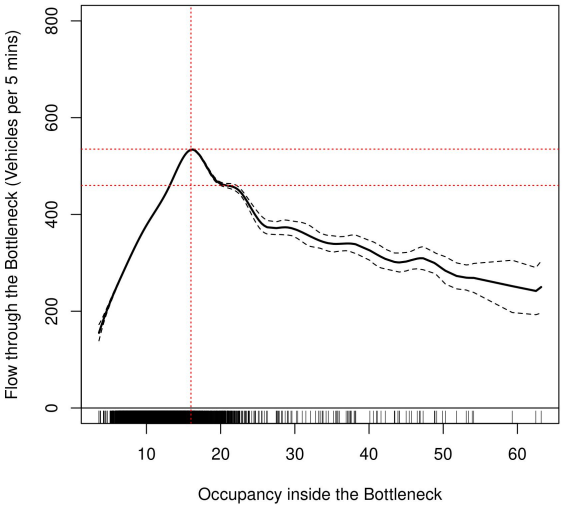


(b) Non-parametric IV-based estimator.

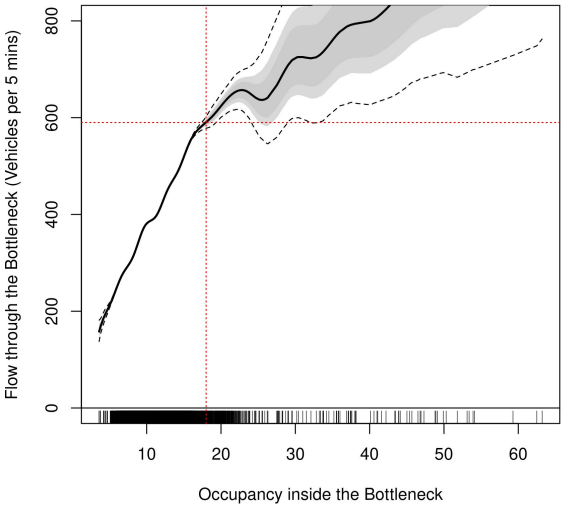


(c) Comparison of different estimators.

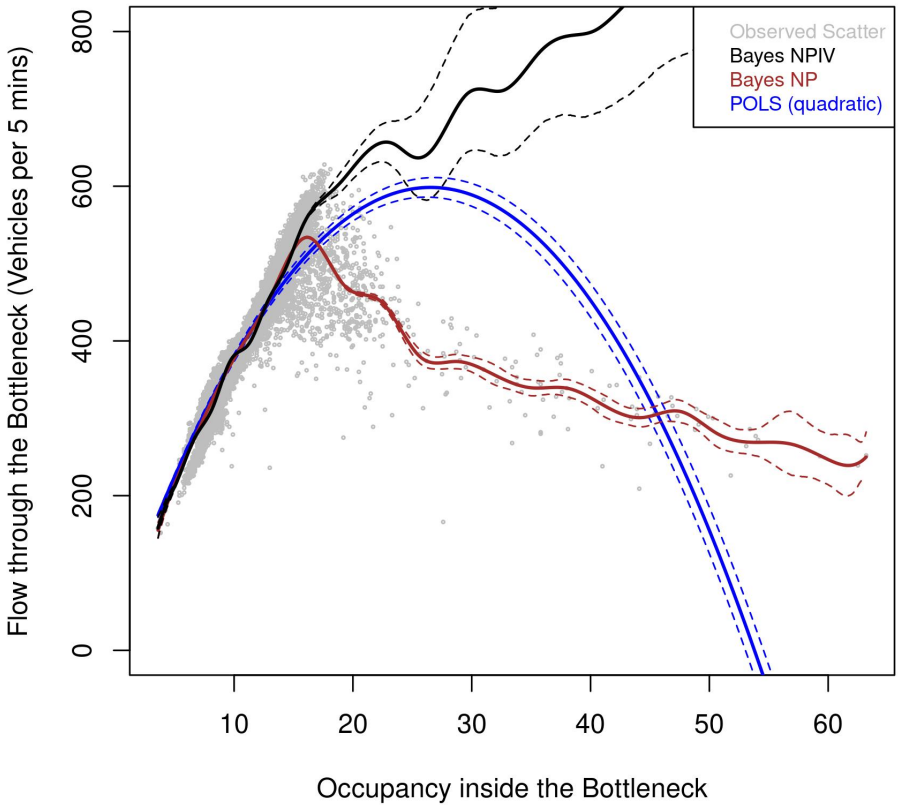
Figure 4.5: Estimated flow-occupancy curves for Westbound SR-24.



(a) Non-parametric non-IV estimator.

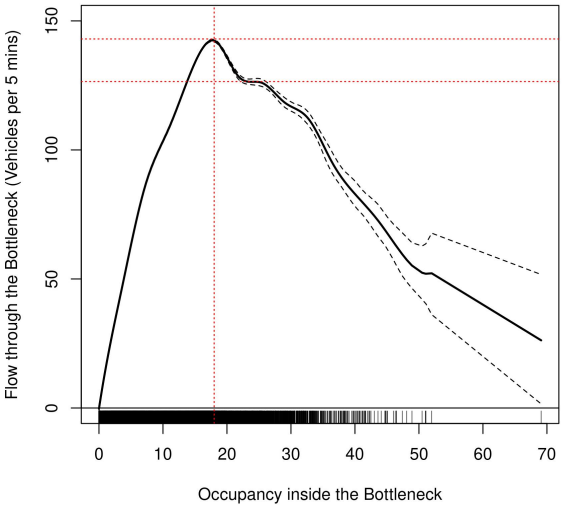


(b) Non-parametric IV-based estimator.

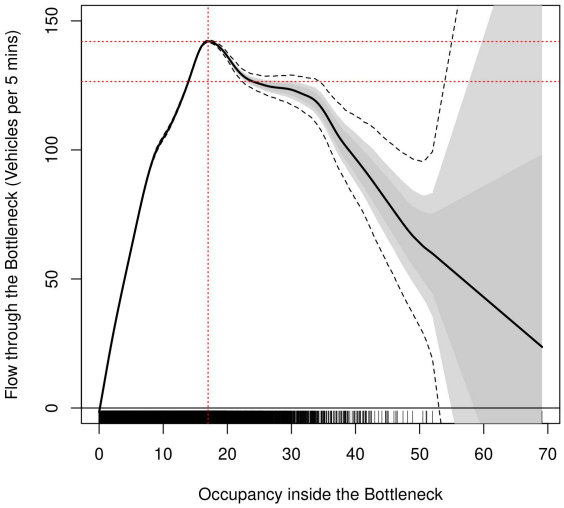


(c) Comparison of different estimators.

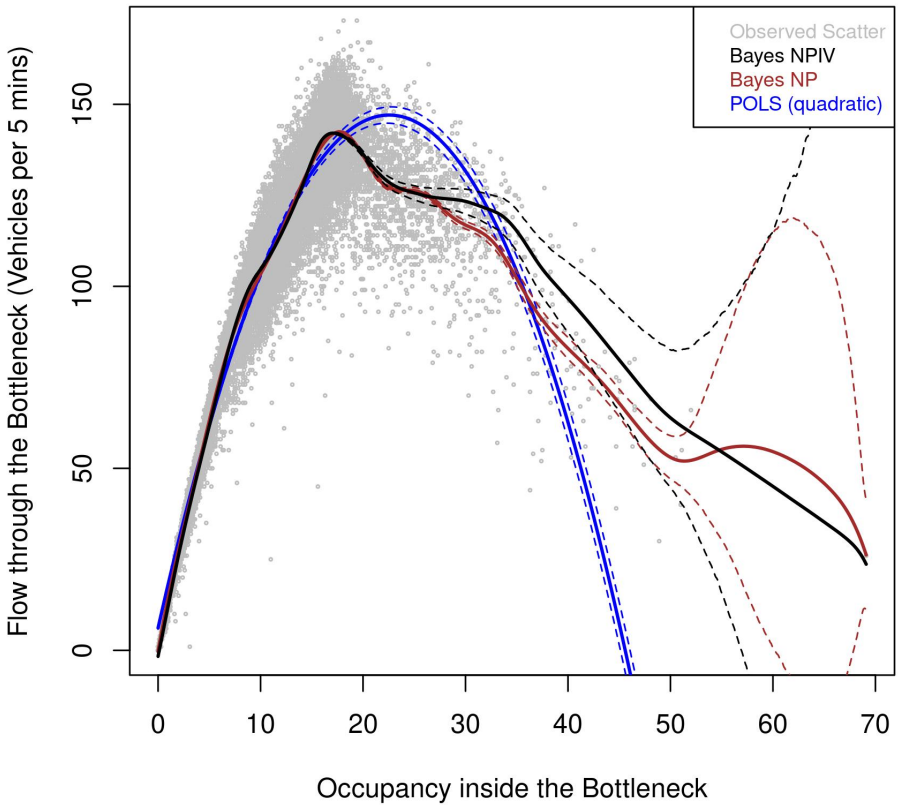
Figure 4.6: Estimated flow-occupancy curves for Eastbound SR-91.



(a) Non-parametric non-IV estimator.



(b) Non-parametric IV-based estimator.



(c) Comparison of different estimators.

Figure 4.7: Estimated flow-occupancy curves for Eastbound SR-12.

Table 4.3: Summary of Results.

(a) Comparison of estimators.

	Estimated Capacity		Estimated Capacity-drop	
Highway Section	Bayes NP	Bayes NPIV	Bayes NP	Bayes NPIV
Westbound SR-24	6564 veh/hr	6156 veh/hr	27.42 percent	7.80 percent
Eastbound SR-91	6420 veh/hr	7080 veh/hr	14.02 percent	n.s.
Eastbound SR-12	1716 veh/hr	1704 veh/hr	11.54 percent	10.92 percent

*n.s. stands for not statistically significant.

(b) Estimated capacity and comparison with the literature.

Highway Section	Estimated Capacity	Capacity reported in the Engineering literature
Westbound SR-24	6156 veh/hr	4100 veh/hr (Chung et al. 2007)
Eastbound SR-91	7080 veh/hr	7200 veh/hr (Oh & Yeo 2012)
Eastbound SR-12	1704 veh/hr	NA

*NA stands for not available.

(c) Estimated capacity-drop and comparison with the literature.

	Estimated	Average Capacity-drop as reported in the	
Highway Section	Capacity-drop	Engineering literature	Economics literature
Westbound SR-24	7.80 percent	5.10 to 8.40 percent	n.s.
Eastbound SR-91	n.s.	13.50 percent	NA
Eastbound SR-12	10.92 percent	NA	n.s.

*n.s. stands for not statistically significant; NA stands for not available.

(d) Activation of the bottleneck.

	Occupancy corresponding to capacity	
Highway Section	non-IV-based	IV-based
Westbound SR-24	17.5	17.5
Eastbound SR-91	16.0	17.5
Eastbound SR-12	18.0	17.0

Estimated capacity

Table 4.3a summarises the estimated capacity for each highway section. For Westbound SR-24 that features a lane-drop bottleneck with number of lanes reducing from four to two, the capacity estimated via the Bayesian NP estimator, that is, 547 vehicles per five-minutes or 6564 vehicles per hour, is significantly more that the Bayesian NPIV-based estimate, that is, 513 vehicles per five-minutes or 6156 vehicles per hour (see Figure 4.5). The capacity reported in the engineering literature is 4100 vehicles per hour (refer to Table 4.3b), which is much lower than both of these estimates.

For Eastbound SR-91 that features a merge bottleneck, Bayesian NPIV-based estimate of capacity is 590 vehicles per five-minutes or 7080 vehicles per hour, which is significantly higher that the Bayesian NP-based estimate of 535 vehicles per five-minutes or 6420 vehicles per hour (see Figure 4.6) but is consistent with the value reported in the engineering literature (7200 vehicles per hour, see Table 4.3b).

For Eastbound SR-12 that features a lane-drop bottleneck with number of lanes reducing from two to one, the capacity estimated via the Bayesian NP-based estimator, that is, 143 vehicles per five-minutes or 1716 vehicles per hour, is similar to the Bayesian NPIV-based estimate of 142 vehicles per five-minutes or 1704 vehicles per hour (see Figure 4.7). We do not note any previous estimate of capacity of this section from the literature.

The above comparison does not point towards a clear direction of bias in the Bayesian NP-based estimate of capacity with respect to the Bayesian NPIV-based estimate, rather it varies on a case-by-case basis depending upon the data generating process. Failing to address endogeneity bias leads to an over-estimation, an under-estimation and no difference in the estimated capacity for the first, second and third sections, respectively. We also find the Bayesian NPIV-based estimates to be much closer to the previous estimates from the engineering literature, particularly for Eastbound SR-91. However, a substantial difference between our Bayesian NPIV estimate and the one reported in the engineering literature for Westbound SR-24 can be attributed to the bias in previous estimates due to

minute-to-minute fluctuations in flow which might have caused due to the use of only a few days of observations (Anderson & Davis 2020). We emphasise that our causal estimates of capacity are more representative of the actual capacity value as they are based on several months of observations and also adjusted for any potential confounding biases.

Activation of the bottleneck and estimated capacity-drop

Table 4.3a summarises the estimated capacity-drop in each highway section obtained via the non-IV-based and IV-based estimators. Table 4.3d reports the occupancy values corresponding to the drop in capacity.

For Westbound SR-24, we observe a statistically significant drop of 27.42 percent in capacity at an occupancy level of 17.5 (the point of bottleneck activation) using the Bayesian NP estimator (see Figure 4.5(a)). Figure 4.5(a) also shows a recovery in capacity to a value close to 600 vehicles per five-minutes following which there is a huge drop of about 50 percent. However, the evidence of recovery, followed by another drop, seems to be weak as the credible bands in this region are not tight. On the other hand, the Bayesian NPIV estimates show a statistically significant drop in flow from 513 to 473 vehicles per five minutes at an occupancy level of 17.5 (see Figure 4.5(b)). This fall in capacity corresponds to a statistically significant drop of 7.80 percent.

For Eastbound SR-91, the results of the Bayesian NP-based estimation shown in Figure 4.6(a) suggests a statistically significant drop of 14.02 percent in capacity at an occupancy level of 16.0. However, the Bayesian NPIV-based estimate shown in Figure 4.6(b) illustrates that there is no statistically significant drop in capacity. From this figure, we note a lack of statistical evidence to suggest any change in flow beyond an occupancy level of 17.5.

For Eastbound SR-12, we observe a statistically significant drop of 11.54 percent in capacity at an occupancy level of 17.5 using the Bayesian NP-based estimator (see Figure 4.7(a)). This Bayesian NP estimate of capacity drop is close to the Bayesian NPIV-based

estimate (10.92 percent), which occurs at an occupancy level of 17.0 (see Figure 4.7(b)).

The above comparison shows that the capacity-drop is overestimated by the Bayesian NP-based estimator in two out of the three sections, but the Bayesian NPIV and Bayesian NP estimates concur for the third section. In contrast to a recent study in the economics literature by Anderson & Davis (2020) that rules out the existence of a statistically significant capacity-drop in highway bottlenecks, we do find sufficient statistical evidence of capacity-drop in two out of the three sections. Thus, the existence of capacity drop (or two-capacity phenomenon) must be evaluated on a case-by-case basis. It is worth noting that consistent with Cassidy & Bertini (1999), we also find that in all the three sections, the level of occupancy corresponding to a drop in capacity are almost similar, that is, ~ 17.0.

The estimated congested (or hypercongested) regime of the flow-occupancy curve

Figures 4.5(c), 4.6(c) and 4.7(c)) illustrate that the non-IV-based estimators (Bayesian NP and POLS) underestimate the congested regime of the flow-density curve that lies beyond the capacity point. These figures show that non-IV-based estimators exhibit a statistically significant backward bending relationship between flow and occupancy following the capacity-drop for all the three sections. However, the IV-based based estimator rules out the possibility of any statistically significant changes in flow with increase in occupancy following the capacity-drop in two out of the three sections. For Eastbound SR-12, we do find a statistically significant evidence of a backward bending relationship up to a certain level of occupancy after the initial capacity-drop.

It is worth noting that the IV-based estimates statistically reinforce some previous observations from the engineering literature. Our estimates support the study by Daganzo et al. (1999), which presents empirical evidence to show that the entire backward bending part of the fundamental diagram for a uniform highway section arises due to the presence

of downstream disturbances or weather related events that might affect average driving behaviour. Such obstructions give rise to a predictable flow-density or flow-speed relationship; otherwise, the capacity of the section does not drop even when the demand is high (Daganzo et al. 1999). Our Bayesian NPIV estimates support this observation – we do not observe a backward bending relationship beyond the capacity-drop for Westbound SR-24 and Eastbound SR-91 as these sections are perfectly isolated from any downstream bottlenecks, but a statistically significant backward bending relationship is evident for Eastbound SR-12 because SR-12 merges with I-80 just downstream of the analysed site.

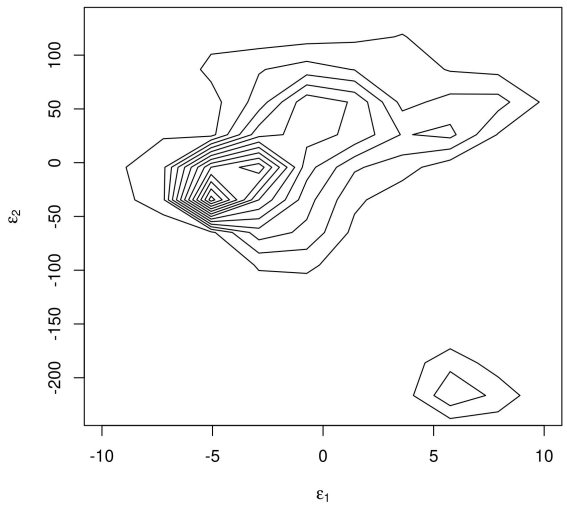
Our empirical IV-based estimates of the flow-occupancy curve are also consistent with the amendment in the fundamental speed-flow relationship proposed in the economics literature by Verhoef (2001). Based on car following theory, Verhoef (2001) found that the entire backward bending part of the speed-flow relationship for a uniform highway section is dynamically unstable. Thus, consistent with Daganzo et al. (1999), the amendment by Verhoef (2001) suggests the absence of a backward bending part in the speed-flow curve, instead points towards a constant outflow from a uniform highway section upon onset of congestion.

These results thus indicate the presence of large endogeneity biases in the non-IV-based estimates of the flow-occupancy curve (particularly in the congested regime), and thus, the advantages of adopting NPIV are apparent.

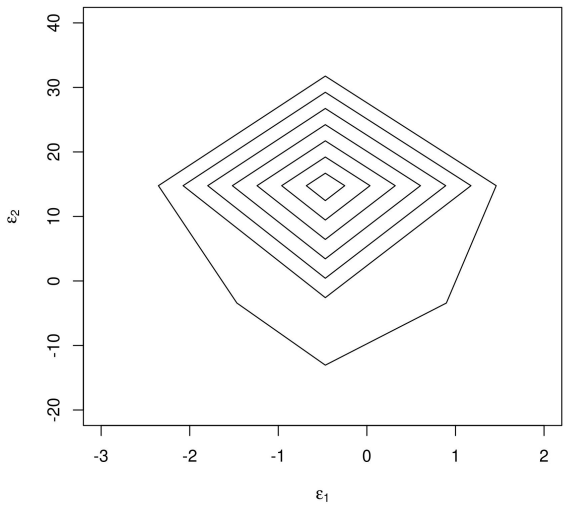
4.5.2 Robustness Tests

Distribution of Errors

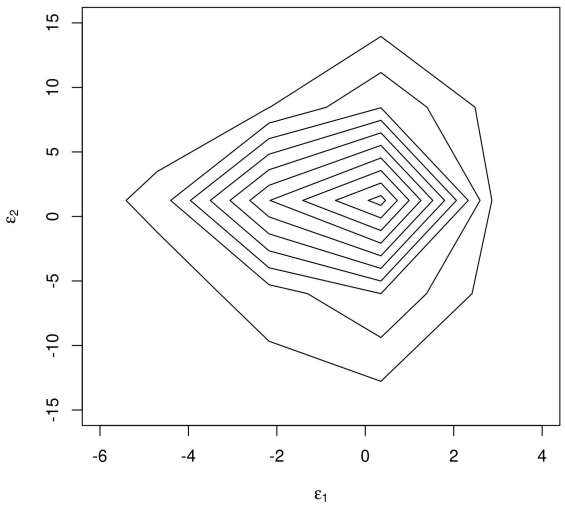
Figure 4.8 shows the contour plot of the joint distribution of errors from the first stage (ϵ_1) and the second stage (ϵ_2). These figures show that the joint error distribution is either uni-modal asymmetric or bi-modal.



(a) Westbound SR-24



(b) Eastbound SR-91



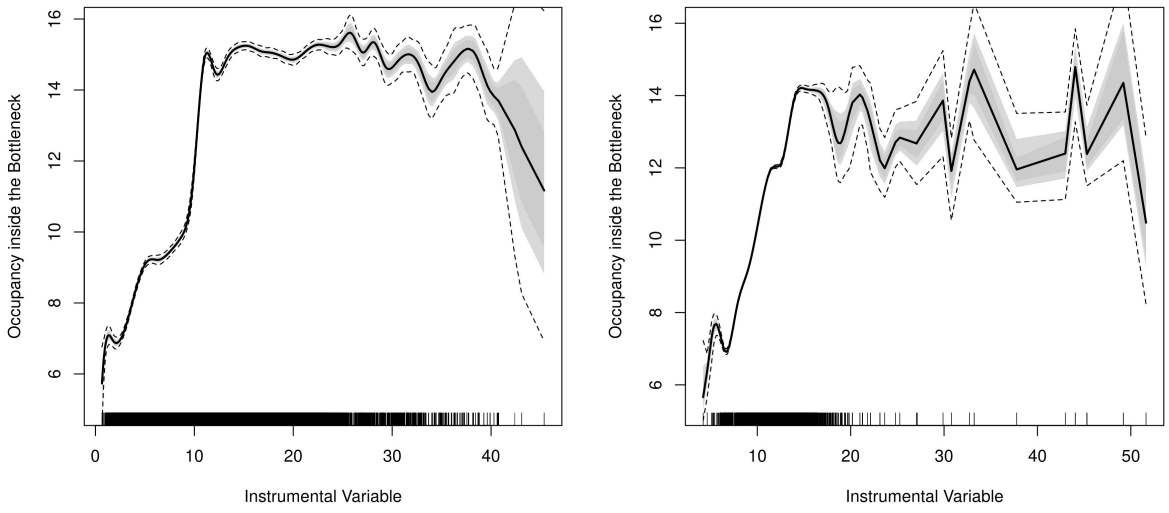
(c) Eastbound SR-12

Figure 4.8: Distribution of errors.

These results suggest that the estimates of $S(.)$ and inference could have poor statistical properties if the error is assumed to follow a uni-modal symmetric and thin-tailed Gaussian error distributions. The adopted Bayesian NPIV method addresses all these potential challenges by allowing for a flexible distribution of errors, instead of assuming a restrictive parametric error distribution.

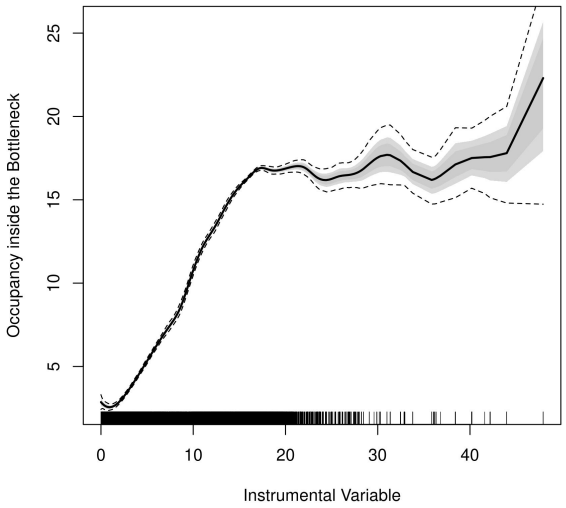
Relevance of Instruments

Figure 4.9 illustrates the results (that is, the estimated $h(.)$) from regression of the endogenous covariate on the instrument for the three highway sections.



(a) Westbound SR-24

(b) Eastbound SR-91



(c) Eastbound SR-12

Figure 4.9: Relevance of instruments.

These figures show a strong correlation between the instrument and the endogenous

covariate for values of the IV less than 15, but appears relatively weaker in the remaining domain of IV for SR-24 and SR-91. We thus carry out complimentary weak instrument tests to evaluate the relevance of the chosen instruments.

To this end, we use the traditional F-tests Stock & Yogo (2005) at different parts of the IV's domain to test for the relevance of the chosen IV at a local level. Specifically, we divide the IV's domain into two bins and carry out the weak instrument test in each bin. The corresponding F-statistics values, along with the one for the entire domain of the IV, are reported in Table 4.4. For all three study sites, F-statistics values across all considered domains of the IV are above the critical value of 10, as reported in Stock & Yogo (2005).

Table 4.4: Summary of results from the Stock and Yogo instrument F-test.

	F-statistic		
Highway Section	for full support of IV	for IV ≤ 15	for IV > 15
Westbound SR-24	5.18×10^4	2.76×10^4	3179.00
Eastbound SR-91	1.98×10^4	3.45×10^4	56.58
Eastbound SR-12	4.82×10^4	3.75×10^4	182.10

Thus, Figure 4.9 and results from the Stock and Yogo weak instrument F-test summarised in Table 4.4 provide supporting evidence that the selected IVs satisfy the relevance condition.

4.6 Conclusions and Future Work

The contributions of this research are two-fold. Our methodological contributions reside in developing a comprehensive understanding of the fundamental relationship of traffic flow in a highway section by adopting a causal econometric framework to determine a novel causal relationship between traffic flow and occupancy in a highway section with a downstream bottleneck. We apply a Bayesian non-parametric instrumental variables (NPIV) estimator on data from three highway bottlenecks in California. The use of NPIV is attractive as it allows us to capture non-linearities in the fundamental relationship with a fully flexible non-parametric specification and adjusts for confounding bias via the inclusion of relevant and exogenous instruments. Such confounding biases may occur because of many external observed or unobserved factors such as driver behaviour, heterogeneous vehicles, weather and demand, that are correlated with both observed traffic variables. We thus deliver a more robust characterisation of traffic flow in a highway section that is reproducible and is not sensitive to these extraneous influences. As a by-product of the estimation, we produce novel quantitative estimates of capacity drop in the three bottlenecks.

Our theoretical contributions emerge from reconciling the economics and engineering approaches to estimate the empirical fundamental relationship of traffic flow. One prominent economic approach is based on a demand-supply framework where users of the highway section are treated as suppliers of travel in the section and outflow from the highway section in turn represents the travel supplied. However, we note that the equivalence of the fundamental relationship of traffic flow and the supply curve for travel in a highway section can only be considered under stationary state traffic conditions, which seldom exist particularly under congested traffic conditions. We thus argue that the demand-supply framework may lead to misrepresentation in developing a causal understanding of the empirical fundamental diagram. We instead adopt causal statistical modelling within the engineering framework which is based on the physical laws that govern the movement of vehicles in a traffic stream.

The above themes are important as a recent study in the economics literature examines the changes in outflow with increasing demand for three different highway bottlenecks in California and finds no evidence of drop in capacity or in other words, hypercongestion during periods of high demand. The study concludes that the fundamental (flow-density of flow-speed) diagram for a highway section should not exhibit a backward bending part and also questions the applicability of traffic control measures and congestion pricing policies that are aimed at regulating demand to avoid hypercongestion. Based on our estimated causal fundamental relationship, we re-evaluate the existence of capacity-drop in highway bottlenecks, which is a well-established phenomenon in the engineering literature.

Our empirical results show a statistically significant decrease in flow upon activation of the bottleneck in two out of three analysed bottlenecks, thus supporting the existence of capacity drop. The estimated capacity-drop varies on a case-to-case basis depending upon the geometry of the bottleneck as well as the characteristics of the average traffic stream passing through it. However, after this drop in capacity, we do not find sufficient statistical evidence to support any changes in flow with further increase in occupancy in isolated highway sections. We thus argue that as the flow through the bottleneck remains constant following the capacity-drop, the flow-occupancy curve is not actually backward bending. However, a statistically-significant backward bending relationship exists only when the highway section is not perfectly isolated from downstream obstacles that cause traffic flow through the section to decrease over occupancy in a predictable way.

It is important to note that the empirical results discussed in this chaper apply only to a highway section with a standard bottleneck. These results are encouraging and the framework can be directly adopted to estimate a causal model of traffic flow for a uniform highway section. Our theoretical conclusions on the association between the fundamental relationship and the analysis of travel supply applies to both of these scenarios, that is, highway section with or without a bottleneck.

Our causal estimates of the fundamental relationship are crucial from a policy point

of view in case of the design of highways and devising traffic control strategies, as these estimates provide a more generalised and robust characterisation of the traffic flow in a highway section and adjusts for any potential confounding biases. Our causal models are, therefore, more suited for standard reference manuals like the highway capacity manual (HCM) and the UK-CoBA. Our theoretical and empirical conclusions also have important implications for deriving highway tolls and congestion pricing policies, which we plan to undertake in future work.

References

Anderson, M. L. & Davis, L. W. (2018), 'Does hypercongestion exist?: New evidence suggests not', *National Bureau of Economic Research* .

Anderson, M. L. & Davis, L. W. (2020), 'An empirical test of hypercongestion in highway bottlenecks', *Journal of Public Economics* **187**, 104197.

Banks, J. H. (1990), 'Flow processes at a freeway bottleneck', *Transportation Research Record* (1287).

Banks, J. H. (1991), 'Two-capacity phenomenon at freeway bottlenecks: a basis for ramp metering?', *Transportation Research Record* (1320).

Bertini, R. L. & Leal, M. T. (2005), 'Empirical study of traffic features at a freeway lane drop', *Journal of Transportation Engineering* **131**(6), 397–407.

Bertini, R. L. & Malik, S. (2004), 'Observed dynamic traffic features on freeway section with merges and diverges', *Transportation Research Record* **1867**(1), 25–35.

Boardman, A. E. & Lave, L. B. (1977), 'Highway congestion and congestion tolls', *Journal of Urban Economics* **4**(3), 340–359.

Cameron, A. C. & Trivedi, P. K. (2005), *Microeconometrics: methods and applications*, Cambridge university press.

Cassidy, M. J. (1998), 'Bivariate relations in nearly stationary highway traffic', *Transportation Research Part B: Methodological* **32**(1), 49–59.

Cassidy, M. J. & Bertini, R. L. (1999), 'Some traffic features at freeway bottlenecks', *Transportation Research Part B: Methodological* **33**(1), 25–42.

Cassidy, M. J. & Rudjanakanoknad, J. (2005), 'Increasing the capacity of an isolated merge by metering its on-ramp', *Transportation Research Part B: Methodological* **39**(10), 896–913.

Ceder, A. & May, A. D. (1976), 'Further evaluation of single-and two-regime traffic flow models', *Transportation Research Record* (567).

Chen, X. M., Li, L. & Shi, Q. (2015), Stochastic fundamental diagram based on headway/spacing distributions, *in* 'Stochastic Evolutions of Dynamic Traffic Flow', Springer, pp. 81–115.

Chib, S., Greenberg, E. & Jeliazkov, I. (2009), 'Estimation of semiparametric models in the presence of endogeneity and sample selection', *Journal of Computational and Graphical Statistics* **18**(2), 321–348.

Chin, H. C. & May, A. D. (1991), 'Examination of the speed-flow relationship at the caldecott tunnel', *Transportation Research Record* **1320**, 75–82.

Chung, K., Rudjanakanoknad, J. & Cassidy, M. J. (2007), 'Relation between traffic density and capacity drop at three freeway bottlenecks', *Transportation Research Part B: Methodological* **41**(1), 82–95.

Coifman, B. (2014*a*), 'Jam occupancy and other lingering problems with empirical fundamental relationships', *Transportation Research Record* **2422**, 104–112.

Coifman, B. (2014*b*), 'Revisiting the empirical fundamental relationship', *Transportation Research Part B: Methodological* **68**, 173–184.

Conley, T. G., Hansen, C. B., McCulloch, R. E. & Rossi, P. E. (2008), 'A semi-parametric bayesian approach to the instrumental variable problem', *Journal of Econometrics* **144**(1), 276–305.

Couture, V., Duranton, G. & Turner, M. A. (2018), 'Speed', *Review of Economics and Statistics* **100**(4), 725–739.

Daganzo, C. F. (1997), *Fundamentals of transportation and traffic operations*, Vol. 30, Pergamon Oxford.

Daganzo, C. F. (2002), 'A behavioral theory of multi-lane traffic flow. part ii: Merges and the onset of congestion', *Transportation Research Part B: Methodological* **36**(2), 159–169.

Daganzo, C. F., Cassidy, M. J. & Bertini, R. L. (1999), 'Possible explanations of phase transitions in highway traffic', *Transportation Research Part A: Policy and Practice* **33**(5), 365–379.

Diakaki, C., Papageorgiou, M. & McLean, T. (2000), 'Integrated traffic-responsive urban corridor control strategy in glasgow, scotland: Application and evaluation', *Transportation Research Record* **1727**(1), 101–111.

Drake, J. L. & Schofer, J. L. (1966), 'A statistical analysis of speed-density hypotheses', *Highway Research Record 154* pp. 53–87.

Drake, J. S. (1967), 'A statistical analysis of speed density hypothesis', *HRR* **154**, 53–87.

Edie, L. C. (1961), 'Car-following and steady-state theory for noncongested traffic', *Operations research* **9**(1), 66–76.

Fosgerau, M. & Small, K. A. (2013), 'Hypercongestion in downtown metropolis', *Journal of Urban Economics* **76**, 122–134.

Greenberg, H. (1959), 'An analysis of traffic flow', *Operations research* **7**(1), 79–85.

Greenshields, B. D., Channing, W. & Miller, H. (1935), A study of traffic capacity, *in* 'Highway research board proceedings', Vol. 14, National Research Council (USA), Highway Research Board, p. 448–477.

Hadiuzzaman, M., Siam, M. R. K., Haque, N., Shimu, T. H. & Rahman, F. (2018), 'Adaptive neuro-fuzzy approach for modeling equilibrium speed–density relationship', *Transportmetrica A: Transport Science* **14**(9), 784–808.

Hall, F. L. & Agyemang-Duah, K. (1991), 'Freeway capacity drop and the definition of capacity', *Transportation Research Record* (1320).

Hall, F. L., Allen, B. L. & Gunter, M. A. (1986), 'Empirical analysis of freeway flow-density relationships', *Transportation Research Part A: General* **20**(3), 197–210.

Hall, F. L. & Hall, L. M. (1990), 'Capacity and speed-flow analysis of the queen elizabeth way in ontario', *Transportation Research Record* (1287).

Hall, F. L., Hurdle, V. F. & Banks, J. H. (1993), 'Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways', *Transportation Research Record* **1365**, 12–18.

Hall, J. D. (2018*a*), 'Can tolling help everyone? estimating the aggregate and distributional consequences of congestion pricing'.
URL: *individual.utoronto.ca/jhall/documents/Can_Tolling_Help_Everyone.pdf*

Hall, J. D. (2018*b*), 'Pareto improvements from lexus lanes: The effects of pricing a portion of the lanes on congested highways', *Journal of Public Economics* **158**, 113–125.

HCM (2016), 'Highway capacity manual: A guide for multimodal mobility analysis', *Transportation Research Board, Washington, DC* .

Horowitz, J. L. (2011), 'Applied nonparametric instrumental variables estimation', *Econometrica* **79**(2), 347–394.

Jabari, S. E., Zheng, J. & Liu, H. X. (2014), 'A probabilistic stationary speed–density relation based on newell's simplified car-following model', *Transportation Research Part B: Methodological* **68**, 205–223.

Jin, W.-L., Gan, Q.-J. & Lebacque, J.-P. (2015), 'A kinematic wave theory of capacity drop', *Transportation Research Part B: Methodological* **81**, 316–329.

Kerner, B. S. (2009), *Introduction to modern traffic flow theory and control: the long road to three-phase traffic theory*, Springer Science & Business Media.

Kidando, E., Karaer, A., Kutela, B., Kitali, A. E., Moses, R., Ozguven, E. E. & Sando, T. (2020), 'Novel approach for calibrating freeway highway multi-regimes fundamental diagram', *Transportation Research Record* p. 0361198120930221.

Kidando, E., Kitali, A. E., Lyimo, S. M., Sando, T., Moses, R., Kwigizile, V. & Chimba, D. (2019), 'Applying probabilistic model to quantify influence of rainy weather on stochastic and dynamic transition of traffic conditions', *Journal of Transportation Engineering, Part A: Systems* **145**(5), 04019017.

Kidando, E., Moses, R. & Sando, T. (2019), 'Bayesian regression approach to estimate speed threshold under uncertainty for traffic breakdown event identification', *Journal of Transportation Engineering, Part A: Systems* **145**(5), 04019013.

Kockelman, K. M. (2001), 'Modeling traffic's flow-density relation: Accommodation of multiple flow regimes and traveler types', *Transportation* **28**(4), 363–374.

Kondyli, A., George, B. S., Elefteriadou, L. & Bonyani, G. (2017), 'Defining, measuring, and modeling capacity for the highway capacity manual', *Journal of Transportation Engineering, Part A: Systems* **143**(3), 04016014.

Laval, J. A. & Daganzo, C. F. (2006), 'Lane-changing in traffic streams', *Transportation Research Part B: Methodological* **40**(3), 251–264.

Leclercq, L., Laval, J. A. & Chiabaut, N. (2011), 'Capacity drops at merges: An endogenous model', *Procedia-Social and Behavioral Sciences* **17**, 12–26.

Li, J., Chen, Q.-Y., Wang, H. & Ni, D. (2012), 'Analysis of lwr model with fundamental diagram subject to uncertainties', *Transportmetrica* **8**(6), 387–405.

Liu, X., Xu, J., Li, M., Wei, L. & Ru, H. (2019), 'General-logistic-based speed-density relationship model incorporating the effect of heavy vehicles', *Mathematical Problems in Engineering* **2019**.

Mahnke, R. & Kaupužs, J. (1999), 'Stochastic theory of freeway traffic', *Physical Review E* **59**(1), 117.

May, A. D. (1990), *Traffic flow fundamentals*, Englewood Cliffs, N.J. : Prentice Hall.

Munjal, P. & Pipes, L. A. (1971), 'Propagation of on-ramp density perturbations on unidirectional two-and three-lane freeways', *Transportation Research/UK/* .

Muralidharan, A., Dervisoglu, G. & Horowitz, R. (2011), 'Probabilistic graphical models of fundamental diagram parameters for simulations of freeway traffic', *Transportation research record* **2249**(1), 78–85.

Newbery, D. M. (1989), 'Cost recovery from optimally designed roads', *Economica* pp. 165–185.

Newell, G. F. (1993), 'A simplified theory of kinematic waves in highway traffic, part ii: Queueing at freeway bottlenecks', *Transportation Research Part B: Methodological* **27**(4), 289–303.

Newey, W. K. & Powell, J. L. (2003), 'Instrumental variables estimation of nonparametric models', *Econometrica* **71**(5), 1565–1578.

Newman, L. (1961), 'Study of traffic capacity and delay at the merge of the north sacramento and elvas freeways', *Final Report, California Division of Highways, USA* .

Ni, D. (2015), *Traffic flow theory: Characteristics, experimental methods, and numerical techniques*, Butterworth-Heinemann.

Oh, S. & Yeo, H. (2012), 'Estimation of capacity drop in highway merging sections', *Transportation Research Record* **2286**(1), 111–121.

Oh, S. & Yeo, H. (2015), 'Impact of stop-and-go waves and lane changes on discharge rate in recovery flow', *Transportation Research Part B: Methodological* **77**, 88–102.

Payne, H. J. (1977), 'Discontinuity in equilibrium freeway traffic flow', *Transportation Research Record* .

Persaud, B. N. (1987), Study of a freeway bottleneck to explore some unresolved traffic flow issues., PhD thesis.

Persaud, B., Yagar, S. & Brownlee, R. (1998), 'Exploration of the breakdown phenomenon in freeway traffic', *Transportation Research Record* **1634**(1), 64–69.

Pipes, L. A. (1966), 'Car following models and the fundamental diagram of road traffic', *Transportation Research/UK/* .

Qu, X., Wang, S. & Zhang, J. (2015), 'On the fundamental diagram for freeway traffic: a novel calibration approach for single-regime models', *Transportation Research Part B: Methodological* **73**, 91–102.

Qu, X., Zhang, J. & Wang, S. (2017), 'On the stochastic fundamental diagram for freeway traffic: model development, analytical properties, validation, and extensive applications', *Transportation research part B: methodological* **104**, 256–271.

Siebel, F., Mauser, W., Moutari, S. & Rascle, M. (2009), 'Balanced vehicular traffic at a bottleneck', *Mathematical and Computer Modelling* **49**(3-4), 689–702.

Small, K. A. & Chu, X. (2003), 'Hypercongestion', *Journal of Transport Economics and Policy (JTEP)* **37**(3), 319–352.

Small, K. A. & Verhoef, E. T. (2007), *The economics of urban transportation*, Routledge, New York.

Smaragdis, E., Papageorgiou, M. & Kosmatopoulos, E. (2004), 'A flow-maximizing adaptive local ramp metering strategy', *Transportation Research Part B: Methodological* **38**(3), 251–270.

Sopasakis, A. (2004), 'Stochastic noise approach to traffic flow modeling', *Physica A: Statistical Mechanics and its Applications* **342**(3-4), 741–754.

Srivastava, A. & Geroliminis, N. (2013), 'Empirical observations of capacity drop in freeway merges with ramp control and integration in a first-order model', *Transportation Research Part C: Emerging Technologies* **30**, 161–177.

Stock, J. & Yogo, M. (2005), *Testing for Weak Instruments in Linear IV Regression*, Cambridge University Press, New York, pp. 80–108.

Sun, L. & Zhou, J. (2005), 'Development of multiregime speed–density relationships by cluster analysis', *Transportation Research Record* **1934**(1), 64–71.

Verhoef, E. T. (2001), 'An integrated dynamic model of road traffic congestion based on simple car-following theory: exploring hypercongestion', *Journal of Urban Economics* **49**(3), 505–542.

Walters, A. A. (1961), 'The theory and measurement of private and social cost of highway congestion', *Econometrica: Journal of the Econometric Society* pp. 676–699.

Wang, H., Li, J., Chen, Q.-Y. & Ni, D. (2011), 'Logistic modeling of the equilibrium speed–density relationship', *Transportation research part A: policy and practice* **45**(6), 554–566.

Wang, H., Ni, D., Chen, Q.-Y. & Li, J. (2013), 'Stochastic modeling of the equilibrium speed–density relationship', *Journal of Advanced Transportation* **47**(1), 126–150.

Wiesenfarth, M., Hisgen, C. M., Kneib, T. & Cadarso-Suarez, C. (2014), 'Bayesian nonparametric instrumental variables regression based on penalized splines and dirichlet process mixtures', *Journal of Business and Economic Statistics* **32**(3), 468–482.

Wooldridge, J. M. (2010), *Econometric analysis of cross section and panel data*, MIT press.

Yuan, K., Knoop, V. L. & Hoogendoorn, S. P. (2015), 'Capacity drop: Relationship between speed in congestion and the queue discharge rate', *Transportation Research Record* **2491**(1), 72–80.

Zhang, L. & Levinson, D. (2004), 'Some properties of flows at freeway bottlenecks', *Transportation Research Record* **1883**(1), 122–131.

Chapter 5

Understanding the production of travel in urban road networks

This chapter quantifies the production of vehicular travel in urban road networks. To do so, it estimates macroscopic fundamental relationships for homogeneously congested sub-networks (reservoirs) in thirty-four cities around the globe. We adopt a causal approach to obtain unbiased estimates of the reservoir-level flow-density relationship using large-scale traffic sensor data. In particular, we apply a Bayesian non-parametric spline-based regression approach with instrumental variables to adjust for potential confounding/endogeneity biases due to simultaneity and/or omitted variables such as vehicle interactions and traffic controls. Our estimates suggest that the provision of vehicular travel in cities is subject to decreasing returns to density and network size. As a by product of the estimation, we also deliver estimates of important traffic control inputs such as capacity and critical occupancy for these reservoirs. Our results are important both for traffic engineers and transport economists. The core findings of this chapter are under review as:

> Anupriya, Graham, D.J. & Bansal, P. (under review). Understanding the production of travel in urban road networks. in *The Review of Economics and Statistics*

5.1 Introduction

Road transportation is vital for economic development as it provides a major means to transport people from one place to another. However, an unprecedented rise in urban population has led to an unmanageable increase in road traffic, resulting in high levels of congestion. The transport consulting firm INRIX reported that the direct and indirect costs of congestion in the US cities amounted to 88 billion US dollars in 2019 (INRIX 2019). Providing sustainable urban transportation solutions to limit congestion thus remains one of the most pressing challenge for governments and decision makers all around the globe (UN-DESA 2018). While investing in road or public transportation infrastructure are both viable solutions, they each require significant spending of public money. Thus, understanding the efficiency that arise from increasing the provision of such services in cities continues to be an important area of research. To this end, Anupriya et al. (2020), Graham (2008) and Savage (1997), among others, analyse the existence of scale economies in short-run urban rail transport (metro) operations and find that metro systems are highly productive in areas with high density, for instance, in city centres. However, empirical evidence on existence of such scale economies in the provision of road-based urban transport is limited.

In this study, we aim to address this gap by quantifying the technical efficiency of production of travel in urban road networks. An integral part of this analysis is to characterise the technology driving congestion in the road network because in transport economics, user time is the primary input factor to the production of vehicular travel on roads (Small & Verhoef 2007). Small & Chu (2003) suggest that modelling congestion in an urban road network requires a complete dynamic representation of traffic flow in the network. Developing such a model will require very detailed inputs including (i) highly dis-aggregated and time-dependent origin-destination (OD) demand data, and (ii) a psychological model representing driver's route choice in existing and anticipated congestion conditions (Daganzo 2005). However, highly congested networks can be very

sensitive to small changes in input OD table or disturbances in driver's route choice, thus posing strict requirements to precisely estimate numerical values for (i) and (ii). Considering these challenges, Daganzo (2005, 2007) propose an alternative observation-based approach that uses aggregate measures of traffic flow, or in other words, the key determinants of city mobility: (a) aggregate vehicular accumulations, and (b) cumulative traffic outflows from spatial units (e.g., district) by time-of-day. Daganzo (2005) suggests that these measures are potentially ideal policy indicators because they do not need any models for disaggregate-level measurements, and they show high correlations with measures like the aggregate number of vehicle-hours-travelled (VHT) and the aggregate vehicle-miles-travelled (VMT) that are central to describing the physics of congestion. For all these reasons, we also adopt the observation-based framework to model the congestion technology in an urban road network. This framework leads us to estimate a macroscopic fundamental relationship between the total outflow of traffic from a homogeneously congested system (or reservoir) and its aggregate accumulation.

To empirically estimate this relationship, the engineering literature mostly adopts a pooled ordinary least squares (OLS) based regression and fits a curve to the observed data of traffic variables (see, for instance, Kouvelas et al. 2017, Geroliminis et al. 2014, Mariotte et al. 2017, among others). We argue that this estimated relationship, however, is *potentially spurious* due to several possible sources of endogeneity/confounding biases. For instance, there are many external factors such as applied traffic controls, route choice and driver adaptation, that are correlated with the observed traffic variables (Geroliminis & Sun 2011, Mahmassani et al. 2013, Leclercq et al. 2014), but are often omitted in the estimation of this relationship. Fitting a pooled OLS regression curve to the observed scatter plot of traffic variables fails to adjust for the above-mentioned sources of confounding, which may bias the estimated relationship (Wooldridge 2010, Cameron & Trivedi 2005).

To address these shortcomings of this widely-adopted engineering approach, in this chapter, we empirically estimate the macroscopic fundamental relationship using a flexible

causal framework. In particular, we adopt a Bayesian non-parametric instrumental variables (NPIV) estimator (Wiesenfarth et al. 2014) that allows us to (1) capture non-linearities in the relationship with a non-parametric specification that does not require an assumed a-priori functional form; and, (2) adjust for any confounding/ simultaneity bias via the use of instrumental variables (IVs). We apply this approach on a unique large-scale multi-city traffic sensor dataset to estimate empirical macroscopic fundamental diagrams (MFDs) for thirty-four homogeneously congested networks in eleven cities across the globe. From the estimated MFDs, we derive novel causal estimates of returns to density (RTD) and returns to network size (RTS) for the analysed networks (or reservoirs).

We emphasise that our causal approach to empirical estimation of the MFD is based on the physics of movement of vehicles in a reservoir as proposed by Daganzo (2005). Some studies in transport economics have also adopted a causal framework for this problem (see, for instance, Couture et al. 2018, Akbar & Duranton 2017, Russo et al. 2019), but based on the interpretation of the speed-flow fundamental relationship in an urban network as the supply curve for travel in the network under stationary (steady) state traffic conditions (Small & Verhoef 2007). Accordingly, Akbar & Duranton (2017) and Russo et al. (2019) use exogenous shifters in demand as instruments to estimate an underlying supply-side relationship from the observed data on speed and flow. We argue that the economic representation of this model as a supply curve can lead to ambiguity since it requires stationary state traffic conditions, which seldom exist, particularly under congested conditions (Daganzo 1998). Consequently, the estimates of RTS derived in Couture et al. (2018) may also be non-representative of the production efficiency of urban road networks.

The rest of this chapter is organised as follows. Section 5.2 reviews the relevant engineering and economics literature on empirical estimation of the MFD. Section 5.3 demonstrates the equivalence of the macroscopic fundamental relationship and the production function for travel in an urban road network. Section 5.4 details the model

specification and explains the econometric method used in this study and describes the data processing and summary statistics of important variables. Section 5.5 presents our results and benchmarks them against those derived using a standard non-parametric estimator without adjustment for endogeneity. Conclusions and implications are discussed in the final section.

5.2 Literature Review

Daganzo (2005, 2007) extend the idea of the existence of a macroscopic model of steady state urban traffic, which was originally proposed by Herman & Prigogine (1979) and further developed by Ardekani & Herman (1987), to predict outflows in a dynamic environment. Daganzo (2005) hypothesises that under slowly varying demand conditions (that is, near steady state conditions) and recurring traffic patterns, the functional relationship between network accumulation and traffic parameters like average flows and average vehicular speeds is insensitive to the OD demands and thus could be viewed as properties of the network itself. Consequently, Daganzo (2007) proposed that a macroscopic relationship exists between total outflow from the system and its aggregate accumulation 'n' for a single system or 'reservoir' or 'neighbourhood'. The dynamics of a reservoir can be described by:

$$\frac{dn}{dt} = f(t) - G(n(t)), \quad for \quad t \geq 0. \tag{5.1}$$

where, $f(t)$ and $n(t)$ describe the input flow and the accumulation in the system at time 't'. $G(n(t))$ represents a non-negative, uni-modal 'exit' function. Daganzo (2007) assumes that this function applies to both steady state and when conditions change smoothly over time.

Geroliminis & Daganzo (2007) further prove the existence of macroscopic fundamental diagrams using simulated traffic data for a homogeneously loaded and evenly congested

traffic network. They state that a city can either be modelled as a single or multi-reservoir system depending upon the geometry, the demand patterns, the distributions of trip destinations among the city and the homogeneity in traffic loads. They also validate that the original equation proposed by Daganzo (2007) is robust to different OD demand tables and a range of traffic conditions. Thus, each reservoir can be described by an invariant macroscopic fundamental diagram (MFD) with a well-defined maximum and insensitive to demand changes. Geroliminis & Daganzo (2008) back up the findings on the existence of MFDs by using observational datasets from detectors and GPS-equipped taxis giving a full representation of traffic network in Yokohama, Japan. They also find that a fixed relationship exists between the network-level space-mean flows and the trip completion rates, which dynamically measure travel production. Previous studies suggest that the traffic relationship underlying a MFD depends on the attributes of the links (that is, the fundamental diagrams of the individual roads comprising the reservoir), the reservoir layout and signal settings or applied traffic controls, as well as route choice and driver adaptation (see, for instance, Geroliminis & Daganzo 2008, Geroliminis & Sun 2011, Wu et al. 2011, Leclercq & Geroliminis 2013, Mahmassani et al. 2013, Leclercq et al. 2014, Laval & Castrillón 2015, and other references therein).

Although the theoretical side of MFDs is well-established and the existence of MFDs has been demonstrated experimentally, estimating MFDs empirically and obtaining reliable quantitative estimates of scale economies from MFDs is not straightforward. The engineering literature mostly adopts a pooled ordinary least squares (OLS) based regression to fit a curve to the observed data. The adopted OLS estimator has three main issues.

First, in most cases, this approach involves a priori parametric assumptions about the shape of the curve. For instance, Kouvelas et al. (2017), Ramezani et al. (2015), Lamotte & Geroliminis (2018), Amirgholy & Gao (2017) and Zhong et al. (2018) choose a polynomial function, Amirgholy et al. (2017), Ampountolas et al. (2017), Geroliminis et al.

(2014), Zheng & Geroliminis (2016) and Liu & Geroliminis (2017) choose an exponential function, Mariotte et al. (2017) and Gao & Gayah (2017) choose a multi-regime function, and, Ambühl et al. (2018) choose a trapezoidal function to obtain an empirical estimate of the MFD. We argue that such an analysis that presumes pre-defined functional forms may fail to capture the non-linearities in the MFD, thus producing biased estimates of the key features of the reservoir such as capacity.

Second, many unaccounted factors may lead to a confounding bias in the empirical estimate of the MFD obtained via pooled OLS regression. As illustrated in Geroliminis & Daganzo (2008), Buisson & Ladier (2009) and Ambühl et al. (2018), the flow-density scatter-plot for any reservoir reveals a range of flows for any given density, which may arise from: (i) non-steady state behaviour (dynamics) of urban traffic (Mariotte et al. 2017, Gao & Gayah 2017, Gayah & Daganzo 2011), (ii) lower flows in heterogeneously congested traffic as compared to a homogeneously distributed traffic (Daganzo et al. 2011, Ji & Geroliminis 2012, Doig et al. 2013, Mazloumian et al. 2010, Geroliminis & Sun 2011), and, (iii) diurnal variation in the interference to vehicular flows caused by public transport operations, for instance, public transport priority or rigid timetables (Arnet et al. 2015, Castrillon & Laval 2018), among other factors. The confounding bias occurs because these factors are not accounted in regression models but are likely to be highly correlated with the observed traffic state variables underlying the MFD (see Section 5.4.1 for details).

Third, some urban economic studies have pointed out that the pooled OLS estimate of MFD may suffer from a simultaneity bias (Couture et al. 2018, Akbar & Duranton 2017). The main identification challenge is that in real urban networks, the reservoir accumulations and exit flows may be simultaneously determined and any supply shock is likely to affect both. Such supply shocks include road works, accidents, weather shocks and time of travel, among many other factors. Moreover, a shock to exit flows can affect the input of the reservoir and in turn the accumulation of the reservoir if the shock is known to travellers prior to their departure. This is because the shock affects the

decision of the traveller to travel (Akbar & Duranton 2017). Akbar & Duranton (2017) use counterfactual data on travel times to overcome the simultaneity bias. However, their study does not consider homogeneous reservoirs, but rather considers an aggregated analysis of the entire network. As originally argued in the transport engineering literature by Daganzo (2005) and further shown empirically by Buisson & Ladier (2009), Geroliminis & Sun (2011), Mazloumian et al. (2010), among many others, MFDs may not even exist for heterogeneously congested networks. Moreover, as argued in the Introduction Section, empirical estimation of the MFD based on the interpretation of the speed-flow fundamental relationship as the supply curve for travel may lead to ambiguity.

We, thus, aim to merge these two strands in the literature, one from engineering and the other from economics, and estimate the MFD giving due attention to all these issues – parametric functional form assumptions, omitted variable biases, simultaneity biases, and homogeneity considerations. The adopted NPIV approach incorporates non-linearities in the MFD non-parametrically without assuming any pre-defined functional form and addresses different endogeneity biases using IVs. We also use data from homogeneous reservoirs to ensure the existence of MFDs. From the estimated MFDs, we derive the first estimates of RTS and RTD in the literature.

5.3 The Production of Travel in Urban Road Networks

In this section, we discuss the relationships between: (i) aggregate vehicle accumulations and vehicle hours travelled (VHT), and, (ii) flow sums and vehicle miles travelled (VMT). The derivations presented in this section are adopted from Daganzo (2005). The aim of this discussion is to understand the equivalence between the macroscopic fundamental relationship and the production function for travel in an urban road network, thus unifying the economics and engineering interpretations.

We consider a reservoir, r with a set of directed links L^r. We define $n_i(t)$ as the number of vehicles travelling on link i at time t and $A_i(t)$ and $D_i(t)$ as the corresponding cumulative arrivals and departures such that $n_i(t) = A_i(t) - D_i(t)$. Link arrivals and departures consist of exogenous (from/to other links) portions $U_i(t)$ and $L_i(t)$ respectively, representing upstream arrivals and downstream departures. Similarly the respective endogenous (from/to the origins/destinations in the link) portions are $O_i(t)$ and $E_i(t)$, representing the trips originated and ended within i respectively. Thus, $A_i(t) = O_i(t) + U_i(t)$, and $D_i(t) = E_i(t) + L_i(t)$.

5.3.1 Aggregate accumulations and VHT

The total VHT in link i during an infinitesimally small time interval dt equals $n_i(t)dt$, given that no vehicle enters or exits the link. Thus, the total number of vehicle-hours for a time interval t equals $VHT_i = \int_t n_i(t)dt$ or equivalently $VHT_i = \sum_t n_i(t)\Delta t$ for short time slices Δt satisfying the entry-exit condition. The total VHT in a reservoir VHT^r is simply the sum of the VHT_i over all links $i \in A^r$. As Daganzo (2005) suggests, VHT^r can be approximated by sampling $n_i(t)$ every Δt time units and evaluating $VHT_i = \sum_t \sum_i n_i(t)\Delta t$ for $i \in A^r$.

5.3.2 Flow-sums and VMT

We assume that a link i of length l_i is unoccupied at both ends of our time interval of interest $(0, t)$, such that, $A_i(t) = D_i(t)$. We reasonably ignore the number of trips that both begin and end in the link. Then, the total number of endogenous link visits (trips with at least one end rooted in the link) by time t is $O_i(t) + E_i(t)$. Since the total number of link visits (that is, the sum of endogenous link visits and through trips) is $A_i(t) = O_i(t) + U_i(t) = D_i(t) = E_i(t) + L_i(t)$, the number of through trips is: $U_i(t) - E_i(t) = L_i(t) - O_i(t)$.

As every through trip covers a distance of l_i in the link i and assuming that each

endogenous trip covers a distance $l_i/2$, the VMT for link i equals: $l_i(U_i(t) + \frac{1}{2}[O_i(t) - E_i(t)]) = l_i(L_i(t) + \frac{1}{2}[O_i(t) - E_i(t)])$. Thus, the total VMT for a reservoir VMT^r equals $= VMT^r = \sum_i l_i U_i(t) + \sum_i \frac{1}{2} l_i O_i(t) - \sum_i \frac{1}{2} l_i E_i(t)])$. As Daganzo (2005) point out, for highway sections, major arterials and collector streets, the number of endogenous trips are usually much smaller than the number of exogenous trips. Thus, for a city region that includes freeway portions, arterials and spans many blocks, for instance, with a diameter just a few times smaller than a typical trip, the sum of exogenous trips $\sum_i U_i(t)$ is substantially larger than the endogenous trips $O_i(t)$ and $E_i(t)$. Thus, VMT^r approximately equals $VMT^r = \sum_i l_i U_i(t)$, that is, the flow-sums in the reservoir r. However, this approximation does not hold for a reservoir with many local streets as for local streets, the number of endogenous trips are quite high.

5.4 Model and Data

This section is divided into three subsections. In the first subsection, we discuss the model specification and explain potential endogeneity bias in estimation of the MFD. In the second subsection, we briefly describe the Bayesian NPIV method in the context of this study. In the final subsection, we describe the data and the relevant variables used in this analysis.

5.4.1 Model Specification

As discussed in the Introduction section (Section 5.1), we aim to estimate an input-output production relationship between the amount of travel time spent in a road network, that is, vehicle hours travelled (VHT), and the amount of travel produced in the network, or in other words, vehicle miles travelled (VMT). Consistent with Daganzo (2005) and Geroliminis & Daganzo (2008), this relationship is equivalent to estimating a macroscopic

relationship between the weighted average accumulation (density or occupancy[1] and the total outflow (weighted average flow-sums) from the road network, given that the road network is roughly homogeneously congested. Weighted average flow and occupancy are defined as $q^r = \frac{\sum_i q_i l_i}{\sum_i l_i}$ and $o^r = \frac{\sum_i o_i l_i}{\sum_i l_i}$ respectively, where i represents each link (that is, a road lane segment between intersections) in the reservoir r and q_i, o_i, l_i denote its flow, occupancy and length respectively. Geroliminis & Daganzo (2008) suggests that these weighted averages would be space-means for all links in the reservoir r, if the detectors are located at representative locations within each link. This is because for time intervals of the order of a typical traffic cycle, flows are roughly the same regardless of where they are measured within a link.

Accordingly, we estimate a *causal relationship* between weighted average occupancy in the reservoir, o^r_{jt}, in the 5-minutes interval j, $j = 1, .., N$, on a particular day t, $t = 1, ..., T$, and the weighted average flow in the reservoir, q^r_{jt}. We consider q^r_{jt} to be a function of o^r_{jt}, conditional on the properties of individual links in the reservoir, the reservoir layout, applied traffic controls, route choice and driver adaptation, among other factors.

$$q^r_{jt} = S^r(o^r_{jt}) + \eta^r_{jt} + \delta^r_{jt} + \xi^r_{jt} \tag{5.2}$$

where δ_{jt} is the unobserved (to researchers) traffic specific component common to all drivers or any traffic specific operational characteristic that applies to all vehicles in the observed traffic stream. η_{jt} represents the degree of homogeneity of the reservoir. ξ_{jt} represents a idiosyncratic error term representing all random shocks to the dependent variable. The exact structural form of how o^r_{jt} enters the equation is unknown, so we adopt a non-parametric specification $S(.)$ in which the shape of the relationship is delivered from the data and regression splines.

We expect δ_{jt} to be correlated with o^r_{jt}. This correlation follows from the omitted

[1]In line with previous studies (see, for instance, Geroliminis & Daganzo 2008, Ambühl et al. 2018, and other reference therein), we use occupancy o as a proxy for density k as the latter cannot be measured directly via traffic detectors. Density at a detector location equals $k = \frac{o}{s}$, where s is the space-mean effective vehicle length. In their study, Geroliminis & Daganzo (2008) use a value of $s \cong 5.5m$.

variables and simultaneity biases discussed in Section 5.2. Moreover, we also expect η_{jt} to be positively correlated with o^r_{jt} and q^r_{jt}. The unavailability of a measure for δ^r_{jt} and η^r_{jt} may lead to a confounding bias in the estimates of $S(.)$. In particular, in the absence of a suitable measure or proxy for δ_{jt} and η_{jt}, an ordinary least squares estimation may under- or over-estimate $S(.)$ if $S(.)$ is a linear function. Therefore, we adopt a nonparametric instrumental variable (NPIV) regression, which not only enables non-parametric specification of $S(.)$ but also addresses potential/ simultaneity confounding biases.

From the estimated $S(.)$, we deliver estimates of capacity (q^r_c) and critical occupancy (o^r_c). We also produce novel quantitative estimates of the returns to density (RTD), that is, the percentage change in VMT (or, equivalently q^r) with respect to percentage change in VHT (or, equivalently o^r), at two points of the estimated curve (that is, for levels of o^r): (i) at the average level of reservoir occupancy, o^r_{avg}, and, (ii) at the average level of peak-hour reservoir occupancy (that is, average occupancy between 06:30 hours to 09:30 hours and 16:00 hours to 19:00 hours), $o^r_{avg,peak}$. To do so, we use the mid-point formula (Varian 2014). Thus,

$$\begin{aligned} \text{RTD} &= \frac{\text{percentage change in } q^r}{\text{percentage change in } o^r} \\ &= \frac{\frac{q^r_2 - q^r_1}{(q^r_2 + q^r_1)/2}}{\frac{o^r_2 - o^r_1}{(o^r_2 + o^r_1)/2}} \end{aligned} \tag{5.3}$$

where, q^r_1 and q^r_2, and, o^r_1 and o^r_2 denote the respective flows and occupancies at the two points in the vicinity of the point at which we compute a representative value of RTD. We assume the flow-occupancy relationship is approximately linear between these two points.

To quantify the returns to network size (RTS), we pool together data from all reservoirs to estimate an extended version of equation 5.2 which includes a function $f(.)$ of network size l_r (that is, sum of lengths of all links in the reservoir).

$$q_{r,jt} = S(o_{r,jt}) + f(l_r) + \eta_{r,jt} + \delta_{r,jt} + \xi_{r,jt} \tag{5.4}$$

To adjust for the unobserved time-invariant reservoir-specific heterogeneity (for instance, reservoir design), we estimate a correlated random effects model as follows:

$$q_{r,jt} = S_1(o_{r,jt}) + S_2(\bar{o_r}) + f(l_{r,jt}) + \eta_{r,jt} + \delta_{r,jt} + u_r + \xi_{r,jt} \tag{5.5}$$

where, $\bar{o_r}$ represents the mean occupancy of the reservoir and u_r represents the reservoir-specific random effect.

From the estimated $S(.)$ and $f(.)$, we calculate the elasticity of q with respect to o and l at various levels of o and l, using the mid-point formula mentioned above. We add these two elasticities to obtain the value of RTS. The estimated value of RTS represents the effect of equi-proportionate increases in network size and network density on system output (that is, flow-sums). In other words, RTS describes the relationship between system output and the overall scale of operations.

5.4.2 Bayesian Nonparametric Instrumental Variable Approach

To capture the salient features of $S(.)$ in a data-driven manner without making a priori assumptions on the functional form of the relationship, and to address endogeneity/simultaneity biases, we a nonparametric instrumental variable (NPIV) regression. There are several approaches to NPIV regression proposed in the econometrics literature, but such methods have not been considered in the estimation of macroscopic fundamental diagram. Extensive reviews can be found in Newey & Powell (2003) and Horowitz (2011).

Classical (frequentist) NPIV regression approaches are popular in theoretical econometrics (such as, Newey & Powell 2003, Horowitz 2011, Newey 2013, Chetverikov & Wilhelm 2017), but they are challenging to apply in practice due to two main reasons. First, tuning parameters to monitor the flexibility of $S(.)$ are often required to be specified by the

analyst. Second, standard errors are generally computed using bootstrap, making these methods computationally prohibitive for large datasets. Therefore, we adopt a scalable *Bayesian* NPIV approach, proposed by Wiesenfarth et al. (2014), that can produce a consistent estimate of non-parametric $S(.)$, even if the analyst does not observe η_{jt} and δ_{jt}. This Bayesian method addresses both challenges of the frequentist estimation because it *learns* tuning parameters related to $S(.)$ during estimation and uncertainty in parameters estimates is inherently captured by credible intervals (analogous to classical confidence intervals). In addition, it also enables nonparametric specification of the unobserved error component ξ_{jt}, precluding the need for making additional assumptions.

In Section 4.4.4 of Chapter 4, we benchmark the performance of the Bayes NPIV estimator against state-of-the-art estimators in a Monte Carlo study and illustrate its ability to adjust for endogeneity bias and recover complex functional forms of $S(.)$.

Adopted Bayesian NPIV approach

We re-discuss the Bayesian NPIV approach (Wiesenfarth et al. 2014) for a model with a single endogenous covariate, that is,

$$q = S(n) + \epsilon_2, \quad n = h(z) + \epsilon_1 \tag{5.6}$$

Note that ω and ξ are encapsulated in ϵ_2, and z is an instrument for the endogenous regressor n. The relationship between n and z is represented by an unknown functional form $h(.)$ and ϵ_2 is an idiosyncratic random error term. For notational simplicity, we drop time-day subscripts. Bayesian NPIV is a control function approach, and assumes the following standard identification restrictions:

$$E(\epsilon_1|z) = 0 \quad \text{and} \quad E(\epsilon_2|\epsilon_1, z) = E(\epsilon_2|\epsilon_1), \tag{5.7}$$

which yields

$$\begin{aligned} E(q|n,z) &= S(n) + E(\epsilon_2|\epsilon_1, z) = S(n) + E(\epsilon_2|\epsilon_1) \\ &= S(n) + \nu(\epsilon_1), \end{aligned} \tag{5.8}$$

where $\nu(\epsilon_1)$ is a function of the unobserved error term ϵ_1. This function is known as the control function.

To satisfy the identification restrictions presented in equation 4.3, we need an instrumental variable (IV) z. The IV should be (i) exogenous, that is, uncorrelated with ϵ_2; (ii) relevant, that is, correlated with the endogenous covariate o, conditional on other covariates in the model. Due to the absence of suitable external instruments, we use an aggregate lagged level of the endogenous covariate (occupancy) as an instrument, that is, for occupancy observed in the 5-minutes interval j on day t, we consider the observation on the covariate from the same interval j from the previous workday $t-1$ as its instrument. We argue that the occupancy o^r_{jt} in the 5-minutes interval j on day t is correlated with the occupancy $o^r_{j,t-1}$ in the same 5-minutes interval j on the previous day $t-1$. This correlation follows from the influence of time-of-the-day on demand for travel in urban road networks. However, these lagged occupancy values $o^r_{j,t-1}$ are exogenous because they do not directly determine the response variable q^r_{jt} in equation 5.2 and would never feature in the model for that response. To justify the relevance of the considered instrument, we present the estimated $h(.)$ in equation 5.6 in the Results and Discussion Section (Section 5.5.4).

Further details of the Bayesian NPIV estimator along with the estimation practicalities are discussed in Section 4.4.3 of Chapter 4.

5.4.3 Data

We use a unique and extensive dataset comprising billions of vehicle observations from stationary traffic sensors located in forty cities worldwide. This largest publicly available[2] multi-city traffic dataset has been assembled by researchers at ETH Zurich for their work on understanding the traffic capacity of urban networks (Loder et al. 2019). For the purpose of this study, we select thirty-four cities for which atleast three days of observations are recorded in the data[3]. Table 5.1 summarises the data used in this analysis.

The dataset reports at least two out of the three fundamental traffic variables speed, flow and occupancy (proxy for density) collected via the traffic sensors, where occupancy represents the fraction of time a traffic sensor is occupied during an observation period. The dataset is enriched with information on the location (latitude-longitude coordinates) of each detector and the attributes of the link, for instance, length of the link, in which the detector is located.

For empirical estimation of the MFD, we first need to identify homogeneously congested sub-networks or reservoirs within each regional networks. The literature suggests different partitioning algorithms (see, for instance, Ji & Geroliminis 2012, Saeedmanesh & Geroliminis 2016, among others) which allow for systematic zoning of the network based on the variation in density between consecutive links. However, in the dataset used in this study, the spatial coverage of the detectors is limited with respect to the whole city road network, preventing application of such algorithms. In their study, Ambühl et al. (2018) and Loder et al. (2019) define different reservoirs heuristically based on the patterns of flow and density in different parts of the network. We use the reservoirs defined by Loder et al. (2019) for the purpose of our study. In each city, we identify one representative reservoir in the central business district (CBD) for further analyses.

Moreover, we also filter the data to remove observations from malfunctioning detectors,

[2]Available at https://utd19.ethz.ch/.

[3]To adjust for reservoir-specific effects and to derive suitable IVs from the panel nature of the dataset (as explained in Section 5.4.2, we need atleast three days of observations for each reservoir.

Table 5.1: Summary of data.

Sl. No.	City	Country	Detectors	Days
1	Augsburg	Germany	777	20
2	Basel	Switzerland	83	7
3	Bern	Switzerland	769	7
4	Birmingham	United Kingdom	114	6
5	Bolton	United Kingdom	202	22
6	Bordeaux	France	591	7
7	Bremen	Germany	583	14
8	Cagliari	Italy	133	50
9	Constance	Germany	129	7
10	Darmstadt	Germany	393	5
11	Essen	Germany	38	36
12	Graz	Austria	300	10
13	Groningen	Netherlands	55	6
14	Hamburg	Germany	419	105
15	Innsbruck	Austria	49	30
16	Kassel	Germany	601	4
17	London	United Kingdom	5804	22
18	Los Angeles	USA	4072	14
19	Luzern	Switzerland	159	361
20	Madrid	Spain	2123	20
21	Manchester	United Kingdom	221	22
22	Marseille	France	178	32
23	Paris	France	513	366
24	Rotterdam	Netherlands	227	6
25	Santander	Spain	378	3
26	Speyer	Germany	199	14
27	Strasbourg	France	220	25
28	Stuttgart	Germany	298	8
29	Tokyo	Japan	2111	30
30	Torino	Italy	787	21
31	Toronto	Canada	298	61
32	Toulouse	France	910	7
33	Wolfsburg	Germany	405	14
34	Zurich	Switzerland	1225	7

over-sampled lanes, and from detectors located in residential roads as in Loder et al. (2019) and Ambühl et al. (2018). Figures C.1-C.34 attached in the appendix show the scatter plots of weighted average flow versus weighted average occupancy in each reservoir.

5.5 Results and Discussion

This section is divided into four subsections. In the first subsection, we compare the results of the MFD obtained via the adopted Bayesian NPIV with three other estimators: (i) a Bayesian NP estimator, (ii) a pooled ordinary least squares (POLS) with a quadratic specification, and, (iii) a two stage least squares (2SLS) estimator with a quadratic specification. The Bayesian NP estimator is a counterpart of the Bayesian NPIV, which does not address confounding bias (that is, $z = x; \epsilon_1 = 0; h(.)$: identity function in Equation 5.6). In the next subsection, we report the estimates of the capacity and critical density delivered by the estimated MFD. In the penultimate subsection, we discuss the estimates of returns to scale extracted from the empirically estimated macroscopic fundamental diagram (MFD). In the final subsection, we present the estimated kernel error distributions to illustrate the importance of the non-parametric DPM specification. The relevance of our instruments is also demonstrated in this subsection.

5.5.1 Estimated Fundamental Relationship

We present the estimates of $S(.)$ (see equation 5.6, second-stage) using Bayesian NPIV, Bayesian NP, 2SLS and POLS in Figures C.1-C.34 for the different reservoirs. For reasons discussed earlier and demonstrated via simulation, our preferred model is Bayesian NPIV. Most discussion would revolve around comparing results of Bayesian NPIV and its non-IV counter part (that is, Bayesian NP).

From each of these figures, we do not observe any notable differences between the Bayesian NPIV and Bayesian NP estimate of the free-flow regime of the flow-occupancy

curve. In this regime, the Bayesian NPIV estimate of $S(.)$ is as efficient as its Bayesian NP counterpart, as evidenced by tight credible bands in the domain of occupancy where we have sufficient number of observations. However, in most cases, we observe substantial differences near the saturation (capacity) point and in the congested (or hypercongested as per the economics literature) regime of the estimated curve. Moreover, these figures also suggest that if models with a pre-specified functional form (such as POLS and 2SLS regression models with a quadratic specification) have not faithfully represented the shape of the relationship, then they lead to non-representative estimates of capacity, critical density and returns to density. Therefore, these results highlight the limitation of economics and engineering studies such as Couture et al. (2018), Russo et al. (2019) and Ambühl et al. (2018), which use pre-specified functional forms for an empirical MFD. We further discuss the Bayesian NPIV estimates in next sub-sections.

5.5.2 Estimated Capacity and Critical Occupancy

Table 5.2 reports the estimated values of capacity and critical occupancy for different reservoirs obtained via the Bayes NPIV estimator. Note that these values are extracted from that part of the estimated curve where the associated credible bands are tight, that is, the estimated relationship between flow and occupancy is statistically significant. For instance, in Figure C.4, our estimates are based on occupancy level below 0.17.

From Table 5.2, we note that the Bayes NP estimator under-estimates the value of capacity in most cases as compared to the Bayes NPIV estimator. These differences result from the stricter exogeneity assumptions implicit in the former method. Moreover, we also note a substantial variation in the estimated capacities of different reservoirs. Loder et al. (2019) show that a major portion of this variation can be explained by factors such as road and bus network topology.

Table 5.2: Summary of estimated capacity and critical occupancy for different reservoirs.

City	Estimated Capacity (q_c^r)		Critical Occupancy (o_c^r)	
	Bayes NPIV	Bayes NP	Bayes NPIV	Bayes NP
Augsburg	496.41	479.91	27.80	27.80
Basel	394.01	377.54	18.96	18.79
Bern	452.06	45.86	436.95	45.86
Birmingham	568.71	568.68	15.52	12.19
Bolton	220.28	220.28	12.10	12.10
Bordeaux	566.51	563.28	16.00	16.00
Bremen	647.48	647.48	22.42	22.42
Cagliari	529.78	372.81	14.37	13.31
Constance	325.77	325.78	10.16	10.16
Darmstadt	414.98	414.98	36.74	36.74
Essen	513.52	498.68	6.24	6.07
Graz	438.53	438.53	14.60	14.60
Groningen	527.77	539.36	19.24	19.24
Hamburg	339.85	331.41	43.50	43.21
Innsbruck	811.48	897.88	16.48	27.34
Kassel	333.93	326.91	46.03	45.61
London	379.62	376.33	19.36	19.54
Los Angeles	1189.03	1254.06	14.47	14.47
Luzern	615.76	617.57	20.71	18.88
Madrid	765.27	737.10	23.59	23.59
Manchester	625.52	606.21	33.62	33.62
Marseille	658.97	631.79	21.30	21.30
Paris	700.30	663.30	7.50	7.20
Rotterdam	920.75	920.75	33.64	33.64
Santander	780.73	724.63	17.12	17.08
Speyer	479.45	465.58	41.90	41.90
Strasbourg	606.37	554.42	12.69	12.94
Stuttgart	462.18	312.03	9.00	7.74
Tokyo	369.94	369.94	14.41	14.41
Torino	659.96	663.81	20.93	20.93
Toronto	487.32	412.95	19.99	14.06
Toulouse	1009.49	7.99	955.85	11.95
Wolfsburg	837.87	745.67	27.68	27.68
Zurich	340.04	313.63	35.90	35.90

*Capacity values are reported in vehicles/hour-lane.

**Critical occupancy levels are reported in percentages.

5.5.3 Returns to scale

In this subsection, we discuss the (i) returns to density (RTD), and, (ii) returns to network size (RTS) estimates derived from the estimated reservoir-level and city-level MFDs respectively.

Returns to density

As mentioned in Section 5.4.1, we provide estimates of RTD for each reservoir at two different levels of occupancy: (i) at the average level of reservoir occupancy, o^r_{avg}, and, (ii) at the average level of peak-hour reservoir occupancy (that is, average occupancy between 06:30 hours to 09:30 hours and 16:00 hours to 19:00 hours), $o^r_{avg,peak}$. Table 5.3 summarises these estimates.

From this table, we note that the provision of vehicular travel in homogeneously congested reservoirs is subject to a decreasing returns to density in most cases, both at o^r_{avg} and $o^r_{avg,peak}$. Thus, operating a fixed road network at a higher traffic density results in decrease in technical efficiency of the network. At o^r_{avg}, the average estimate of RTD is 0.779 and the corresponding standard deviation is 0.151. At $o^r_{avg,peak}$, the average estimate of RTD is 0.631 and the corresponding standard deviation is 0.208. In addition, we note that the marginal returns to density are diminishing. Furthermore, the table also suggests that heavily congested reservoirs such as London have lower RTD estimates.

While the provision of vehicular travel in cities is subject to decreasing RTD, that is, vehicular travel is technically less efficient in dense city centres, the weight of evidence in the transportation literature supports increasing RTD in public transport services like buses and metros (see, Anupriya et al. 2020, Graham 2008, for further details).

Returns to network size

Figure 5.1 presents the estimates of $S_1(.)$, $S_2(.)$ and $f(.)$ in equation 5.5 obtained using the Bayesian NPIV and the Bayesian NP estimators. Note that the covariates $\bar{o_r}$ and $l_{r,jt}$

Table 5.3: Summary of RTD estimates for different reservoirs.

City	Returns to Density (RTD)					
		at o^r_{avg}			at $o^r_{avg,peak}$	
	o^r_{avg}	Bayes NPIV	Bayes NP	$o^r_{avg,peak}$	Bayes NPIV	Bayes NP
Augsburg	7.11	0.645	0.577	11.33	0.625	0.571
Basel	5.83	0.662	0.655	10.57	0.456	0.424
Bern	14.99	0.852	0.903	24.72	0.707	0.648
Birmingham	9.51	0.669	0.707	12.60	0.228	0.183
Bolton	2.85	0.940	0.786	4.22	0.940	0.786
Bordeaux	5.99	0.712	0.793	11.09	0.332	0.371
Bremen	7.16	0.769	0.769	11.34	0.610	0.610
Cagliari	3.18	1.038	0.968	4.78	0.725	0.542
Constance	3.50	0.832	0.844	5.44	0.825	0.804
Darmstadt	5.14	1.003	1.003	6.79	1.106	1.106
Essen	2.01	0.865	0.754	2.09	0.900	0.784
Graz	5.94	0.667	0.667	9.84	0.464	0.464
Groningen	7.54	0.883	0.883	8.88	0.832	0.819
Hamburg	13.66	0.574	0.543	20.53	0.408	0.412
Innsbruck	6.23	0.953	0.974	9.11	0.818	0.842
Kassel	24.65	0.595	0.652	38.03	0.595	0.296
London	10.61	0.515	0.503	12.25	0.433	0.409
Los Angeles	3.70	0.665	0.720	6.89	0.393	0.388
Luzern	6.16	0.664	0.745	9.53	0.522	0.632
Madrid	9.54	0.781	0.755	11.52	0.378	0.391
Manchester	10.96	0.744	0.755	13.94	0.794	0.751
Marseille	6.83	0.662	0.626	9.13	0.448	0.384
Paris	2.90	0.988	0.932	3.50	0.852	0.806
Rotterdam	14.00	0.811	0.811	20.88	0.605	0.605
Santander	6.52	0.885	0.807	9.19	0.438	0.400
Speyer	12.62	0.615	0.642	20.04	0.636	0.536
Strasbourg	3.22	0.822	0.721	5.62	0.602	0.527
Stuttgart	1.33	1.116	0.832	6.33	0.931	0.252
Tokyo	1.95	0.842	0.842	6.95	0.695	0.695
Torino	7.92	0.844	0.841	11.95	0.743	0.686
Toronto	6.44	0.630	0.551	10.13	0.478	0.374
Toulouse	3.27	0.959	0.876	5.67	0.679	0.530
Wolfsburg	6.80	0.653	0.571	11.31	0.844	0.742
Zurich	9.43	0.616	0.637	17.08	0.418	0.337
Mean RTD	7.338	0.779	0.754	11.273	0.631	0.562
Std. Dev.	4.724	0.151	0.131	7.046	0.208	0.206

*Occupancy levels are reported in percentages.

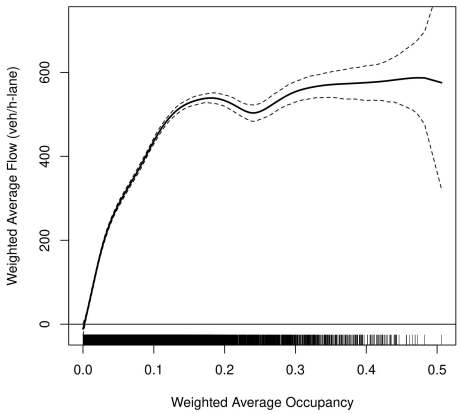
are not continuously distributed, which leads to over-fitting (less precise estimates) of $S_2(.)$ and $f(.)$ in equation 5.5. To improve our estimates, we adopt a quadratic specification for $f(.)$ and re-estimate equation 5.5. From the estimated $S_1(.)$, $S_2(.)$ and $f(.)$, we derive the elasticity of flow with respect to occupancy and network length at the mean level of these covariates. As mentioned in Section 5.4.1, we sum these two elasticities to obtain the RTS estimate. Table 5.4 summarises these results.

Table 5.4: Estimated RTS at average level of covariates.

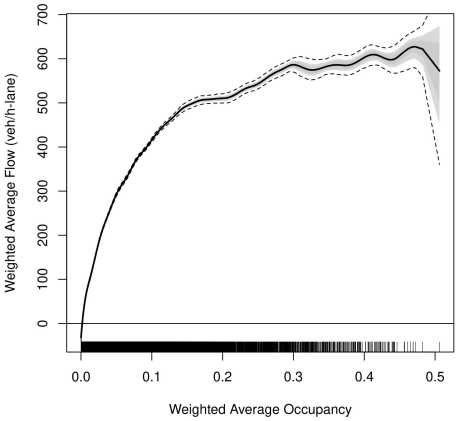
	Average Covariate-level	Bayes NPIV	Bayes NP
Elasticity w.r.t. occupancy	7.54 (%)	0.524	0.635
Elasticity w.r.t. network length	21.44 (km)	-0.224	0.798
Estimated RTS	-	0.300	1.433

Our Bayesian NPIV results suggest that the provision of vehicular travel in cities is subject to decreasing RTS. The estimated RTS equals 0.300. A decreasing RTS implies that private vehicular travel on urban road networks becomes technically less efficient, that is, there is less than proportionate increase in the vehicle kilometres travelled in the network with equi-proportionate increase in traffic density (or equivalently, vehicle hours travelled) and network length. A previous study by Couture et al. (2018) also reports decreasing returns to network size for travel in US cities, although their estimate of -0.04 is negative. We identify two key reasons for the observed difference in the RTS estimates: First, as mentioned in Section 5.2, Couture et al. (2018) derive the RTS estimate from fundamental relationships for the whole network as opposed to considering homogeneous reservoirs within the network. Second, the evidence in their study is limited to data from US cities only.

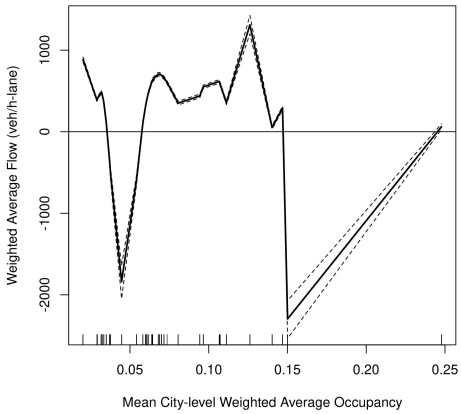
Interestingly, where road-based urban travel is produced with decreasing RTS, evidence in the literature suggests that the provision of public transport services such as metros



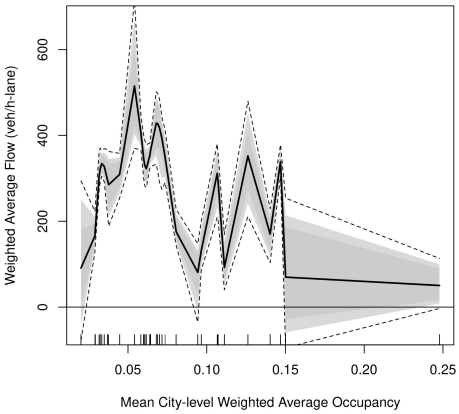
(a) Bayes NP



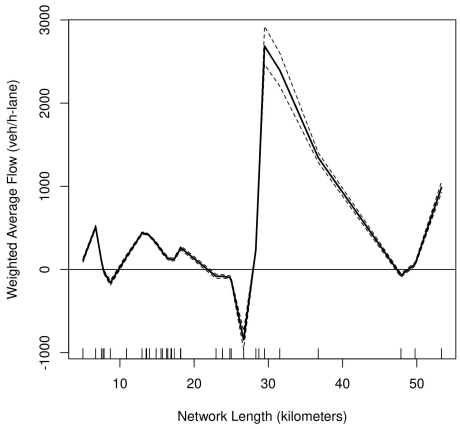
(b) Bayes NPIV



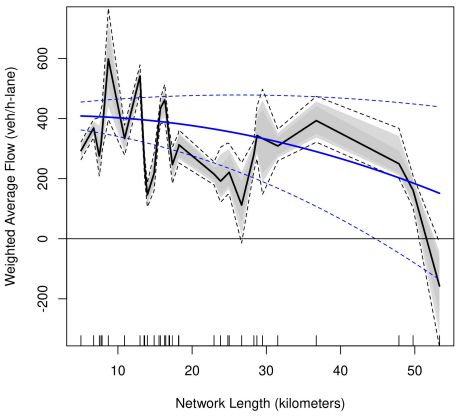
(c) Bayes NP



(d) Bayes NPIV



(e) Bayes NP



(f) Bayes NPIV

Figure 5.1: Returns to Network Size.

and buses is associated with either constant or increasing RTS (see, Anupriya et al. 2020, Graham 2008, for further details).

5.5.4 Robustness Tests

Distribution of errors

Figure 5.2 and Figure C.35 attached in the appendix show the contour plot of the joint distribution of errors from the first stage (ϵ_1) and the second stage (ϵ_2). These figures show that the joint error distribution is either uni-modal asymmetric or bi-modal.

These results suggest that the estimates of $S(.)$ and inference could have poor statistical properties if the error is assumed to follow a uni-modal symmetric and thin-tailed Gaussian error distributions. The adopted Bayesian NPIV method addresses all these potential challenges by allowing for a flexible distribution of errors, instead of assuming a restrictive parametric error distribution.

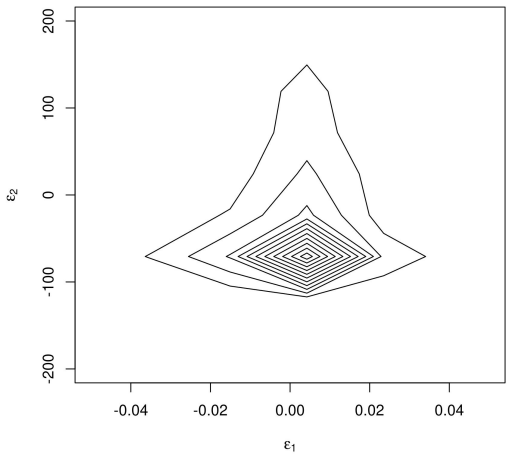


Figure 5.2: Distribution of errors in equation 5.4.

Relevance of instruments

Figure 5.3 and Figure C.36 attached in the appendix illustrate the results (that is, the estimated $h(.)$) from regression of the endogenous covariate on the instrument for the

three highway sections. This figure shows a strong correlation between the instrument and the endogenous covariate for all reservoirs, thus providing satisfactory evidence that the selected IVs satisfy the relevance condition.

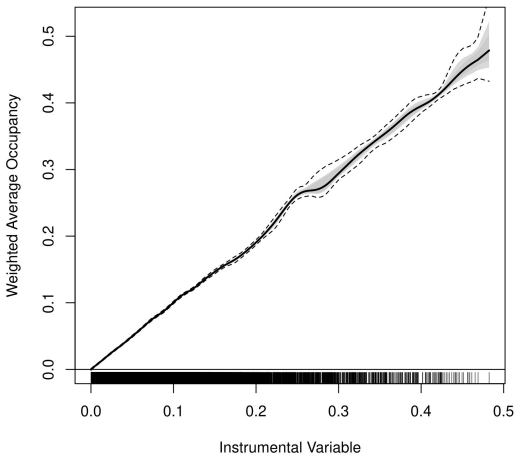


Figure 5.3: Relevance of instruments in equation 5.4.

5.6 Conclusions and Future Work

The contributions of this research are two-fold. Our methodological contributions lie in determining novel causal estimates of the macroscopic fundamental relationship, or equivalently, the production function for travel in urban road networks. We apply a Bayesian non-parametric instrumental variables (NPIV) estimator on a unique large-scale traffic sensor dataset from multiple cities across the globe. The use of NPIV is attractive as it allows us to capture non-linearities in the fundamental relationship with a fully flexible non-parametric specification, and adjusts for confounding bias via the inclusion of relevant and exogenous instruments. Such confounding biases may occur because of many external observed or unobserved factors such as driver adaption and route choice, among others, that are correlated with both observed traffic variables. We thus deliver a more robust characterisation of traffic flow in an urban road network that is reproducible and is not sensitive to these extraneous influences. As a by-product of the estimation, we produce novel quantitative estimates of returns to density and scale in increasing the provision of road-based vehicular travel in cities.

Our theoretical contributions emerge from reconciling the economics and engineering approaches to estimate the macroscopic fundamental relationship of traffic flow. One prominent economic approach is based on a demand-supply framework where users of the road section are treated as suppliers of travel in the section and outflow from the road section in turn represents the travel supplied. However, we note that the equivalence of the macroscopic fundamental relationship and the supply curve for travel in an urban road network can only be considered under stationary state traffic conditions, which seldom exist particularly under congested traffic conditions. We thus argue that the demand-supply framework may lead to misrepresentation in developing a causal understanding of the empirical macroscopic fundamental diagram (MFD). We instead adopt causal statistical modelling within the engineering framework which is based on the physical laws that govern the movement of vehicles in an urban road network.

Our empirical results show the presence of decreasing returns to density in the provision of vehicular travel in cities. Thus, any increase in vehicle hours travelled in a fixed road network results in less than proportionate increase in vehicle kilometres travelled in the network. Across the thirty-four reservoirs analysed, the mean estimate of RTD at the average-level of occupancy is 0.779 and the associated standard deviation is 0.151. At the mean-level of peak-hour occupancy across all reservoirs, the average estimate of RTD is 0.631 with a standard deviation of 0.208. Furthermore, we also find that vehicular travel is produced with decreasing returns to scale in cities. Our estimated RTS of 0.300 implies a less than proportionate increase in the vehicle kilometres travelled in the network with equi-proportionate increase in vehicle hours travelled and network length.

The empirical estimates derived in this study suggest that urban road networks with high density of usage are technically less efficient. This could be taken as evidence in support of policies that aim to reduce car usage in city centres where demand (and therefore, traffic density) is usually high. In addition, the presence of network size dis-economies may be relevant from a policy point of view, particularly for the economic appraisal of large infrastructure projects that lead to network expansion. Decreasing returns to network size implies that such investments may generate external dis-benefits in the form of a network-wide increase in travel time per unit distance for road users. It would be interesting to quantify this external dis-benefit and assess whether it could have a significant impact on the outcome of traditional cost-benefit analyses. We aim to undertake this analysis in future.

Finally, the causal estimates of MFD delivered by this study are crucial from a policy point of view in case of the design of urban road networks, devising traffic control strategies and congestion pricing policies, as these estimates provide a more generalised and robust characterisation of the traffic flow in the network and adjusts for any potential confounding biases.

References

Akbar, P. & Duranton, G. (2017), 'Measuring the cost of congestion in highly congested city: : Bogotá'.
URL: *scioteca.caf.com*

Ambühl, L., Loder, A., Bliemer, M. C., Menendez, M. & Axhausen, K. W. (2018), 'A functional form with a physical meaning for the macroscopic fundamental diagram', *Transportation Research Part B: Methodological* .

Amirgholy, M. & Gao, H. O. (2017), 'Modeling the dynamics of congestion in large urban networks using the macroscopic fundamental diagram: User equilibrium, system optimum, and pricing strategies', *Transportation Research Part B: Methodological* **104**, 215–237.

Amirgholy, M., Shahabi, M. & Gao, H. O. (2017), 'Optimal design of sustainable transit systems in congested urban networks: A macroscopic approach', *Transportation Research Part E: Logistics and Transportation Review* **103**, 261–285.

Ampountolas, K., Zheng, N. & Geroliminis, N. (2017), 'Macroscopic modelling and robust control of bi-modal multi-region urban road networks', *Transportation Research Part B: Methodological* **104**, 616–637.

Anupriya, Graham, D. J., Carbo, J. M., Anderson, R. J. & Bansal, P. (2020), 'Understanding the costs of urban rail transport operations', *Transportation Research Part B: Methodological* **138**, 292–316.

Ardekani, S. & Herman, R. (1987), 'Urban network-wide traffic variables and their relations', *Transportation Science* **21**(1), 1–16.

Arnet, K., Guler, S. I. & Menendez, M. (2015), 'Effects of multimodal operations on urban roadways', *Transportation Research Record* **2533**(1), 1–7.

Buisson, C. & Ladier, C. (2009), 'Exploring the impact of homogeneity of traffic measurements on the existence of macroscopic fundamental diagrams', *Transportation Research Record* **2124**(1), 127–136.

Cameron, A. C. & Trivedi, P. K. (2005), *Microeconometrics: methods and applications*, Cambridge university press.

Castrillon, F. & Laval, J. (2018), 'Impact of buses on the macroscopic fundamental diagram of homogeneous arterial corridors', *Transportmetrica B: Transport Dynamics* **6**(4), 286–301.

Chetverikov, D. & Wilhelm, D. (2017), 'Nonparametric instrumental variables estimation under monotonicity', *Econometrica* **85**(4), 1303–1320.

Couture, V., Duranton, G. & Turner, M. A. (2018), 'Speed', *Review of Economics and Statistics* **100**(4), 725–739.

Daganzo, C. F. (1998), 'Queue spillovers in transportation networks with a route choice', *Transportation Science* **32**(1), 3–11.

Daganzo, C. F. (2005), 'Improving city mobility through gridlock control: an approach and some ideas'.

Daganzo, C. F. (2007), 'Urban gridlock: Macroscopic modeling and mitigation approaches', *Transportation Research Part B: Methodological* **41**(1), 49–62.

Daganzo, C. F., Gayah, V. V. & Gonzales, E. J. (2011), 'Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability', *Transportation Research Part B: Methodological* **45**(1), 278–288.

Doig, J. C., Gayah, V. V. & Cassidy, M. J. (2013), 'Inhomogeneous flow patterns in undersaturated road networks: Implications for macroscopic fundamental diagram', *Transportation research record* **2390**(1), 68–75.

Gao, X. S. & Gayah, V. V. (2017), 'An analytical framework to model uncertainty in urban network dynamics using macroscopic fundamental diagrams', *Transportation research procedia* **23**, 497–516.

Gayah, V. V. & Daganzo, C. F. (2011), 'Clockwise hysteresis loops in the macroscopic fundamental diagram: an effect of network instability', *Transportation Research Part B: Methodological* **45**(4), 643–655.

Geroliminis, N. & Daganzo, C. F. (2007), Macroscopic modeling of traffic in cities, *in* 'TRB 86th annual meeting'.

Geroliminis, N. & Daganzo, C. F. (2008), 'Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings', *Transportation Research Part B: Methodological* **42**(9), 759–770.

Geroliminis, N. & Sun, J. (2011), 'Properties of a well-defined macroscopic fundamental diagram for urban traffic', *Transportation Research Part B: Methodological* **45**(3), 605–617.

Geroliminis, N., Zheng, N. & Ampountolas, K. (2014), 'A three-dimensional macroscopic fundamental diagram for mixed bi-modal urban networks', *Transportation Research Part C: Emerging Technologies* **42**, 168–181.

Graham, D. J. (2008), 'Productivity and efficiency in urban railways: Parametric and non-parametric estimates', *Transportation Research Part E: Logistics and Transportation Review* **44**(1), 84–99.

Herman, R. & Prigogine, I. (1979), 'A two-fluid approach to town traffic', *Science* **204**(4389), 148–151. pmid:17738075.

Horowitz, J. L. (2011), 'Applied nonparametric instrumental variables estimation', *Econometrica* **79**(2), 347–394.

INRIX (2019), Traffic scorecard 2019, Technical report, INRIX, Washington, United States.
URL: *https://inrix.com/press-releases/2019-traffic-scorecard-us/*

Ji, Y. & Geroliminis, N. (2012), 'On the spatial partitioning of urban transportation networks', *Transportation Research Part B: Methodological* **46**(10), 1639–1656.

Kouvelas, A., Saeedmanesh, M. & Geroliminis, N. (2017), 'Enhancing model-based feedback perimeter control with data-driven online adaptive optimization', *Transportation Research Part B: Methodological* **96**, 26–45.

Lamotte, R. & Geroliminis, N. (2018), 'The morning commute in urban areas with heterogeneous trip lengths', *Transportation Research Part B: Methodological* **117**, 794–810.

Laval, J. A. & Castrillón, F. (2015), 'Stochastic approximations for the macroscopic fundamental diagram of urban networks', *Transportation Research Procedia* **7**, 615–630.

Leclercq, L., Chiabaut, N. & Trinquier, B. (2014), 'Macroscopic fundamental diagrams: A cross-comparison of estimation methods', *Transportation Research Part B: Methodological* **62**, 1–12.

Leclercq, L. & Geroliminis, N. (2013), 'Estimating mfds in simple networks with route choice', *Procedia-Social and Behavioral Sciences* **80**, 99–118.

Liu, W. & Geroliminis, N. (2017), 'Doubly dynamics for multi-modal networks with park-and-ride and adaptive pricing', *Transportation Research Part B: Methodological* **102**, 162–179.

Loder, A., Ambühl, L., Menendez, M. & Axhausen, K. W. (2019), 'Understanding traffic capacity of urban networks', *Scientific reports* **9**(1), 1–10.

Mahmassani, H. S., Saberi, M. & Zockaie, A. (2013), 'Urban network gridlock: Theory, characteristics, and dynamics', *Procedia-Social and Behavioral Sciences* **80**, 79–98.

Mariotte, G., Leclercq, L. & Laval, J. A. (2017), 'Macroscopic urban dynamics: Analytical and numerical comparisons of existing models', *Transportation Research Part B: Methodological* **101**, 245–267.

Mazloumian, A., Geroliminis, N. & Helbing, D. (2010), 'The spatial variability of vehicle densities as determinant of urban network capacity', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **368**(1928), 4627–4647.

Newey, W. K. (2013), 'Nonparametric instrumental variables estimation', *American Economic Review* **103**(3), 550–556.

Newey, W. K. & Powell, J. L. (2003), 'Instrumental variables estimation of nonparametric models', *Econometrica* **71**(5), 1565–1578.

Ramezani, M., Haddad, J. & Geroliminis, N. (2015), 'Dynamics of heterogeneity in urban networks: aggregated traffic modeling and hierarchical control', *Transportation Research Part B: Methodological* **74**, 1–19.

Russo, A., Adler, M., Liberini, F. & van Ommeren, J. N. (2019), 'Welfare losses of road congestion'.
URL: *https://ssrn.com/abstract=3416866*

Saeedmanesh, M. & Geroliminis, N. (2016), 'Clustering of heterogeneous networks with directional flows based on "snake" similarities', *Transportation Research Part B: Methodological* **91**, 250–269.

Savage, I. (1997), 'Scale economies in United States rail transit systems', *Transportation Research Part A: Policy and Practice* **31**(6), 459–473.

Small, K. A. & Chu, X. (2003), 'Hypercongestion', *Journal of Transport Economics and Policy (JTEP)* **37**(3), 319–352.

Small, K. A. & Verhoef, E. T. (2007), *The economics of urban transportation*, Routledge, New York.

UN-DESA (2018), The 2018 revision of world urbanization prospects, Technical report, Department of Economic and Social Affairs, United Nations.
URL: *https://esa.un.org/unpd/wup*

Varian, H. R. (2014), *Intermediate Microeconomics: A Modern Approach: Ninth International Student Edition*, WW Norton and Company.

Wiesenfarth, M., Hisgen, C. M., Kneib, T. & Cadarso-Suarez, C. (2014), 'Bayesian nonparametric instrumental variables regression based on penalized splines and dirichlet process mixtures', *Journal of Business and Economic Statistics* **32**(3), 468–482.

Wooldridge, J. M. (2010), *Econometric analysis of cross section and panel data*, MIT press.

Wu, X., Liu, H. X. & Geroliminis, N. (2011), 'An empirical analysis on the arterial fundamental diagram', *Transportation Research Part B: Methodological* **45**(1), 255–266.

Zheng, N. & Geroliminis, N. (2016), 'Modeling and optimization of multimodal urban networks with limited parking and dynamic pricing', *Transportation Research Part B: Methodological* **83**, 36–58.

Zhong, R., Huang, Y., Chen, C., Lam, W., Xu, D. & Sumalee, A. (2018), 'Boundary conditions and behavior of the macroscopic fundamental diagram based network traffic dynamics: A control systems perspective', *Transportation Research Part B: Methodological* **111**, 327–355.

Chapter 6

Congestion in near capacity metro operations: optimum boardings and alightings at bottleneck stations

During peak hours, metro systems often operate at high service frequencies to transport large volumes of passengers. However, the punctuality of such operations can be severely impacted by a vicious circle of passenger congestion and train delays. In particular, high volumes of passenger boardings and alightings may lead to increased dwell times at stations, that may eventually cause queuing of trains in upstream. Such stations act as active bottlenecks in the metro network and congestion may propagate from these bottlenecks to the entire network. Thus, understanding the mechanism that drives passenger congestion at these bottleneck stations is crucial to develop informed control strategies, such as control of inflow of passengers entering these stations. To this end, we conduct the first station-level econometric analysis to estimate a causal relationship between boarding-alighting movements and train flow using data from entry/exit gates and train movement data of the Mass Transit Railway, Hong Kong. We adopt a Bayesian non-parametric spline-based regression approach and apply instrumental variables estimation to control for confounding

bias that may occur due to unobserved characteristics of metro operations. Through the results of the empirical study, we identify bottleneck stations and provide estimates of optimum passenger movements per train and service frequencies at the bottleneck stations. These estimates, along with real data on daily demand, could assist metro operators in devising station-level control strategies. The core findings of this chapter are under revision as:

> Anupriya, Graham, D.J., Bansal, P., Hörcher, D. & Anderson, R.J. (under revision). Congestion in near capacity metro operations: optimum boardings and alightings at bottleneck stations. in *Transportation Research Part C: Emerging Technologies*

6.1 Introduction

As metro systems around the world face an unprecedented growth in peak-hour ridership, passengers increasingly experience congestion and frequent scheduling delays (Tirachini et al. 2013, Seo et al. 2017). For instance, the London Underground reported 504 congestion-related delays of two minutes or more in 2018 (London Assembly 2019), and passengers lost almost 400,000 hours due to these delays (Independent 2017). The congestion in metros can be classified into two main categories: (1) passenger-congestion due to longer boarding and alighting times, and (2) train-congestion due to queuing and reduction in train velocity.

During peak hours, high volumes of passenger boardings and alightings may lead to substantial increases in dwell times of trains at stations, which gives rise to passenger-congestion at stations (Seo et al. 2017). As transit systems are operating at high, often near-capacity service frequencies, increased and irregular dwell times of trains may eventually disrupt service frequencies due to queuing of trains. This queuing phenomenon is referred to as train-congestion or knock-on-delays (Carey & Kwieciński 1994). Since the headway of train arrivals at stations increase as a result of train-congestion, passenger-

congestion at stations intensifies due to further accumulation of passengers on the platform (Seo et al. 2017, Keiji et al. 2015, Daganzo 2009). Thus, passenger-congestion and train-congestion develop into a vicious cycle, and passenger-congestion is generally the root cause of this phenomenon (Seo et al. 2017, Daganzo 2009, Zhang & Wada 2019). The stations where passenger-congestion arises can be characterised as active bottlenecks in the transit network. Based on network configuration and operational attributes, congestion may spread from these bottleneck stations to other parts of the network, resulting in larger overall delays and degradation in system-wide performance.

In recent years, a few studies have modelled the dynamics of metro operations while considering the physical interaction between train-congestion and passenger-congestion (Seo et al. 2017, Zhang & Wada 2019). A similar literature is available for mainline railway operations (Keiji et al. 2015, Wada et al. 2012, and other references therein). These studies suggest various headway-based control strategies to recover the system from knock-on delays such as keeping a moderate separation between trains with necessary adjustment in departure time from origin stations (Keiji et al. 2015), dwelling time extension at some control stations located at the upstream of the bottleneck station (Wada et al. 2012) and increase of free flow speed (Daganzo 2009). However, most of these strategies address train-congestion delays without targeting the root cause of these delays – increased passenger movements at bottleneck stations. Another strand of the literature develops optimisation-based passenger inflow control strategies to minimise the impact of recurrent congestion in metro networks during peak hours. These studies mainly focus on minimising the total waiting time experienced by passengers in the network during peak hours (Shi et al. 2018, Guo et al. 2015, Yuan et al. 2020), reducing the safety risks imposed by passengers waiting on the platform (Jiang et al. 2018, Zou et al. 2018), or minimising the number of stranded passengers (Wang et al. 2020, Yuan et al. 2020).

Unlike previous studies, we focus on understanding the mechanism that drives passenger-congestion at bottleneck stations within an econometric framework. In par-

ticular, we aim to estimate a *causal relationship* between the total number of boardings and alightings per train (passenger movements per train, hereafter), and train flows at each station. Since excessive passenger movement at bottleneck stations is the primary driver of congestion, we expect that a *critical passenger movement* level exists in metros at bottleneck stations, above which train flow or throughput of the station reduces. This intuition is analogous to the road traffic flow theory, which presents evidence of a drop in traffic flow through a road section above a critical vehicular density (see Daganzo 1997, for the fundamental diagram of traffic flow). Therefore, the objective of this exercise is to identify active bottlenecks in the metro network and empirically estimate the optimal passenger movements per train and frequency at bottleneck stations. Such estimates would be instrumental for metro operators in developing data-driven station-based control strategies to avoid both passenger- and train-congestion, and corresponding delays.

To study this congestion phenomenon, we use automated fare collection and train movement data provided by the Mass Transit Railway (MTR), Hong Kong. To find a causal relationship between passenger movements per train and train flows, we adopt an approach that is similar to the estimation of the fundamental diagram of traffic flow in the road traffic theory. However, we argue that estimates derived by simply fitting a pooled ordinary least square regression curve to the observed scatter plot of train flows versus passenger movements per train may be confounded by unobserved characteristics of metro operations such as any existing station-level control measures adopted by the operator. Moreover, the functional form of the estimated causal relationship is not known a priori. Therefore, we adopt a Bayesian nonparametric instrumental variables (NPIV) approach, proposed by Wiesenfarth et al. (2014), that adjusts for such confounding biases. In addition, we also simulate a synthetic metro system to demonstrate the vicious circle of passenger-train-congestion and the importance of estimating optimal passenger movements in developing station-level control strategies. While the main focus of this study remains station-level empirical analysis, the objective of the simulation study is to

provide an intuitive depiction of rail operations using time-space diagrams and illustrate the implications of our empirical analysis in practice.

To summarise, this study contributes with the first station-level empirical analysis of congestion phenomenon in a metro network. Using the estimated relationship between passenger movements and train flow, we identify potential bottleneck stations. We also provide novel estimates of optimal passenger movements at these stations, which could be instrumental for metro operators to develop informed station-level control strategies.

The rest of this chapter is organised as follows. Section 6.2 presents microsimulation of a synthetic metro system under various control strategies. Section 6.3 explains the econometric method used in this study and describes the data processing and summary statistics of important variables. Section 6.4 presents the results of the empirical study. Conclusions and policy relevance of results are discussed in the final section.

6.2 Simulation of passenger congestion and delays

In this section, we reproduce the passenger-train-congestion phenomenon by simulating a typical high-frequency metro operation under pure moving block signalling system[1]. We adopt the model parameters from the study by Yan et al. (2012), which simulates an Asian metro system with near capacity operations. We merge the train operation model from Yan et al. (2012) with a train dwell time model proposed by Zhang & Wada (2019) and Seo et al. (2017) to develop our simulation model. We use this simulation model to demonstrate the efficacy of station-level passenger inflow control measures in avoiding congestion-related delays. The rest of this section is divided into three sub-sections. We describe the train operation model, followed by the main model parameters and the results

[1]We consider a moving block signalling system over a fixed block signalling system as the former permits a more efficient management of queuing and delays by allowing trains to operate at lower headways (Gill 1994, Takeuchi et al. 2003). Several metro lines on the London Underground, the Singapore MRT, the Hong Kong MTR and the New York City subway, among others, use moving block signalling system. Metro systems around the world are increasingly upgrading to such systems to reduce congestion-related delays in the network (Hong Kong MTR 2019).

of the simulation exercise.

6.2.1 The train operation model

To simulate the movement of trains between stations, we adopt a Cellular Automata (CA) model. The CA model was originally developed for simulation of road traffic flow (Nagel & Schreckenberg 1992). However, owing to its ability to reproduce complex real world traffic flow phenomena in a simplistic framework (for instance, see Spyropoulou 2007, Meng & Weng 2011), it has also been widely used to simulate rail traffic flow (refer to Li et al. 2005, Yinping et al. 2008, Xun et al. 2013, Ning et al. 2014, and other references therein).

In the CA model, the rail line i is divided into L cells, each of length 1 metre (that is, $i \in \{1, 2, ...L\}$). The simulation time T_s comprises of discrete time steps t of 1 second each (that is, $t \in \{1, 2, ...T_s\}$). At each time step t, each cell i can either be empty or occupied by the n$^{\text{th}}$ train with integer velocity $v_{n,t}$ (that is, $v_{n,t} \in \{0, 1...v_{max}\}$). Stations are placed at different positions along the line and corresponding dwell time is defined. Each train n is indexed based on its order of entry into the system (that is, $n =\in \{1, 2, ..., N\}$). The position of train n at time t is denoted by $X_{n,t}$. The boundary conditions are open and defined as follows: (i) After each departure interval D, a train with velocity v_{max} enters at the position $i = 1$ given that the train ahead is at a safe breaking distance (given by equation 6.1) from the entry; (ii) At position $i = L$, trains simply exit the system.

At each discrete time step, $t \rightarrow t + 1$, the state of the system is updated according to well-defined rules, mainly governed by the following two situations:

- When the (n-1)$^{\text{th}}$ train is in front of the n$^{\text{th}}$ train at time t, a comparison of the headway distance $\Delta X_{n,t} = X_{n-1,t} - X_{n,t}$ and the minimum instantaneous distance $d_{n,t}$ determines whether the n$^{\text{th}}$ train will accelerate or decelerate in the next time step. The minimum instantaneous distance between successive trains operating

under pure moving-block signalling is given by (Yan et al. 2012):

$$d_{n,t} = \frac{v_{n,t}^2}{2b} + SM \tag{6.1}$$

where, $v_{n,t}$ is the velocity of train n at time t and b is its deceleration. The first term on the right hand side of equation 6.1 represents the breaking distance of train n at time t. A safety margin, SM, is introduced to avoid collision.

- When the n$^{\text{th}}$ train is behind an empty station within the breaking distance, its velocity must vary such that the train can stop at the station. To obtain the updated velocity, we apply the kinematics equation: $v_{n,t+1}^2 - v_o^2 = 2bG_{n,t}$, where v_o is the target velocity which is zero for the train to stop at the station and $G_{n,t}$ is the distance between the station and the train n at time t. As the CA model allows only for integer values of velocity, the velocity update $v_{n,t+1}$ is given by:

$$v_{n,t+1} = \text{int}(\sqrt{2bG_{n,t}}) \tag{6.2}$$

Therefore, the update rules for velocity and position of a train at each time step are as follows:

1. When the n$^{\text{th}}$ train is behind the (n-1)$^{\text{th}}$ train

   Step 1 Velocity update:

   if $\Delta X_{n,t} > d_{n,t}$, $v_{n,t+1} = \min(v_{n,t} + a, v_{max})$;

   elseif $\Delta X_{n,t} < d_{n,t}$, $v_{n,t+1} = \min(v_{n,t} - b, 0)$;

   else $v_{n,t+1} = v_{n,t}$.

   Step 2 Position update:

   $X_{n,t+1} = X_{n,t} + v_{n,t+1}$.

2. When the n$^{\text{th}}$ train is behind a station

(a) When the station is occupied by the (n-1)$^{\text{th}}$ train

The update rules are the same as case 1.

(b) When the station is empty

Step 1 Velocity update:

if $G_{n,t} > d_{n,t}$, $v_{n,t+1} = \min(v_{n,t} + a, v_{max})$;

elseif $G_{n,t} < d_{n,t}$, $v_{n,t+1} = \min(v_{n,t} - b, \text{int}(\sqrt{2bG_{n,t}}), 0)$;

else $v_{n,t+1} = v_{n,t}$.

Step 2 Position update:

$X_{n,t+1} = X_{n,t} + v_{n,t+1}$.

3. When the n$^{\text{th}}$ train is at a station

Step 1 Velocity update:

if $t_{dwell} = T_d$ and $\Delta X_{n,t} > L_s$, $v_{n,t+1} = \min(v_n + a, v_{max})$, $t_{dwell} = 0$;

elseif $t_{dwell} < T_d$, $v_{n,t+1} = 0$, $t_{dwell} = t_{dwell} + 1$.

Step 2 Position update:

$X_{n,t+1} = X_{n,t} + v_{n,t+1}$.

where $L_s = \frac{1}{2a} + SM$ is the safe distance to avoid any collision with the train ahead of the dwelling train, t_{dwell} stores the current dwell time (that is, the time for which the train has stopped at the station until time-step t), and T_d is the planned dwell time.

6.2.2 Model parameters

We consider a metro line of length $L = 4000$ metres with three stations, namely Station 1, Station 2 and Station 3. These stations are located at positions 1000 metres, 2000 metres and 3000 metres respectively. The system is simulated for $T_s = 3600$ seconds. Based on the characteristics of the Asian metro system, we assume that the velocity $v_{n,t}$ of a train in our system varies between 0 and 20 m/s, its acceleration is $a = 1$ m/s^2, and its deceleration is $b = 1$ m/s^2 (Yan et al. 2012). We consider that the safety margin is $SM = 50$ metres.

We assume that the dwell time of trains at Station 1 T_{d_1} and Station 2 T_{d_2} is 30 seconds. We consider Station 3 as an active bottleneck station along the simulated metro line where the dwell time T_{d_3} increases with increasing passenger movements per train N_p (that is, the total boarding and alighting movements). Consistent with Zhang & Wada (2019) and Seo et al. (2017)), we assume that $N_p = A_p H$, where A_p represents the passenger movement rate and H is the time headway of successive trains at Station 3. Following Zhang & Wada (2019) and Keiji et al. (2015), we adopt the following dwell time model for Station 3:

$$T_{d_3} = \begin{cases} 40 \text{ seconds}, & \text{if } (A_p H \leq N_o) \\ 40 + \gamma(A_p H - N_o) \text{ seconds}, & \text{if } (A_p H > N_o) \end{cases}$$

that is, T_{d_3} remains constant until a critical passenger number N_o is reached, following which it starts increasing. γ represents the growth rate of dwell time with the increase in number of passenger movements. We consider γ to be equal to 0.1. Moreover, we assume that the passenger movement rate A_p increases gradually from a value of 0 passenger per second to a maximum value of 10 passengers per second, with an increment of 0.005 passenger per second at each time step.

Furthermore, we assume that the interval D of trains entering into metro system decreases by 5 seconds at every 2 minutes interval, starting from a value of 120 seconds until it attains a value of 60 seconds. This specification implies that with increasing passenger movements, the operator increases train departure frequency up to the maximum value or the capacity value beyond which further increase in frequency is not possible.

6.2.3 Results of simulation

Figure 6.1a shows a time-space diagram representing the train trajectories during the simulation period. From this figure, we note that longer station dwell time of trains at Station 3 eventually leads to queuing of trains or train-congestion in upstream of Station

3, starting at about $t = 2100$ seconds. The queuing-related delays increase the time headway of arrivals of trains at Station 3, which leads to increase in passenger movements because $N_p = A_p H$. As a consequence, station dwell time at Station 3 increases further and queuing of trains in the upstream increases significantly. Based on the spacing of trajectories downstream of Station 3, we note that the throughput of the line first increases as a result of increase in train departure frequency. However, with increasing passenger congestion and delays, this throughput decreases substantially. The number of trains that pass through the system is 39 in an hour.

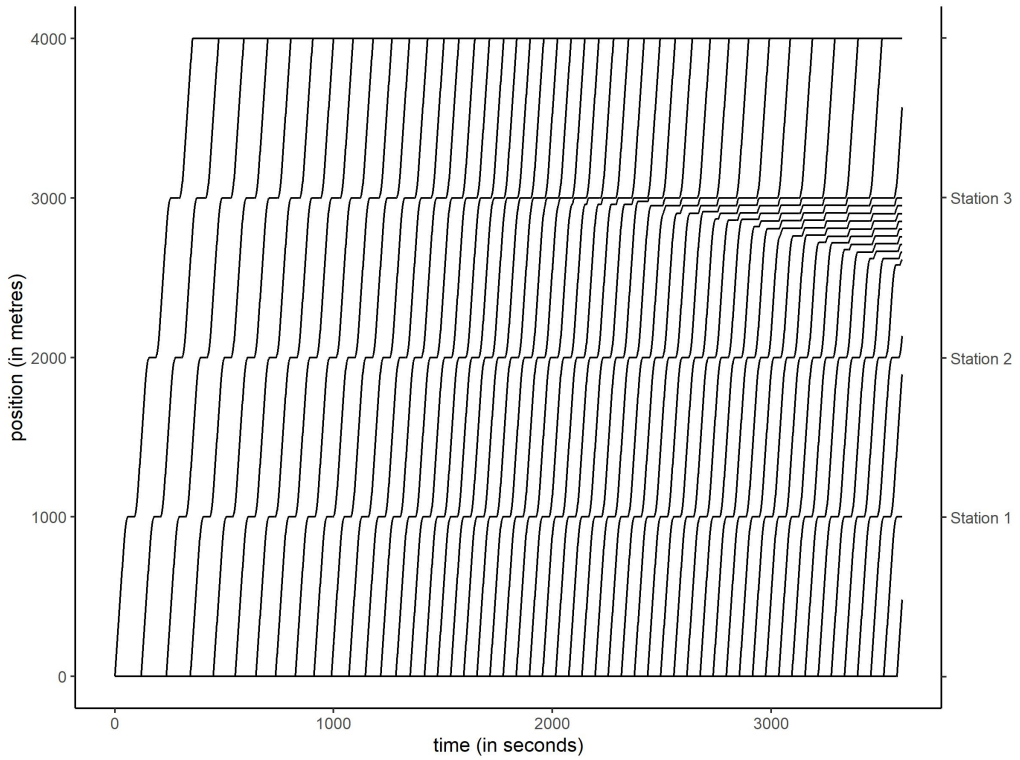
Figure 6.1b shows the effect of increasing passenger movements per train on train flow at Station 3, where train flow is obtained by taking the inverse of time headway of trains arriving at Station 3. The figure illustrates that with increasing passenger movements, train flow first increases as a result of increase in train departure frequency. However, beyond a certain value of passenger movements, train flow decreases due to high levels of passenger-train-congestion. The maximum observed train flow is 0.0167 train per second and the corresponding optimum level of passenger movements per train is around 580 passengers per hour. Using $N_p = A_p H$, the optimum passenger movement rate turns out to be 9.65 passengers per second.

We now consider two passenger inflow control scenarios: (i) when the passenger movement rate is restricted at 9.75 passengers per second (that is, slightly higher than the optimum rate), and (ii) when the passenger movement rate is restricted at the estimated optimal value of 9.65 passengers per second using station-level control strategies. Figure 6.2 shows the time-space diagram for both scenarios. Although we observe lower train-congestion as compared to the *no control* scenario (compare Figures 6.1a and 6.2a), train queuing and decrease in system throughput as compared to optimal are still substantial. The number of trains that pass through the system in the first scenario is 43 per hour. Interestingly, the queues are entirely eliminated, and the hourly system throughput increases to 47 trains in the optimal scenario (see Figure 6.2b).

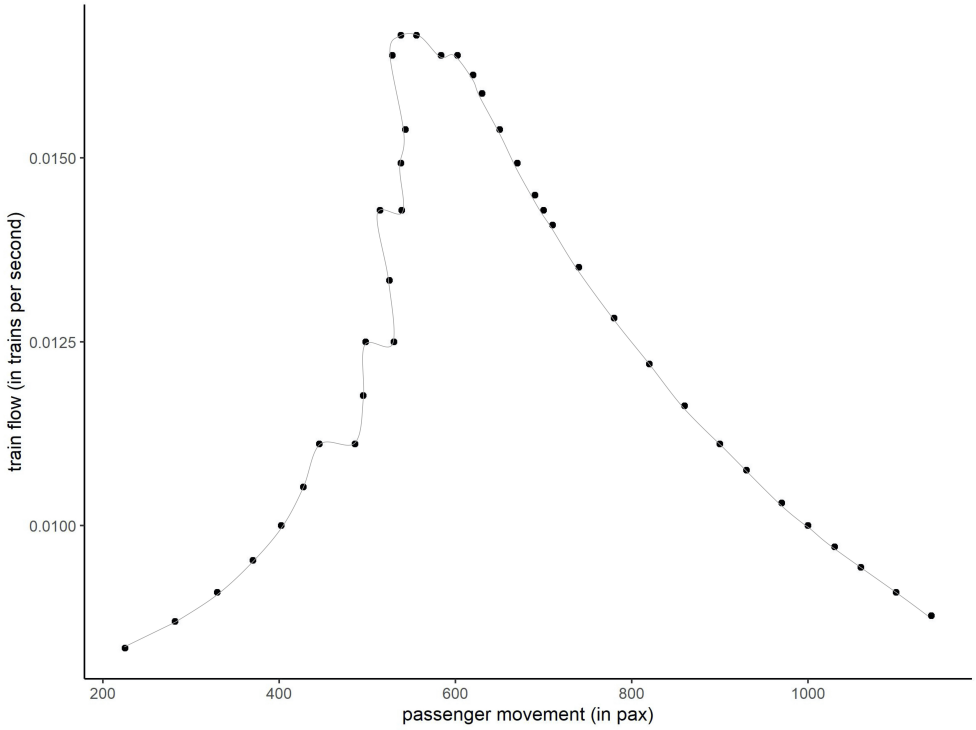
Furthermore, we compare the station-level passenger inflow control and headway-based control strategies. The headway-based strategies have been recommended in the literature (for instance, see Seo et al. 2017, Keiji et al. 2015), which enable operators to avoid train-congestion by moderating train movements per train. In our headway-based strategy, we control train movements by increasing the interval D between successive trains entering into the metro system. In this controlled scenario, we set the minimum value of D as 90 seconds, as opposed to 60 seconds in the *no control* scenario. Moreover, to avoid queuing, we allow trains to be held longer at stations upstream of the bottleneck (that is, at stations 1 and 2) by increasing their dwell time from 30 seconds to 60 seconds. Figure 6.3 shows the time-space diagram for this headway-based control scenario. We note that train-congestion is significantly lower in the headway-based control as compared to the *no control* scenario. However, the throughput of the system decreases from 39 per hour in *no control* to 27 per hour in the headway-based control scenario.

We also consider a combination of station-level passenger inflow control and headway-based strategies. We restrict the maximum passenger movement rate at 9.75 passengers per second and we also set the minimum value of D as 70 seconds, as opposed to 60 seconds in the *no control* scenario. Figure 6.4 shows the time-space diagram for this combined control scenario. The figure illustrates that queuing of trains is completely eliminated under this control strategy and the system throughput increases from 27 per hour in the headway-based control only scenario to 38 trains per hour in the combined control scenario.

The simulation study thus illustrates that ensuring the optimum passenger movement per train at bottleneck stations using station-level controls can be more effective than solely headway-based strategies in reducing congestion-related delays and improving service reliability. This simple exercise also corroborates the findings from some recent studies on bus transit operations, which compare control strategies that combine limiting the number of boarding passengers at stops during peak operations and holding of buses at

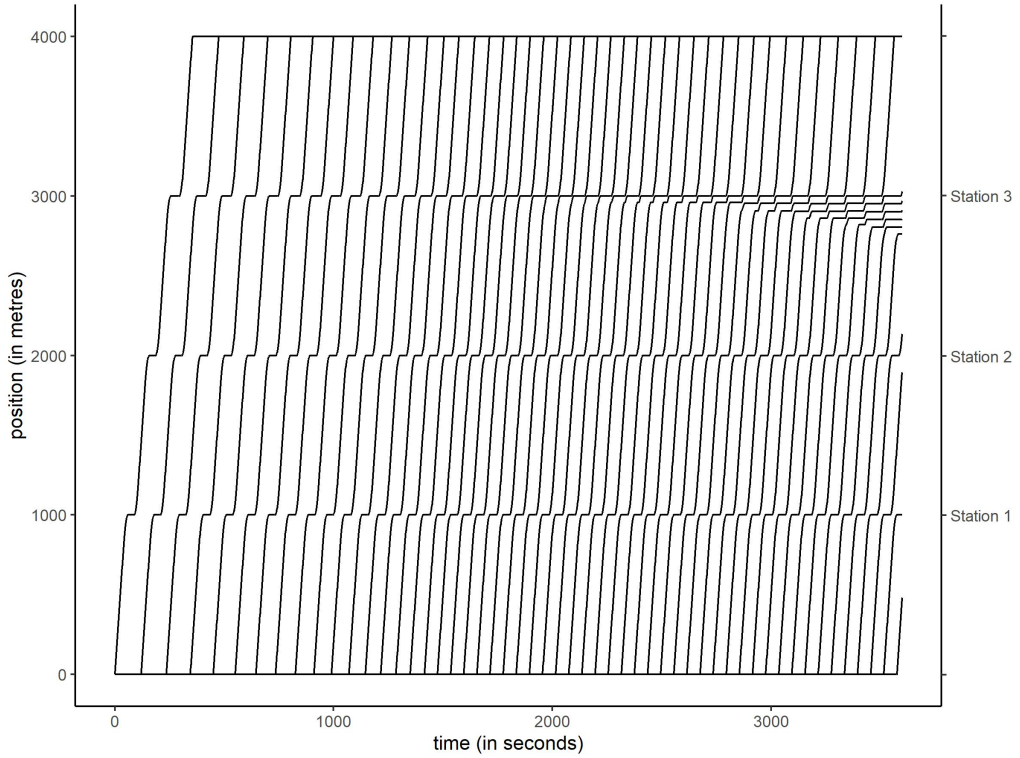


(a) The time-space diagram representing train trajectories.

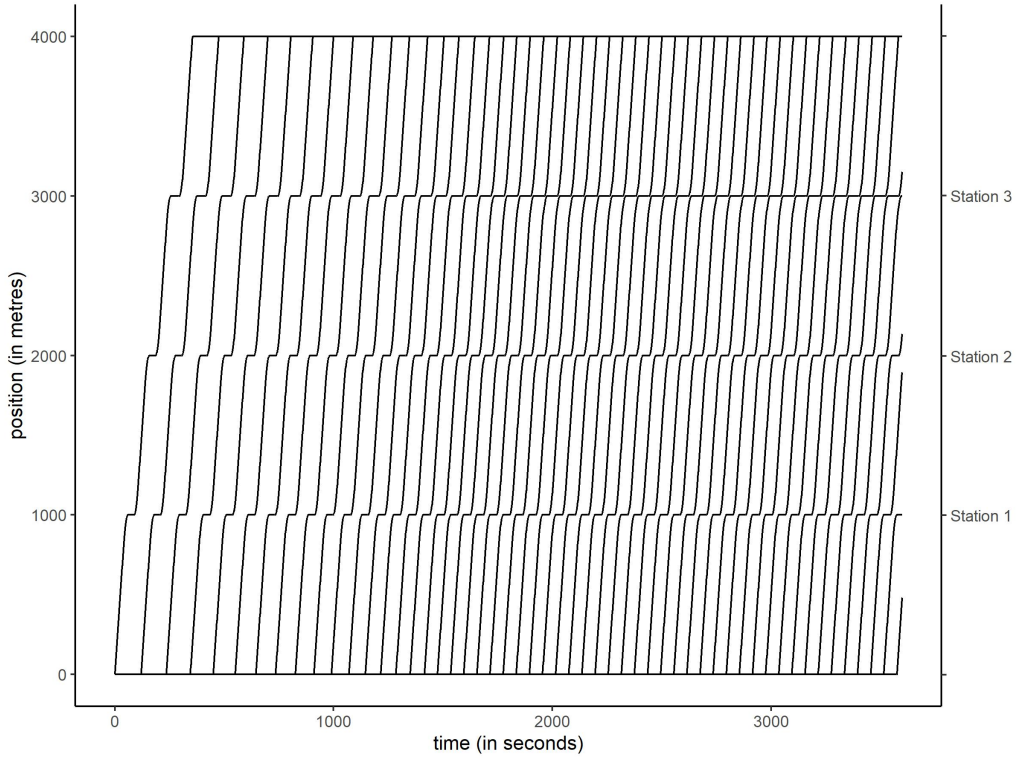


(b) The train flow versus passenger movement per train diagram.

Figure 6.1: Train operations under no control scenario.



(a) When maximum passenger movement rate at the bottleneck station is restricted at 9.75 passengers per second.



(b) When maximum passenger movement rate at the bottleneck station is restricted at the optimal level of 9.65 passengers per second.

Figure 6.2: The time-space diagrams representing train operations under passenger inflow control scenarios.

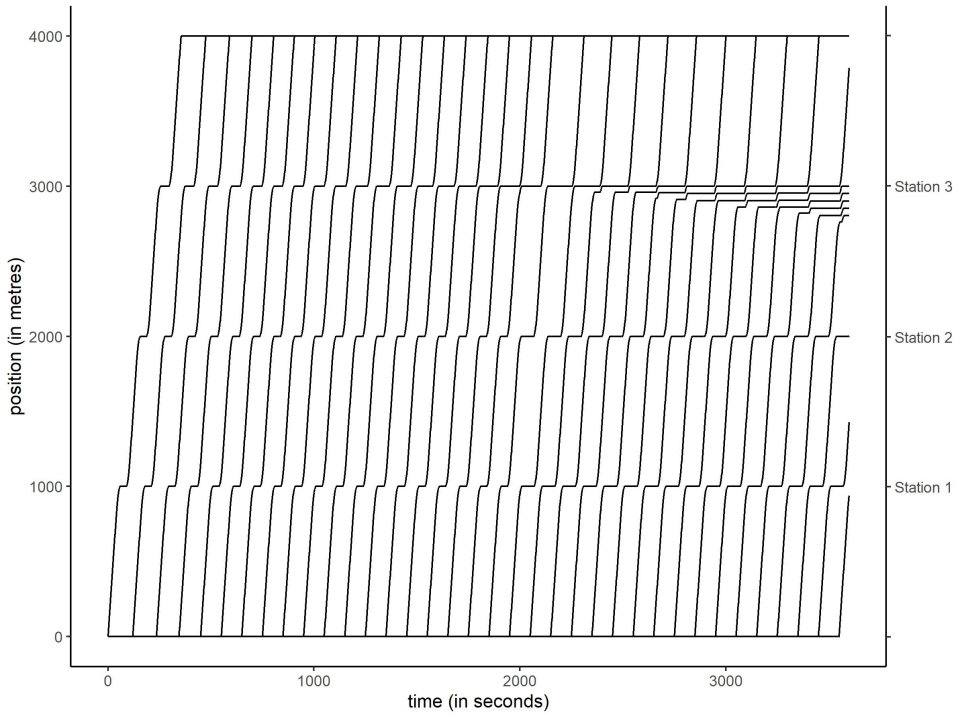


Figure 6.3: The time-space diagram representing train operations under a headway-based control strategy.

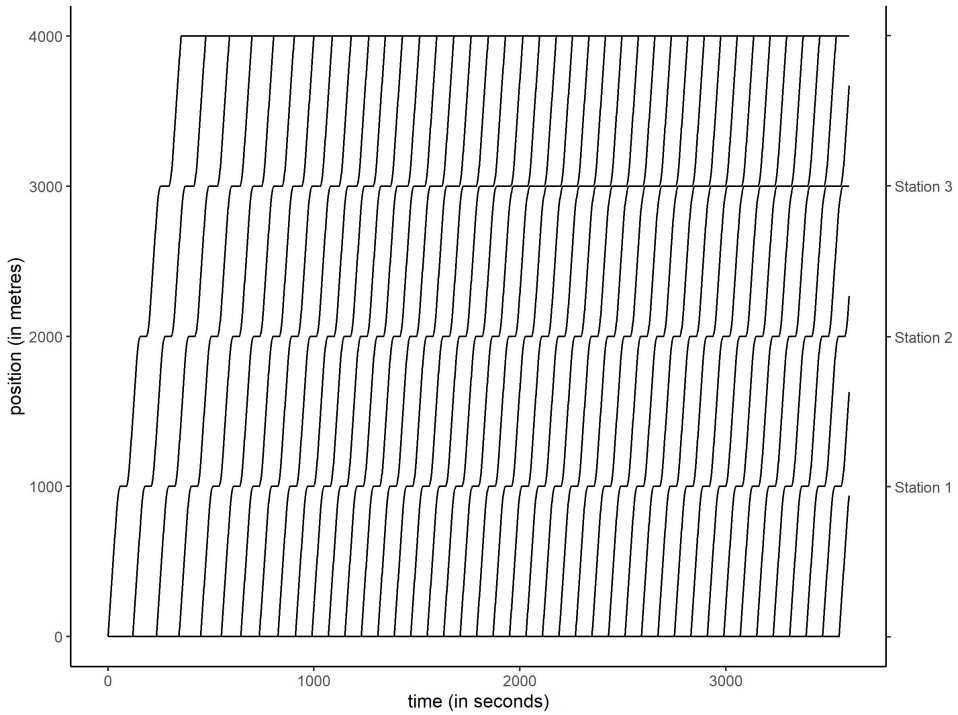


Figure 6.4: The time-space diagram representing train operations under a combination of passenger inflow control and headway-based control strategies.

control stations (headway-control) with those that involve holding of buses only (Delgado et al. 2009, 2012). These studies suggest that the hybrid strategy outperforms the headway-based strategy in improving service reliability by avoiding bus bunching. In the rest of this chapter, we show how the bottleneck stations can be identified and how the optimum passenger movements at these stations can be empirically estimated using automated fare collection and train movement data.

6.3 Model and Data

As discussed in the Introduction section (Section 6.1), we aim to estimate a *causal relationship* between passenger movements per train and train flows at each station. In other words, the objective of the empirical study is to estimate an equivalent of Figure 6.1(b) (presented in the simulation exercise in Section 6.2) for each station using data of Hong Kong MTR. Through these station-level diagrams, we aim to examine whether a unique and optimal passenger volume exists at stations, above which passenger movements negatively affect the train arrival rate.

This section is divided into two subsections. In the first subsection, we discuss the model specification, describe the Bayesian NPIV method in the context of this study and highlight the estimation practicalities. In the second subsection, we describe the data and the relevant variables.

6.3.1 Methodology

Model Specification

We consider that the average train flow q_{it}^s (that is, inverse of headway) at a station s in the ten-minute interval i on a particular day t is a function of the average number of boarding and alighting movements per train n_{it}^s occurring at the station in that interval:

$$q_{it}^{s} = S(n_{it}^{s}) + \omega_{it} + \xi_{it} \tag{6.3}$$

where ω_{it} represents the unobserved properties of the station-specific operations, such as existing control measures adopted by station staff, ξ_{it} is a idiosyncratic error term representing all random shocks to the dependent variable, and the unknown functional relationship of n_{it}^{s} with q_{it}^{s} is denoted by $S(.)$. Based on Figure 6.1(b), we can expect $S(.)$ to be a step function with different steps representing various regimes of the planned train frequencies. Adopting a parametric specification such as a quadratic function may be too restrictive to capture such non-linearities in the estimated relationship. Therefore, a non-parametric specification of $S(.)$ should be considered to obviate the need for defining its functional form a priori.

In addition, it is worth acknowledging that operators often adopt control measures to restrict passenger movements during peak hours so that the planned dwell times and headway of trains can be maintained. For instance, Transport for London often closes entrances/exits at various stations during peak hours to regulate passenger demand (TfL 2018). Considering that ω_{it} represents these control measures, we expect a negative correlation between ω_{it} and n_{it}^{s} and a positive correlation between ω_{it} and q_{it}^{s}. The unavailability of a measure for ω_{it} may lead to a confounding bias in the estimates of $S(.)$, commonly known as omitted variable bias in the econometrics literature (see Cameron & Trivedi 2005, for details). In particular, in the absence of a suitable measure or proxy for ω_{it}, an ordinary least squares estimation may underestimate $S(.)$ if $S(.)$ is a linear function (Cameron & Trivedi 2005). Therefore, we adopt a nonparametric instrumental variable (NPIV) regression, which not only enables non-parametric specification of $S(.)$ but also addresses any potential confounding biases. As discussed in the previous chapters, classical (frequentist) NPIV regression approaches are popular in theoretical econometrics (such as, Newey & Powell 2003, Horowitz 2011, Newey 2013, Chetverikov & Wilhelm 2017), but they are challenging to apply in practice due to two main reasons. First, tuning

parameters to monitor the flexibility of $S(.)$ are often required to be specified by the analyst. Second, standard errors are generally computed using bootstrap, making these methods computationally prohibitive for large datasets. Therefore, we adopt a scalable *Bayesian* NPIV approach, proposed by Wiesenfarth et al. (2014), that can produce a consistent estimate of non-parametric $S(.)$, even if the analyst does not observe ω_{it}. This Bayesian method addresses both challenges of the frequentist estimation because it *learns* tuning parameters related to $S(.)$ during estimation and uncertainty in parameters estimates is inherently captured by credible intervals (analogous to classical confidence intervals). In addition, it also enables nonparametric specification of the unobserved error component ξ_{it}, precluding the need for making additional assumptions. In Section 4.4.4 of Chapter 4, we benchmark the performance of the Bayes NPIV estimator against state-of-the-art estimators in a Monte Carlo study and illustrate its ability to adjust for endogeneity bias and recover complex functional forms of $S(.)$.

Bayesian Nonparametric Instrumental Variable Regression

We revisit the Bayesian NPIV approach (Wiesenfarth et al. 2014) for a model with a single endogenous covariate, that is,

$$q = S(n) + \epsilon_2, \quad n = h(z) + \epsilon_1 \tag{6.4}$$

Note that ω and ξ are encapsulated in ϵ_2, and z is an instrument for the endogenous regressor n. The relationship between n and z is represented by an unknown functional form $h(.)$ and ϵ_2 is an idiosyncratic random error term. For the notational simplicity, we drop time-day subscripts. Bayesian NPIV is a control function approach, and assumes the following standard identification restrictions:

$$E(\epsilon_1|z) = 0 \quad \text{and} \quad E(\epsilon_2|\epsilon_1, z) = E(\epsilon_2|\epsilon_1), \tag{6.5}$$

which yields

$$\begin{aligned} E(q|n,z) &= S(n) + E(\epsilon_2|\epsilon_1, z) = S(n) + E(\epsilon_2|\epsilon_1) \\ &= S(n) + \nu(\epsilon_1), \end{aligned} \tag{6.6}$$

where $\nu(\epsilon_1)$ is a function of the unobserved error term ϵ_1. This function is known as the control function.

To satisfy the identification restrictions presented in equation 6.5, we need an instrumental variable (IV) z. The IV should be (i) exogenous, that is, uncorrelated with ω, ξ, and ϵ_2; (ii) relevant, that is, strongly correlated with the endogenous covariate n. Due to the absence of suitable external instruments, we use the lagged level of the endogenous covariate (average passenger movements per train) as an instrument, that is, for average passenger movements observed in the ten-minute interval i on day t, we consider the observation on the covariate from the same interval i from the previous workday $t-1$ as its instrument. We argue that the average passenger movements n^s_{it} in the ten-minute interval i on day t is highly correlated with the average passenger movements $n^s_{i,t-1}$ in the same ten-minute interval i on the previous day $t-1$. This correlation follows from the influence of time-of-the-day on passenger demand. However, these lagged passenger movements $n^s_{i,t-1}$ are exogenous because they do not directly determine the response variable q^s_{it} in equation 6.3. To justify the relevance of the considered instrument, we present the estimated $h(.)$ in equation 6.4 in the Results and Discussion Section (Section 6.4.1).

Further details of the Bayesian NPIV estimator along with the estimation practicalities are discussed in Section 4.4.3 of Chapter 4.

6.3.2 Data and Relevant Variables

We use the automatic fare collection (AFC) or data from entry/exit gates at the stations and automatic vehicular location (AVL) or train movement datasets provided by Hong Kong MTR, the urban and suburban rail operator of Hong Kong and a member of the Community of Metros facilitated by the Transport Strategy Centre (TSC) at Imperial College London. The MTR dataset is practical for the present analysis because it is a closed system. All stations in the MTR network are fenced, and thus, the AFC data contain information about all transactions at both the origin and destination stations. The data contain a record for millions of entry/exit transactions corresponding to trips occurring in the MTR network over the period from January 1, 2019 to March 31, 2019. The AVL data recovered from the signalling system contains a precise record of departure and arrival times of trains at each station in the MTR network for the above mentioned period. We assign passengers to trains by matching automated fare collection data with train movement data using the methodology detailed in Bansal et al. (2020) (an extension to Hörcher et al. (2017)).

In this analysis, we focus on a group of stations on the Kwun Tong Line (green line) that are located in the central business district of Hong Kong. These stations are highlighted in Figure 6.5.

We analyse the trains moving in both downward and upward directions, that is, towards the Whampao and Choi Hung stations respectively. From the results of passenger-train assignment, we calculate train flow and average number of alightings and boardings per train or passenger movements per train for different consecutive ten-minute time intervals throughout a day for each station. Note that train flow in a ten-minute interval for a station is obtained by taking the mean of the inverse of time headway of trains arriving at that station within that interval.

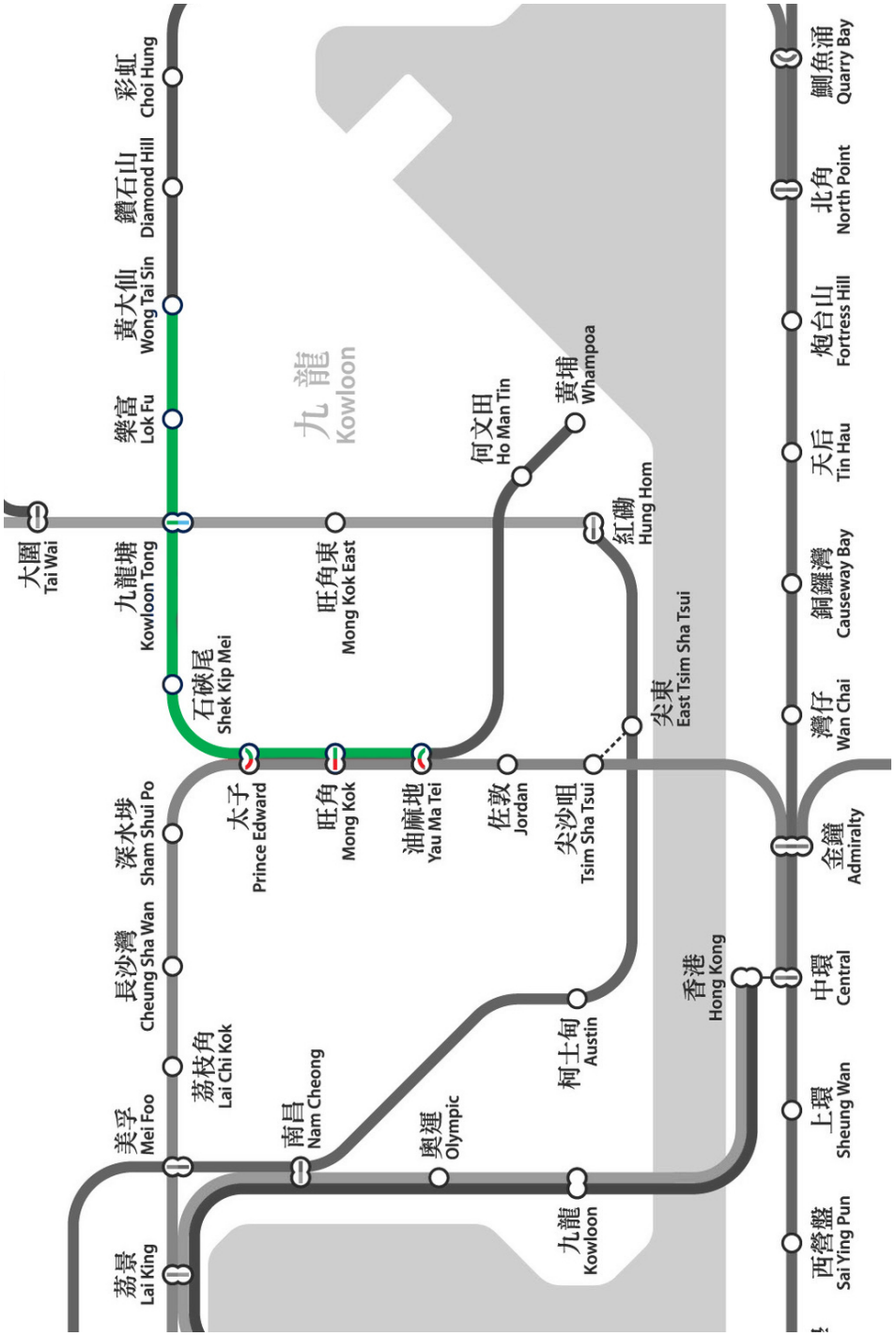


Figure 6.5: A part of the MTR network where the line that we study is highlighted in green.

Table 6.1: Summary statistics for variables used in the analysis.

Station	Direction	Variable	Obs.	Min	Max	Mean	Std.Dev
Wong Tai	downward	train flow (tr/10min)	8932	0.35	8.58	3.23	1.00
		passenger movements	8932	8.00	653.00	224.50	87.62
	upward	train flow (tr/10min)	8943	0.33	9.02	3.40	1.06
		passenger movements	8943	4.00	506.0	210.00	67.36
Lok Fu	downward	train flow (tr/10min)	9007	0.35	7.54	3.33	1.03
		passenger movements	9007	3.00	381.00	105.28	41.24
	upward	train flow (tr/10min)	8977	0.32	9.20	3.38	1.07
		passenger movements	8977	3.00	283.50	106.60	40.79
Kowloon Tong	downward	train flow (tr/10min)	9018	0.33	7.74	3.32	1.02
		passenger movements	9018	19.00	2244.00	551.70	222.28
	upward	train flow (tr/10min)	9039	0.38	8.75	3.34	1.05
		passenger movements	9039	10.00	1574.50	519.50	191.67
Shek Kip Mei	downward	train flow (tr/10min)	8946	0.33	8.70	3.34	1.01
		passenger movements	8946	3.00	293.00	88.67	31.99
	upward	train flow (tr/10min)	9008	0.34	9.68	3.35	1.04
		passenger movements	9008	5.33	406.33	86.10	33.19
Prince Edward	downward	train flow (tr/10min)	8959	0.33	8.70	3.34	1.04
		passenger movements	8959	6.00	1523.00	451.50	175.94
	upward	train flow (tr/10min)	8969	0.36	9.49	3.32	1.02
		passenger movements	8969	3.00	1703.50	416.00	158.57
Mong Kok	downward	train flow (tr/10min)	8946	0.34	8.19	3.35	1.01
		passenger movements	8946	10.00	1697.70	544.60	257.76
	upward	train flow (tr/10min)	8970	0.37	10.35	3.02	1.00
		passenger movements	8970	12.00	1595.00	293.40	146.84
Yau Ma Tei	downward	train flow (tr/10min)	9007	0.34	8.25	3.34	1.02
		passenger movements	9007	2.00	398.00	115.53	46.66
	upward	train flow (tr/10min)	8955	0.33	10.79	3.29	1.00
		passenger movements	8955	29.00	2798.00	524.70	274.03

*Obs.: Number of observations, Std. Dev.: Standard Deviation,tr/10min: trains per ten minutes

**Passenger movements represent average no. of boardings and alightings per train

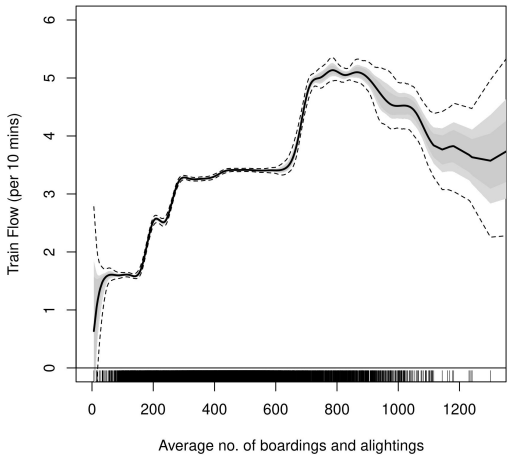
Table 6.1 presents the summary statistics for ten-minute train flows and average number of boardings and alightings at each station. We note that the four interchange stations – Kowloon Tong, Prince Edward, Mong Kok and Yau Ma Tei – are associated with higher level of passenger boardings and alightings in either of the two or both directions of train flow as compared to other stations. We provide the observed scatter plots of train flow versus passenger movements per train for the considered stations in Figures D.2 and D.3 in Appendix D.2.

6.4 Results and Discussion

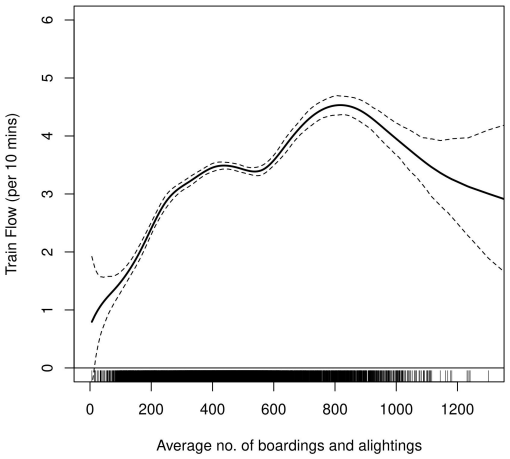
This section is divided into four subsections. In the first subsection, we compare results from our IV-based estimator with those from a non-IV estimator. The non-IV estimator is a counterpart of the Bayesian NPIV, which does not address confounding bias (that is, $z = n; \epsilon_1 = 0; h$: identity function in Equation 6.4). In the next subsection, we discuss the estimated kernel error distributions to illustrate the importance of the non-parametric DPM specification. We discuss the relevance of our instruments in the penultimate subsection. We conclude this section by describing Bayesian NPIV results in detail and discussing how we identify the optimal passenger movement at the bottleneck stations.

6.4.1 Comparison of IV-based and non-IV-based estimators

We compare the estimates of $S(.)$ in equation 6.4 (second-stage), which we obtain using the Bayesian NPIV, and its non-IV-based counterpart, for Prince Edward Station for train flows in both downward and upward direction as shown in Figures 6.6 and 6.7. Both figures suggest that IV-based estimate of $S(.)$ is as efficient as its non-IV counterpart, that is, both have similar and tight credible bands for the domain of passenger movements where we have sufficient number of observations (note that the density of the tick marks on the X-axis represents the number of observations).

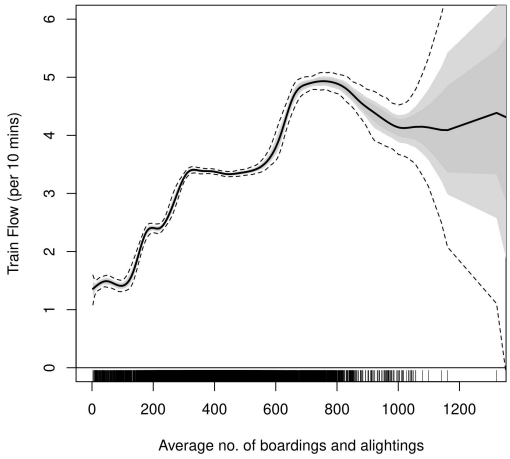


(a) With instrumental variables

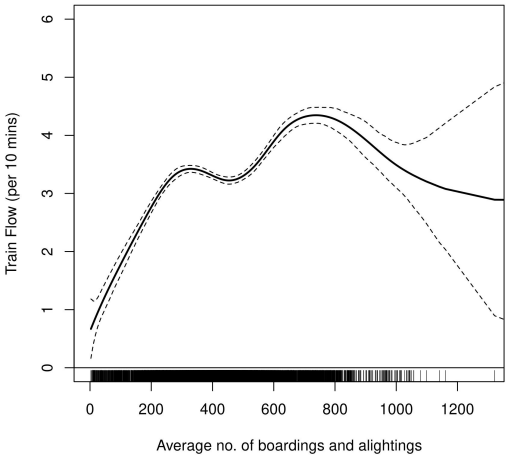


(b) Without instrumental variables

Figure 6.6: Train Flow (per 10 minutes) in downward direction versus Average number of boardings and alightings (in 10 minutes) at Prince Edward Station.



(a) With instrumental variables



(b) Without instrumental variables

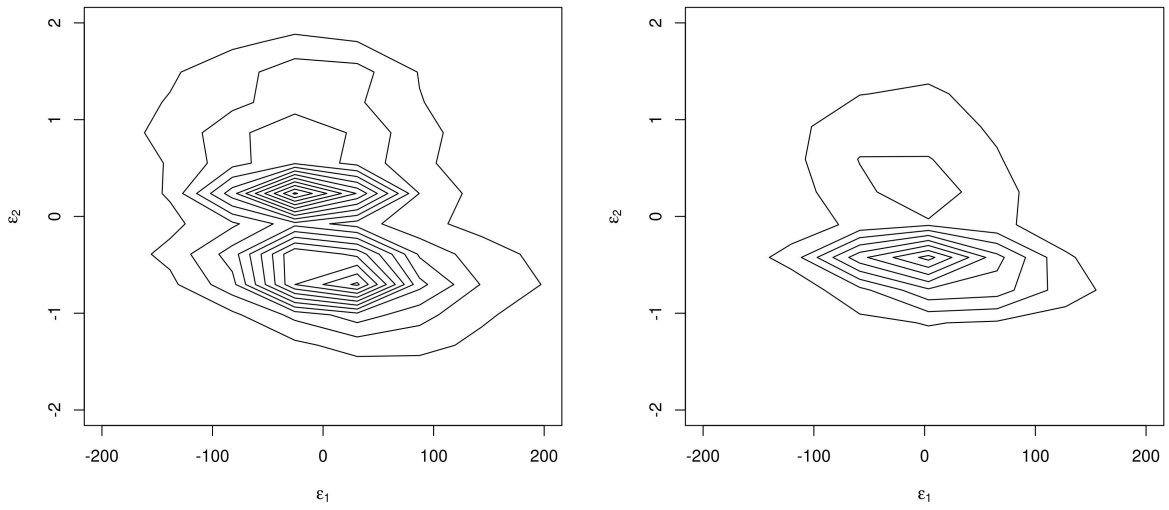
Figure 6.7: Train Flow (per 10 minutes) in upward direction vs Average number of boardings and alightings (in 10 minutes) at Prince Edward Station.

However, the maximum train flow is slightly underestimated by the non-IV estimator in both figures. As discussed in Section 6.3.1, such bias is expected in the non-IV-based estimate due to the absence of suitable control for unobserved characteristics of metro operations. Moreover, the multi-step function estimated from the IV-based method is

more plausible as it could detect multiple regimes of the planned train frequencies, instead of binary regimes illustrated by non-IV estimates (see Travel China Guide 2019, for time-of-day frequency of the Kwun Tong Line). The results indicate that the endogeneity bias is less severe in our case, however, it may be more pronounced in other similar empirical studies depending upon their data generating processes. The advantages of adopting NPIV would be even more apparent in presence of large endogeneity biases.

6.4.2 Distribution of Errors

Figure 6.8 shows the contour plot of the joint distribution of errors from the second stage (ϵ_2) and the first stage (ϵ_1) for both directions of train flows at Prince Edward Station. From these figures, we observe the joint error distribution is bi-modal.



(a) Prince Edward Station (downward). (b) Prince Edward Station (upward).

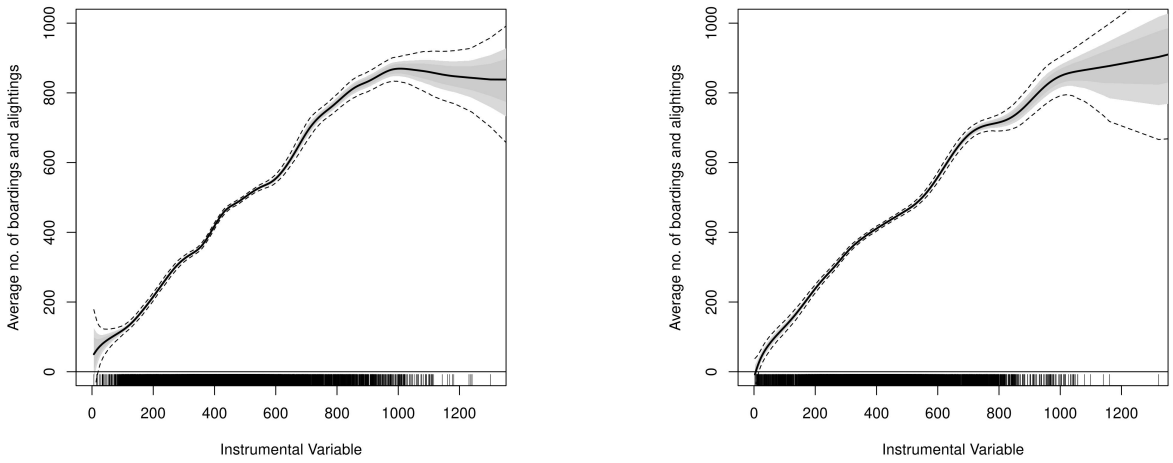
Figure 6.8: Distribution of errors.

The results suggest that the estimates of $S(.)$ from traditional econometric methods could have poor finite sample properties because they generally assume uni-modal symmetric and thin-tailed Gaussian error distributions. The adopted Bayesian NPIV method addresses all these potential challenges by allowing for a flexible distribution of errors,

instead of assuming a restrictive parametric error distribution.

6.4.3 Relevance of Instruments

Figure 6.9 illustrates the results (that is, the estimated $h(.)$) from regression of the endogenous covariate over the chosen instrument for the two directions of train flows at Prince Edward Station.



(a) Prince Edward Station (downward).

(b) Prince Edward Station (upward).

Figure 6.9: Strength of instruments used in this analysis.

From these figures, we notice a strong correlation between the instrument and the endogenous covariate. These figures provide supporting evidence that the selected instruments satisfy the relevance condition. For other stations, we observe similar patterns of correlation between the instrument (that is, passenger movements per train in a given time-of-the-day on the previous workday) and the endogenous covariate, but we omit them here for brevity. Full results are attached in Appendix D.5.

6.4.4 Bottlenecks and station-level optimal passenger movements

Figures 6.10 and 6.11 show the estimated train flow versus passenger movement per train curves (that is, the estimated $S(.)$ in the second stage) for all stations that are

highlighted in Figure 6.5 for downward and upward directions respectively. We observe that all the estimated curves are nearly concave, but the associated credible bands in the backward bending region are very wide for all stations, except for Prince Edward Station in the downward direction and Kowloon Tong Station, Mong Kok Station, Prince Edward Station and Yau Ma Tei Station in the upward direction (see Figures 6.10c, 6.11c, 6.11e, 6.11f, 6.11g). However, the statistical significance of the backward bending part is apparent in a short range of passenger movements at these stations. Thus, these plots provide empirical evidence to support the existence of a unique and optimal passenger volume, above which passenger movements negatively affect the train arrival rate.

The stations with statistically significant backward bending act as active bottlenecks in the associated direction of train flow along Kwun Tong Line in the MTR network. In the downward direction, the optimal number of passenger boardings and alightings at Prince Edward Station is around nine hundred passengers. In the upward direction, the optimal number of passenger boardings and alightings is around nine hundred passengers at Kowloon Tong Station, around eight hundred passengers at Prince Edward Station, around seven hundred passengers at Mong Kok Station, and around eleven hundred passengers at Yau Mai Tei Station. The corresponding maximum train inflow at all the bottleneck stations is around five trains per ten minutes, that is, the estimated minimum headway between trains is around two minutes. The estimated minimum headway value in both directions is consistent with the planned minimum peak headway of 2.1 minutes (Travel China Guide 2019). Table 6.2 summarises these results. Thus, the application of the NPIV approach allows us to adjust for any confounding bias and recover the scheduled peak headway.

The large credible intervals in the backward bending part of the estimated $S(.)$ at all other stations implies that there may be only a handful of instances of the delay propagation from bottlenecks to these stations due to in-place control measures and relatively lower operating frequency of the MTR services. Thus, the estimated relationship

between train flow and passenger movements and optimal/critical passenger movements are not universal, rather they depend upon the characteristics of the metro network such as metro demand, frequency, and spacing between stations, station design among many others.

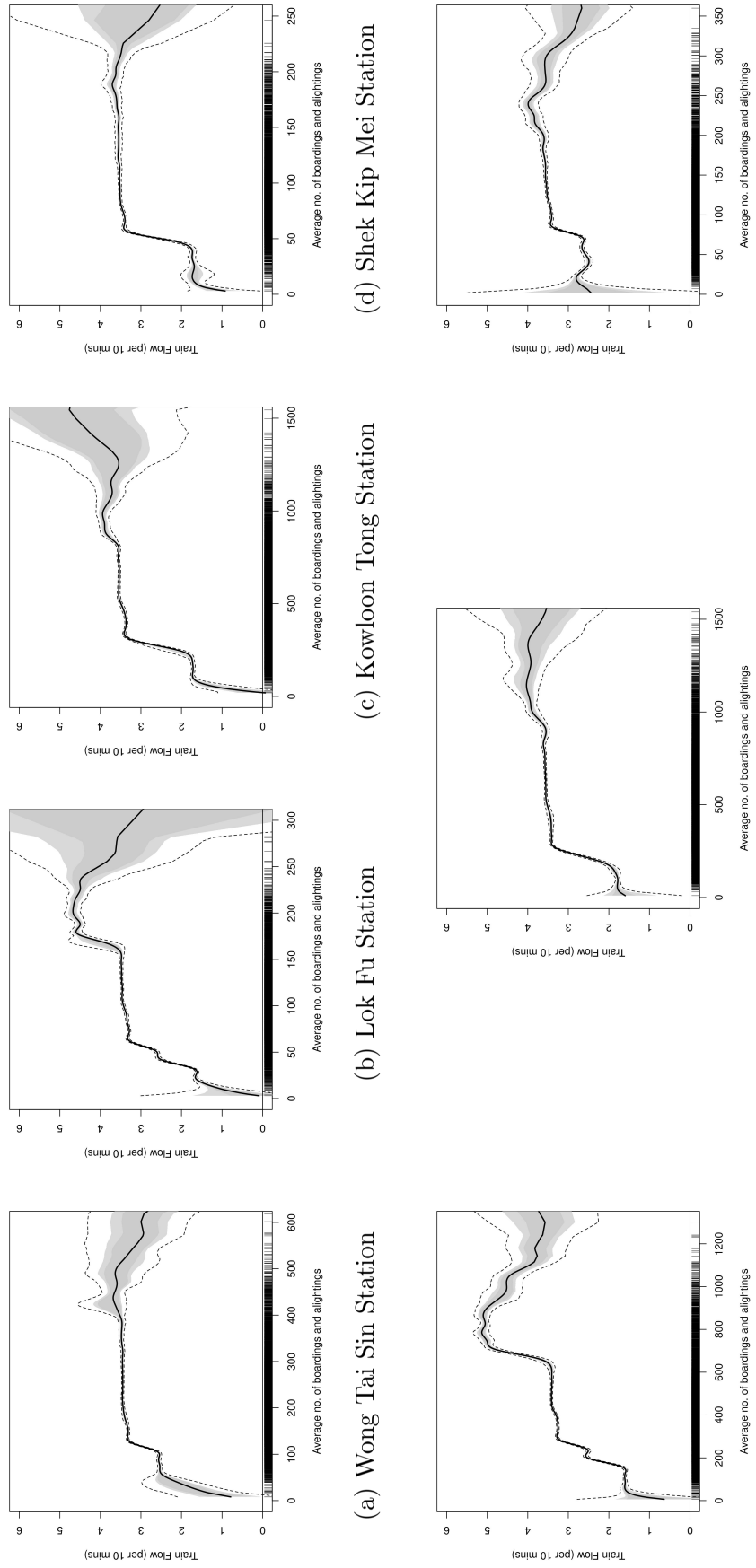
Table 6.2: Summary of results.

Identified Bottleneck Station	Direction of Flow	Optimum passenger movements	Scheduled peak headway
Prince Edward	downward	$\sim$ 900 pax per train	$\sim$ 2 minutes
Kowloon Tong	upward	$\sim$ 900 pax per train	$\sim$ 2 minutes
Mong Kok	upward	$\sim$ 700 pax per train	$\sim$ 2 minutes
Prince Edward	upward	$\sim$ 800 pax per train	$\sim$ 2 minutes
Yau Ma Tei	upward	$\sim$ 1100 pax per train	$\sim$ 2 minutes

To show the relevance of the estimated optimum passenger movement from a conceptual perspective and illustrate its importance in devising control strategies, we briefly discuss the analogous concept of *capacity* and *critical density* from traffic flow theory. Understanding traffic capacity and the corresponding critical density at the link- or network-level has been the main focus in the modelling of traffic flow (Srivastava & Geroliminis 2013, Siebel et al. 2009, Laval & Daganzo 2006, Loder et al. 2019, Geroliminis & Daganzo 2008, Daganzo & Geroliminis 2008). This is because the capacity may be insufficient for the peak-hour demand and the system may transition from a free-flow state to a congested state. With the increase in the number of vehicles in the system, travel production decreases in the congested state but increases in the free-flow state. Traffic control strategies like ramp metering and congestion pricing aim to regulate the demand to maximise the link- or network-level travel production (Papageorgiou et al. 2003, Small & Verhoef 2007). The existence of a well-defined fundamental diagram between traffic flow and traffic density and

the corresponding estimates of traffic capacity and critical density over which productivity of the system falls, have been crucial inputs in the development of such traffic control strategies.

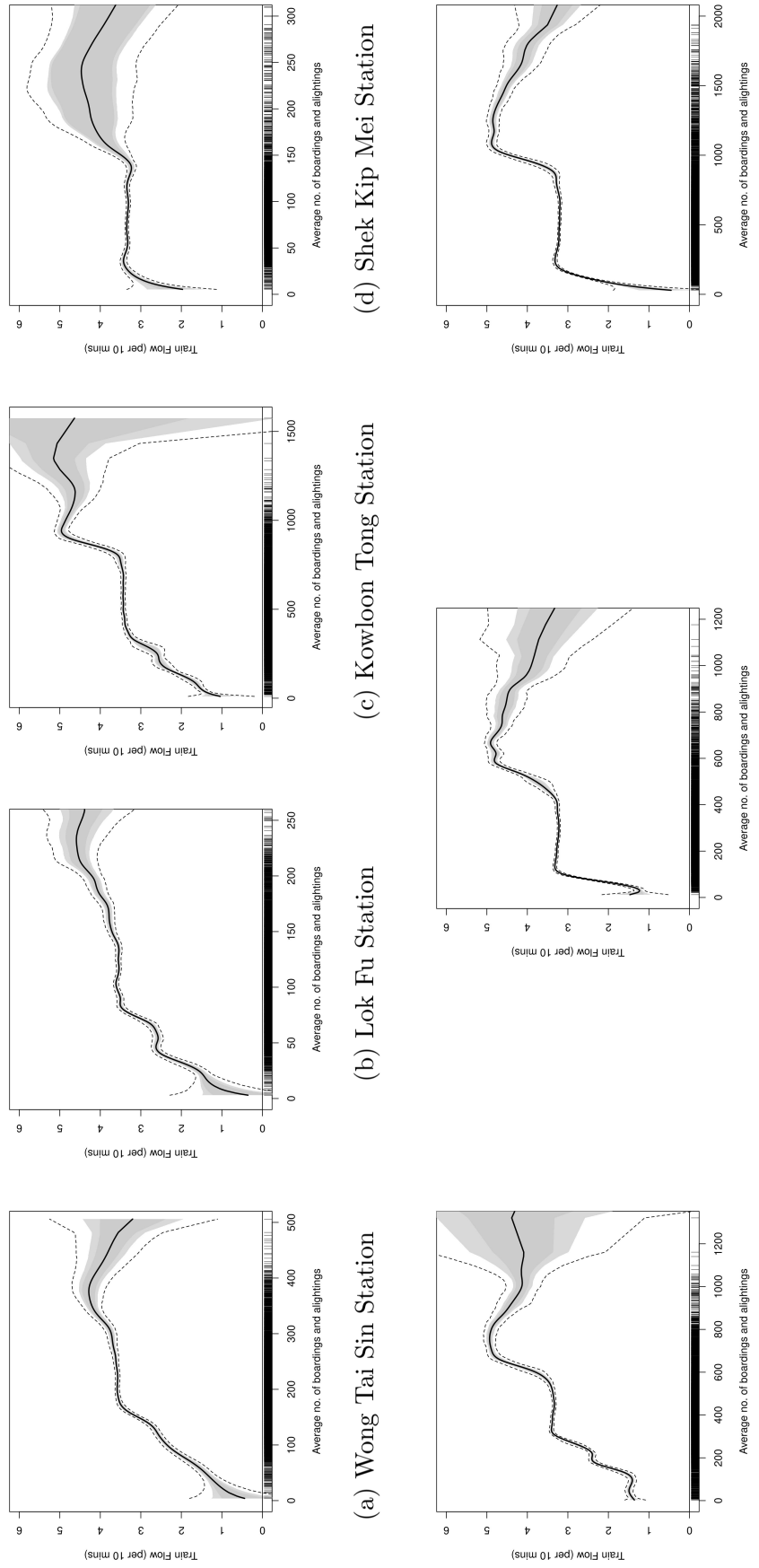
Similar to the fundamental diagram in traffic flow theory, we illustrate the existence of a well-defined relationship between passenger movements and train flow at a station in metros. Furthermore, we find that there also exists a unique critical level of passenger movements, over which the throughput of the metro line decreases. These estimates could be crucial inputs in the design of passenger inflow control strategies that are similar to vehicular control strategies in road traffic.



(a) Wong Tai Sin Station (b) Lok Fu Station (c) Kowloon Tong Station (d) Shek Kip Mei Station

(e) Prince Edward Station (f) Mong Kok Station (g) Yau Ma Tei station

Figure 6.10: Non-parametric Instrumental Variables based estimation results for train movements in the downward direction along the Kwun Tong Line.



(a) Wong Tai Sin Station (b) Lok Fu Station (c) Kowloon Tong Station (d) Shek Kip Mei Station

(e) Prince Edward Station (f) Mong Kok Station (g) Yau Ma Tei station

Figure 6.11: Non-parametric Instrumental Variables based estimation results for train movements in the upward direction along the Kwun Tong Line.

6.5 Conclusions and Relevance

This chapter provides the first station-level analysis of congestion in a metro network by estimating a causal relationship between passenger movement per train and train flow. We use automated fare collection and train movement data from Hong Kong MTR and adopt a data-driven Bayesian non-parametric instrumental variable method to address potential confounding biases in the estimated relationship. This analysis could help metro operators to identify bottleneck stations in the network. Furthermore, optimum passenger movements and corresponding train frequencies are also obtained as a by-product of the estimation.

The short-term prediction of subway passenger demand has received significant attention in recent years (Ding et al. 2016, Ma et al. 2018). This study enhances the value of short-term demand prediction by estimating its causal impact on train frequencies. Understanding the dynamics of passenger movements and train frequencies, along with estimates of optimum passenger movement, can help in designing strategies to control passenger movements and minimise delays. Such control strategies may involve i) adopting platform management practices such as reducing escalator capacity, ii) deployment of staff resources to regulate the entry of passengers into bottleneck stations, and iii) pricing policies. Another strategy could be *ramp metering*[2] of passengers entering stations to increase overall system throughput. Daganzo (2005, 2007) suggest such strategies in the context of vehicular traffic control in urban networks.

We note that metro operators around the world presently implement such strategies based on their day-to-day experience of congestion patterns at various stations. For instance, Transport for London implements different types of station control measures such as avoid train dwelling at particular stations during a specific time of the day,

[2]A ramp meter is a basic traffic light device together with a signal controller, that are used to regulate the flow of vehicular traffic entering freeways according to current traffic conditions. Ramp metering systems have proved to be successful in decreasing traffic congestion on freeways. Similar strategies are being sought after by metro operators to regulate passenger demand in future.

individual platform closures, and closures of gate lines and entrances (TfL 2018). The findings of this study could assist metro operators in improving these control strategies in a data-driven manner.

It is worth noting that the above-discussed strategies rely on controlling passenger boarding movements, but the estimated relationship includes both passenger boarding and alighting movements. This disparity does not restrict the application of the empirical results in practice because metro operators can use short-term demand prediction models to forecast the number of alighting and boarding movements at any station. Subsequently, they can adopt station-level control measures to regulate the number of boarding movements such that the instantaneous sum of boarding and alighting movements remains optimum. Developing and testing such control measures using real data is an important avenue for future research. Another interesting area of future research could be to explore the potential of fundamental diagrams in the long-run to understand the level of operational service and guide improvements in the metro network.

References

Bansal, P., Hörcher, D. & Graham, D. J. (2020), 'A dynamic choice model with heterogeneous decision rules: Application in estimating the user cost of rail crowding', *arXiv preprint arXiv:2007.03682* .

Cameron, A. C. & Trivedi, P. K. (2005), *Microeconometrics: methods and applications*, Cambridge university press.

Carey, M. & Kwieciński, A. (1994), 'Stochastic approximation to the effects of headways on knock-on delays of trains', *Transportation Research Part B: Methodological* **28**(4), 251–267.

Chetverikov, D. & Wilhelm, D. (2017), 'Nonparametric instrumental variables estimation under monotonicity', *Econometrica* **85**(4), 1303–1320.

Daganzo, C. F. (1997), *Fundamentals of transportation and traffic operations*, Vol. 30, Pergamon Oxford.

Daganzo, C. F. (2005), 'Improving city mobility through gridlock control: an approach and some ideas'.

Daganzo, C. F. (2007), 'Urban gridlock: Macroscopic modeling and mitigation approaches', *Transportation Research Part B: Methodological* **41**(1), 49–62.

Daganzo, C. F. (2009), 'A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons', *Transportation Research Part B: Methodological* **43**(10), 913–921.

Daganzo, C. F. & Geroliminis, N. (2008), 'An analytical approximation for the macroscopic fundamental diagram of urban traffic', *Transportation Research Part B: Methodological* **42**(9), 771–781.

Delgado, F., Munoz, J. C. & Giesen, R. (2012), 'How much can holding and/or limiting boarding improve transit performance?', *Transportation Research Part B: Methodological* **46**(9), 1202–1217.

Delgado, F., Munoz, J. C., Giesen, R. & Cipriano, A. (2009), 'Real-time control of buses in a transit corridor based on vehicle holding and boarding limits', *Transportation Research Record* **2090**(1), 59–67.

Ding, C., Wang, D., Ma, X. & Li, H. (2016), 'Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees', *Sustainability* **8**(11), 1100.

Geroliminis, N. & Daganzo, C. F. (2008), 'Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings', *Transportation Research Part B: Methodological* **42**(9), 759–770.

Gill, D. C. (1994), 'Railway signalling system'. US Patent 5,366,183.

Guo, J., Jia, L., Qin, Y. & Zhou, H. (2015), 'Cooperative passenger inflow control in urban mass transit network with constraint on capacity of station', *Discrete Dynamics in Nature and Society* **2015**.

Hong Kong MTR (2019), 'Investing for the future'.
URL: *https://tinyurl.com/y459kjnx*

Horowitz, J. L. (2011), 'Applied nonparametric instrumental variables estimation', *Econometrica* **79**(2), 347–394.

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

Independent (2017), 'Tube passengers wasted 400,000 hours in 2016 because of overcrowding delays', *Independent, UK* .
URL: *http://tiny.cc/0mulsz*

Jiang, Z., Fan, W., Liu, W., Zhu, B. & Gu, J. (2018), 'Reinforcement learning approach for coordinated passenger inflow control of urban rail transit in peak hours', *Transportation Research Part C: Emerging Technologies* **88**, 1–16.

Keiji, K., Naohiko, H. & Shigeru, M. (2015), 'Simulation analysis of train operation to recover knock-on delay under high-frequency intervals', *Case Studies on Transport Policy* **3**(1), 92–98.

Laval, J. A. & Daganzo, C. F. (2006), 'Lane-changing in traffic streams', *Transportation Research Part B: Methodological* **40**(3), 251–264.

Li, K. P., Gao, Z. Y. & Ning, B. (2005), 'Cellular automaton model for railway traffic', *Journal of Computational Physics* **209**(1), 179–192.

Loder, A., Ambühl, L., Menendez, M. & Axhausen, K. W. (2019), 'Understanding traffic capacity of urban networks', *Scientific reports* **9**(1), 1–10.

London Assembly (2019), 'Delays on the london underground caused by overcrowding', *Questions to the Mayor, Mayor of London* .
URL: *http://tiny.cc/4mulsz*

Ma, X., Zhang, J., Du, B., Ding, C. & Sun, L. (2018), 'Parallel architecture of convolutional bi-directional lstm neural networks for network-wide metro ridership prediction', *IEEE Transactions on Intelligent Transportation Systems* **20**(6), 2278–2288.

Meng, Q. & Weng, J. (2011), 'An improved cellular automata model for heterogeneous work zone traffic', *Transportation research part C: emerging technologies* **19**(6), 1263–1275.

Nagel, K. & Schreckenberg, M. (1992), 'A cellular automaton model for freeway traffic', *Journal de physique I* **2**(12), 2221–2229.

Newey, W. K. (2013), 'Nonparametric instrumental variables estimation', *American Economic Review* **103**(3), 550–556.

Newey, W. K. & Powell, J. L. (2003), 'Instrumental variables estimation of nonparametric models', *Econometrica* **71**(5), 1565–1578.

Ning, B., Xun, J., Gao, S. & Zhang, L. (2014), 'An integrated control model for headway regulation and energy saving in urban rail transit', *IEEE Transactions on Intelligent Transportation Systems* **16**(3), 1469–1478.

Papageorgiou, M., Diakaki, C., Dinopoulou, V., Kotsialos, A. & Wang, Y. (2003), 'Review of road traffic control strategies', *Proceedings of the IEEE* **91**(12), 2043–2067.

Seo, T., Wada, K. & Fukuda, D. (2017), A macroscopic and dynamic model of urban rail transit with delay and congestion, *in* '96th Annual Meeting of the Transportation Research Board'.

Shi, J., Yang, L., Yang, J. & Gao, Z. (2018), 'Service-oriented train timetabling with collaborative passenger flow control on an oversaturated metro line: An integer linear optimization approach', *Transportation Research Part B: Methodological* **110**, 26–59.

Siebel, F., Mauser, W., Moutari, S. & Rascle, M. (2009), 'Balanced vehicular traffic at a bottleneck', *Mathematical and Computer Modelling* **49**(3-4), 689–702.

Small, K. A. & Verhoef, E. T. (2007), *The economics of urban transportation*, Routledge, New York.

Spyropoulou, I. (2007), 'Modelling a signal controlled traffic stream using cellular automata', *Transportation Research Part C: Emerging Technologies* **15**(3), 175–190.

Srivastava, A. & Geroliminis, N. (2013), 'Empirical observations of capacity drop in freeway merges with ramp control and integration in a first-order model', *Transportation Research Part C: Emerging Technologies* **30**, 161–177.

Takeuchi, H., Goodman, C. & Sone, S. (2003), 'Moving block signalling dynamics: performance measures and re-starting queued electric trains', *IEE Proceedings-electric power applications* **150**(4), 483–492.

TfL (2018), 'Transport for london customer service and operational performance report', *Transport for London* .
URL: *http://tiny.cc/8ywlsz*

Tirachini, A., Hensher, D. A. & Rose, J. M. (2013), 'Crowding in public transport systems: effects on users, operation and implications for the estimation of demand', *Transportation research part A: policy and practice* **53**, 36–52.

Travel China Guide (2019), 'Kwun Tong Line, Hong Kong MTR'.
URL: *https://tinyurl.com/y45467z4*

Wada, K., Akamatsu, T. & Osawa, M. (2012), A control strategy to prevent propagating delays in high-frequency railway systems, *in* 'The 1st European Symposium on Quantitative Methods in Transportation Systems'.

Wang, X., Wu, J., Yang, X., Guo, X., Yin, H. & Sun, H. (2020), 'Multistation coordinated and dynamic passenger inflow control for a metro line', *IET Intelligent Transport Systems* **14**(9), 1068–1078.

Wiesenfarth, M., Hisgen, C. M., Kneib, T. & Cadarso-Suarez, C. (2014), 'Bayesian nonparametric instrumental variables regression based on penalized splines and dirichlet process mixtures', *Journal of Business and Economic Statistics* **32**(3), 468–482.

Xun, J., Ning, B., Li, K.-p. & Zhang, W.-b. (2013), 'The impact of end-to-end communication delay on railway traffic flow using cellular automata model', *Transportation Research Part C: Emerging Technologies* **35**, 127–140.

Yan, X., Cheng-Xun, C., Ming-Hua, L. & Jin-Long, L. (2012), 'Modeling and simulation for urban rail traffic problem based on cellular automata', *Communications in Theoretical Physics* **58**(6), 847.

Yinping, F., Ziyou, G. & Keping, L. (2008), 'Modeling study for tracking operation of subway trains based on cellular automata', *Journal of Transportation Systems Engineering and Information Technology* **8**(4), 89–95.

Yuan, F., Sun, H., Kang, L. & Wu, J. (2020), 'Passenger flow control strategies for urban rail transit networks', *Applied Mathematical Modelling* **82**, 168–188.

Zhang, J. & Wada, K. (2019), Fundamental diagram of urban rail transit: An empirical investigation by boston's subway data, *in* '8th Symposium of the European Association for Research in Transportation'.

Zou, Q., Yao, X., Zhao, P., Dou, F. & Yang, T. (2018), 'Managing recurrent congestion of subway network in peak hours with station inflow control', *Journal of Advanced Transportation* **2018**.

Chapter 7

Conclusions

7.1 Summary of thesis objectives

The focus of this thesis is to improve the understanding of the technology underlying costs of operation of rail-based public and road-based private modes of urban travel. As introduced in Chapter 1, the objectives of the thesis can be summarised under three main areas:

1. Understand the operational costs of urban rail transport (or metro) systems.
2. Quantify the production of vehicular travel in urban road networks.
3. Determine the mechanism driving congestion in near capacity metro operations.

The thesis objectives are addressed through the application of causal statistical modelling to large scale datasets such as a unique panel data on twenty-four metro operations from all across the globe, entry/exit gates and train movement data of the Mass Transit Railway, Hong Kong and the largest publicly available traffic sensor data comprising billions of vehicle observations from forty different cities. The statistical methods used in this study range from parametric to non-parametric methods.

7.2 Summary of thesis contributions

The research presented in this thesis contributes with novel causal models of operational costs of rail-based public and road-based private modes of urban travel and delivers new causal estimates to characterise the sources of cost-efficiency of these travel modes. The application of advanced causal statistical modelling techniques on large-scale datasets and application of advanced econometric modelling techniques enables a deeper level of understanding of the drivers of operational costs, compared to previous studies in the literature. The analysis presented in the thesis comprises of three parts corresponding to the three main thesis objectives summarised in Section 7.1. The following sections summarise the main contributions by chapter under the three research objectives.

The models estimated in the literature to quantify the operational costs of metro fail to control for observed and unobserved time-invariant and time-variant firm level sources of confounding. Under our first research objective that is addressed in Chapter 3, a rigorous understanding of such endogeneity issues in empirical estimation of a transport cost function is developed and an appropriate econometric framework is applied to address these issues. This chapter also contributes with a unique and very high quality panel data to estimate the technology underlying cost of short-run operations of metro systems. From the estimated cost function, new and more reliable empirical insights into the external sources of cost-efficiency for metro systems are provided.

The second research objective is addressed in Chapters 4 and 5, where a causal econometric framework is proposed to estimate the fundamental relationship (FR) of traffic flow or equivalently the input-out production relationship for travel in a road network. Compared to traditional methods which estimate only an *associational* and possibly *spurious* relationship by fitting a curve to a point cloud of observed traffic state variables, our proposed framework adjusts for potential sources of endogeneity/confounding biases such as observed and unobserved characteristics of driver behaviour, weather and demand. The proposed causal framework is different from the causal framework previously

used by some economists that is based on the interpretation of the speed-flow FR as the supply curve for travel; a limitation of the latter being that it seeks stationary state traffic conditions which seldom exist. As opposed to this economics approach, we present the first application of *causal inference* in empirical estimation of the FR from an engineering perspective that is based on the physics of movement of vehicles in a traffic stream, which is a key contribution of this thesis. We apply a Bayesian non-parametric instrumental variables (NPIV) estimator proposed by Wiesenfarth et al. (2014) that allows us to capture non-linearities in the relationship with a non-parametric specification without presuming the functional form and also adjust for any confounding bias via the use of instrumental variables (IVs).

In Chapter 4, the proposed causal framework is applied large-scale traffic detector data to estimate the flow-density FR for three highway bottlenecks in the US. The estimated relationship is also used to derive estimates of important features of the bottleneck such as capacity and capacity-drop. Previous studies have used different methodologies (for instance, change in cumulative vehicle count) to quantify the capacity-drop phenomenon. Thus, the proposed approach provides a one-stop solution to estimate an unbiased FR for a highway bottlenecks as well as important features such as capacity and capacity-drop. This chapter also contributes by conciliating the engineering and economics approaches to empirical estimation of the FR, the latter of which has recently led to inconclusive empirical evidence on the existence of capacity-drop that is well-established in the engineering literature.

Chapter 5 applies the proposed causal framework to quantify the travel efficiency that arises from increasing the provision of vehicular travel in urban road networks. This chapter estimates macroscopic FRs for homogeneously congested sub-networks in forty cities using the largest publicly-available traffic sensor data consisting of billions of vehicle observations. From the estimated relationships, novel estimates important policy inputs such as returns to density and network size are derived.

The idea of traffic fundamental diagram (FD) and the causal econometric framework for estimating it empirically in Chapters 4 and 5 is further extended to the context of metro operations in Chapter 6 to investigate the existence of FD-like relationships for metro systems. Novel station-level causal relationships between boarding-alighting movements and train flow are estimated using data from entry/exit gates and train movement data of the Mass Transit Railway, Hong Kong, and conclusions are drawn regarding the mechanism that drives passenger-congestion in the network, which addresses the third research objective. Potential bottleneck stations in the network are identified via the estimated relationships and their corresponding optimum boarding-alighting movements are reported. To current knowledge, this chapter presents the first application of econometric modelling to model congestion in high-frequency metro operations, and this is the main contribution of this chapter.

7.3 Summary of main findings

Beyond the initial introductory chapter, the second chapter of this thesis, that is, the Literature Review chapter, provides the context for the research presented in the subsequent analysis chapters.

The first analysis chapter develops a comprehensive understanding of operational costs of urban rail transport (metro) and determines the important aspects of the technology that drives unit cost differences between metro firms. This chapter uses dynamic panel generalised method of moments (DPGMM) with a very high quality panel dataset on metro operations to estimate the underlying cost function. The key methodological improvement offered in this chapter is to control for observed and unobserved time-invariant and time-variant firm level sources of confounding in the estimation of a transport cost function. The DPGMM is illustrated as an attractive tool for the cost function estimation because it permits flexible representation of unobserved productivity level differences between firms and offers better remedies for endogenous covariates. A comparison of our DPGMM results

with the traditional estimation methods like pooled ordinary least squares estimation confirms that failure to account for unobserved productivity differences between firms in empirical cost analysis creates a downward bias in the estimates of RTS and RTD. The estimated RTD is 1.562 as opposed to an estimate of 1.40 from the literature, both of these estimates being statistically greater than one. As the literature suggests, increasing RTD results from the existence a range of fixed and semi-fixed costs are prevalent in the urban rail transport industry that do not vary proportionally with output.

The empirical evidence in this chapter also supports increasing RTS, which justifies the presence of large size firms in urban rail transport industry. The estimated RTD is 1.223, which is again statistically greater than one. The weight of evidence in the urban rail transport literature indicates that the industry is characterised by constant RTS. However, we find that controlling for endogeneity bias in empirical cost analysis and accounting for dynamics in firm-level productivity gives RTS estimates that is consistent with the observed industry behaviour. Furthermore, our data indicates that around eighty-percent of way and structure maintenance costs comprise of labour and electricity costs, which can be varied in the short-run. We, therefore, include infrastructure maintenance costs as a component of variable costs in our short-run operational cost analysis. Increasing returns to scale may have resulted from the presence of cost complementarities between operational and way and track cost components as found in case of mainline railways. We also study other aspects of the underlying production technology. We find that the marginal rate of technical substitution between any two inputs for production of metro output depends on the prices of other inputs, that is, the underlying technology shows non-separability of input factors. Our results also show non-homotheticity implying that changes in factor prices affects both cost elasticity and corresponding factor demand. Therefore, scale economies in provision of urban rail transport services are not independent of input prices.

The second analysis chapter develops a comprehensive understanding of traffic flow in a

highway section by adopting a causal econometric framework to determine the relationship between traffic flow and occupancy in a highway section with a downstream bottleneck. A Bayesian non-parametric instrumental variables (NPIV) estimator is applied on data from three highway bottlenecks in California. The use of NPIV is attractive as it allows us to capture non-linearities in the FR with a fully flexible non-parametric specification and adjusts for confounding bias via the inclusion of relevant and exogenous instruments. Such confounding biases may occur because of many external observed or unobserved factors such as driver behaviour, heterogeneous vehicles, weather and demand, that are correlated with both observed traffic variables.

This chapter also reconciles the economics and engineering approaches to estimate the empirical FR of traffic flow. One prominent economic approach is based on a demand-supply framework where users of the highway section are treated as suppliers of travel in the section and outflow from the highway section in turn represents the travel supplied. However, we note that the equivalence of the FR of traffic flow and the supply curve for travel in a highway section can only be considered under stationary state traffic conditions, which seldom exist particularly under congested traffic conditions. We thus argue that the demand-supply framework may lead to misrepresentation in developing a causal understanding of the empirical FD. We instead adopt causal statistical modelling within the engineering framework which is based on the physical laws that govern the movement of vehicles in a traffic stream.

The above themes are important as a recent study in the economics literature examines the changes in outflow with increasing demand for three different highway bottlenecks in California and finds no evidence of drop in capacity or in other words, hypercongestion during periods of high demand. The study concludes that the fundamental (flow-density or flow-speed) diagram for a highway section should not exhibit a backward bending part and also questions the applicability of traffic control measures and congestion pricing policies that are aimed at regulating demand to avoid hypercongestion. Based on our

estimated causal FR, we re-evaluate the existence of capacity-drop in highway bottlenecks, which is a well-established phenomenon in the engineering literature.

The empirical results in this chapter show a statistically significant decrease in flow upon activation of the bottleneck in two out of three analysed bottlenecks, thus supporting the existence of capacity drop. The estimated capacity-drop varies on a case-to-case basis depending upon the geometry of the bottleneck as well as the characteristics of the average traffic stream passing through it. However, after this drop in capacity, we do not find sufficient statistical evidence to support any changes in flow with further increase in occupancy in isolated highway sections. We thus argue that as the flow through the bottleneck remains constant following the capacity-drop, the flow-occupancy curve is not actually backward bending. However, a statistically-significant backward bending relationship exists only when the highway section is not perfectly isolated from downstream obstacles that cause traffic flow through the section to decrease over occupancy in a predictable way.

The third analysis chapter develops a comprehension of the production of vehicular travel in urban road networks and quantifies the technical efficiency in the travel production process. To do so, it estimates macroscopic fundamental relationships for homogeneously congested sub-networks (reservoirs) in thirty-four cities around the globe. It adopts the causal framework from Chapter 4 to obtain unbiased estimates of the reservoir-level flow-density relationship using large-scale traffic sensor data.

The empirical estimates from this chapter show the presence of decreasing returns to density in the provision of vehicular travel in cities. Thus, any increase in vehicle hours travelled in a fixed road network results in less than proportionate increase in vehicle kilometres travelled in the network. Across the thirty-four reservoirs analysed, the mean estimate of RTD at the average-level of occupancy is 0.779 and the associated standard deviation is 0.151. At the mean-level of peak-hour occupancy across all reservoirs, the average estimate of RTD is 0.631 with a standard deviation of 0.208. Furthermore, we

also find that vehicular travel is produced with decreasing returns to scale in cities. Our estimated RTS of 0.300 implies a less than proportionate increase in the vehicle kilometres travelled in the network with equi-proportionate increase in vehicle hours travelled and network length.

The final analysis chapter develops an understanding of the mechanism driving congestion delays in near capacity metro operations within an econometric framework. In particular, it focuses on how high volumes of passenger boardings and alightings may lead to increased dwell times at stations (passenger-congestion), that may eventually cause queuing of trains in upstream (train-congestion). Such stations act as active bottlenecks in the metro network and congestion may propagate from these bottlenecks to the entire network. This chapter analyses passenger-congestion at stations, which is generally the root cause of the congestion phenomenon. It conducts the first station-level econometric analysis to estimate a causal relationship between boarding-alighting movements and train flow using data from entry/exit gates and train movement data of the Mass Transit Railway, Hong Kong. It adopts a Bayesian non-parametric spline-based regression approach and apply instrumental variables estimation to control for confounding bias that may occur due to unobserved characteristics of metro operations. Our estimates point towards the existence of traffic fundamental-diagram-like-relationships in metro network.

Since excessive passenger movement at bottleneck stations is the primary driver of congestion, we expect a *critical boarding-alighting movement* level exists in metros at bottleneck stations, above which train flow or throughput of the station reduces. This intuition is analogous to the road traffic flow theory, which presents evidence of a drop in traffic flow through a highway bottleneck above a critical vehicular density (see Daganzo 1997, for the fundamental diagram of traffic flow). Based on this analogy, we identify bottleneck stations using the estimated station-level relationships and provide estimates of optimum passenger movements per train and service frequencies at the bottleneck stations. We discuss how these estimates, along with real data on daily demand, could assist metro

operators in devising station-level control strategies.

7.4 Potential applications

There are a number of potential applications of the research presented in the thesis, which can be broadly categorised into applications related to practical policy-making and transport operations. The applications can be summarised as follows:

1. Applications in practical policy-making

    - Appraisal of transport investments - The estimates of returns to scale and density derived in Chapters 3 and 5 are important inputs in the economic appraisal of transportation projects. The presence of increasing returns to network size in metro operations and decreasing returns to network size in increasing provision of vehicle travel in cities may be relevant from a policy point of view, particularly for the economic appraisal of large infrastructure projects that lead to network expansion.
    - Design of highways section and urban road networks - Our causal estimates of the FR from Chapters 5 and 6 are crucial for design of highway sections and urban road networks, as these estimates provide a more generalised and robust characterisation of the traffic flow and adjusts for any potential confounding biases. Our causal models of traffic flow are, therefore, more suited for standard reference manuals like the highway capacity manual (HCM) and the UK Cost Benefit Analysis (UK-CoBA) manual.
    - Design of metro networks - The empirical evidence on existence of traffic FD like relationships in metro networks presented in Chapter 6 opens up a whole new research area with many potential applications that can be borrowed from the traffic flow theory. As the traffic FD is used in design of road networks,

the public transport equivalent of this diagram could be relevant for design of metro networks.

2. Operations related applications

   - Real time information - Improved estimates of the traffic FR, or equivalently, the vehicle kilometres travelled (VKT) and vehicle hours travelled (VHT) relationship from Chapters 4 and 5, when used together with short-term traffic demand prediction models (see, for instance, Van Lint & Van Hinsbergen 2012, and other references therein), could be important inputs to real-time user information systems that provide travel time estimates to road users.
   - Key Performance Indicators - Results from Chapter 6 on modelling of passenger-congestion delays could be used in development of new useful key performance indicators based on large-scale datasets to measure congestion delay and overall reliability of metro services. Analogous to level-of-service indicators for highways and road-networks, these models can also be used to define new indicators of level of service for different parts of a metro network. Moreover, the returns to scale estimates from Chapter 3 could be useful for conditional benchmarking of metro operations, that is, benchmarking of costs after adjusting for external benefits resulting from scale and density of operations.
   - Congestion delay prediction and management in metro networks- The results presented in Chapter 6 enhances the value of short-term demand prediction (see, Ding et al. 2016, Ma et al. 2018, for instance,) by estimating its causal impact on train frequencies. Understanding the dynamics of passenger movements and train frequencies, along with estimates of optimum passenger movement, can help metro operators in designing data-driven strategies to control passenger movements and minimise delays. Such control strategies may involve i) adopting platform management practices such as reducing escalator capacity, ii)

deployment of staff resources to regulate the entry of passengers into bottleneck stations, and iii) pricing policies. Another strategy could be of passengers entering stations to increase overall system throughput. Daganzo (2005) suggests such strategies in the context of vehicular traffic control in urban networks.

7.5 Future Work

A number of potential avenues for future research that is based on the analyses presented in this thesis have been identified. The following paragraphs summarise these potential research questions:

Although Chapter 3 focuses on short-run operational costs of metro systems, the discussion on endogeneity issues in empirical estimation of cost functions, the downward bias in returns to scale estimates in the literature and the treatment of endogeneity via application of application of appropriate statistical tools applies to analyses of the wider transport cost function literature. The methodological framework proposed in provides a general specification that could be useful in cost analysis in other modes of transportation, whether be mainline railways, bus or airline operations. Another important future research question relates to economic appraisal of large infrastructure projects. The presence of network size economies in metro operations estimated in this chapter, may be relevant from a policy point of view, particularly for the economic appraisal of large infrastructure projects that lead to network expansion. Returns to network size implies that such investments may generate external benefits in the form of a network-wide reduction in operational costs. It would be interesting to quantify this external benefit and assess whether it could have significant impact on the outcome of traditional cost-benefit analyses. Furthermore, the application of returns to scale estimates for conditional benchmarking could be another interesting area of future research. One such application is demonstrated in Chapter 3, where unadjusted (unconditional) and adjusted (conditional on the presence of external costs benefits) costs of metro operations are compared.

In Chapter 4, we discuss the endogeneity/confounding biases in empirical estimation of the traffic FR and propose a flexible causal statistical framework to adjust for these biases and produce a more robust characterisation of this relationship. The empirical evidence in this paper is limited to three highway bottlenecks in the US, hence, it would be interesting to extend this framework to analyse highway bottlenecks in other countries. Moreover, although we apply our proposed framework to estimate this relationship for highway bottlenecks, it can be directly adopted to estimate a causal model of traffic flow for a uniform highway section. For a uniform highway section that is well-isolated from downstream influences, it will be interesting to re-evaluate the existence of hypercongestion (the backward-bending of the flow-density or flow-speed relationship with increase in demand). Furthermore, as the demand-supply interpretation of FR may be misleading particularly under congested traffic conditions where traffic conditions are dynamic (changing rapidly), it would be interesting to revisit the literature on congestion pricing that is based on this interpretation.

For estimating returns to network size (RTS) in Chapter 5, we identify and pool data from only one homogeneously congested reservoir in each city. Moreover, Chapter 5 produces estimates of returns to density (RTD) for different homogeneously congested regions in forty cities. We note a substantial variation in these estimates across reservoirs. Similar to the study by Loder et al. (2019) which tries to explain the differences in capacity across reservoirs, it would be interesting to identify the factors that explain the differences in the RTD estimates. These results may have profound implications on how to build and operate cities more efficiently. Furthermore, from policy point of view, it would be interesting to quantify the external dis-benefit from existence of decreasing RTS in increasing provision of vehicular travel in cities and assess whether it could have significant impact on the outcome of traditional cost-benefit analyses.

Chapter 6 presents the first station-level analysis of congestion in a metro network by estimating a causal relationship between passenger movement per train and train flow.

These estimates, along with real data on daily demand, could assist metro operators in devising data-driven station-level passenger inflow control strategies. Such strategies are presently implemented by metro operators around the world based on their day-to-day experience of congestion patterns at various stations. Developing and testing such control measures using real data is an important avenue for future research. Another interesting area of future research could be to explore the potential of the estimated relationships in the long-run to understand the level of operational service and guide improvements in the metro network. Moreover, the empirical evidence presented in this study is limited to real data from the MTR, Hong Kong network. It would be interesting to replicate this study for other metro high-capacity metro operations to reinforce our empirical study. Furthermore, this study focuses on analysis of the root cause of congestion delays in metro networks – increased passenger boarding and alighting movements at bottleneck stations in a metro network. It would be interesting to statistically model the propagation of these delays from the bottleneck stations in order to understand their effect on system-wide performance and reliability.

As a final remark, the Bayesian NPIV estimator adopted in Chapters 4, 5 and 6 of this thesis can currently deal with only one endogenous covariate. As the estimator delivers promising results, it would be worthwhile to extend this estimator to allow for multiple endogenous covariates. Developing tests to quantify the strength of instruments in a non-parametric instrumental variables regression could be another area of future research.

References

Daganzo, C. F. (1997), *Fundamentals of transportation and traffic operations*, Vol. 30, Pergamon Oxford.

Daganzo, C. F. (2005), 'Improving city mobility through gridlock control: an approach and some ideas'.

Ding, C., Wang, D., Ma, X. & Li, H. (2016), 'Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees', *Sustainability* **8**(11), 1100.

Loder, A., Ambühl, L., Menendez, M. & Axhausen, K. W. (2019), 'Understanding traffic capacity of urban networks', *Scientific reports* **9**(1), 1–10.

Ma, X., Zhang, J., Du, B., Ding, C. & Sun, L. (2018), 'Parallel architecture of convolutional bi-directional lstm neural networks for network-wide metro ridership prediction', *IEEE Transactions on Intelligent Transportation Systems* **20**(6), 2278–2288.

Van Lint, J. & Van Hinsbergen, C. (2012), 'Short-term traffic and travel time prediction models', *Artificial Intelligence Applications to Critical Transportation Issues* **22**(1), 22–41.

Wiesenfarth, M., Hisgen, C. M., Kneib, T. & Cadarso-Suarez, C. (2014), 'Bayesian nonparametric instrumental variables regression based on penalized splines and dirichlet process mixtures', *Journal of Business and Economic Statistics* **32**(3), 468–482.

Appendix A

Supplementary Material: Chapter 3

A.1 Description of metro operational cost data

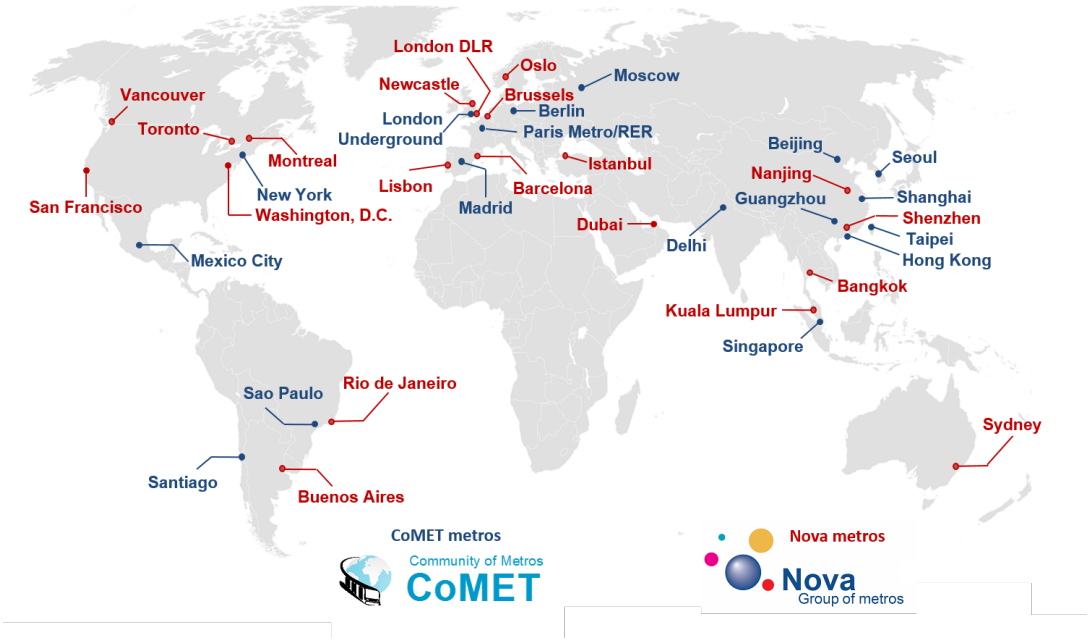


Figure A.1: Metro operations reported in the TSC data.

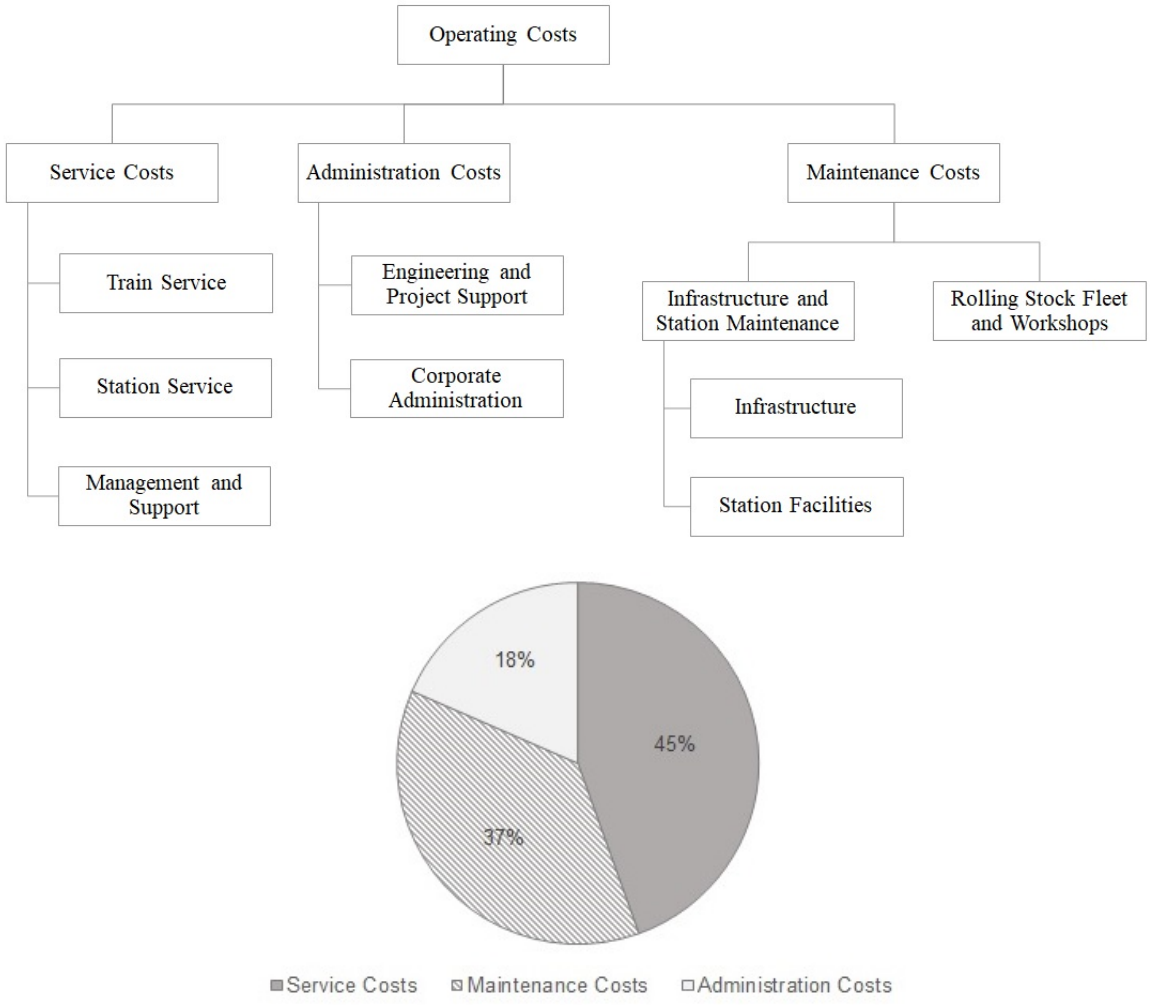


Figure A.2: Components of metro operational costs as in the TSC data.

A.2 Full summary of results

Table A.1: Summary of Results of the Short-run Cost Model.

Explanatory variables (logarithms	Static Panel Models			Dynamic Panel Models	
except for dummy variables)	POLS	FE	IV: Sys GMM	AR(1): Diff GMM	AR(1): Sys GMM
Car kms	0.815***	0.506***	0.838***	0.472***	0.640***
	(0.063)	(0.078)	(0.141)	(0.110)	(0.116)
Network length	0.142**	0.264***	0.140	0.356***	0.177*
	(0.069)	(0.079)	(0.124)	(0.109)	(0.097)
Load Factor	0.340***	0.173**	0.312*	0.261*	0.298*
	(0.093)	(0.070)	(0.186)	(0.136)	(0.160)
Labour Price	0.574***	0.642***	0.578***	0.634***	0.485***
	(0.046)	(0.036)	(0.055)	(0.052)	(0.048)
Energy Price	0.231***	0.177***	0.240***	0.188***	0.208***
	(0.036)	(0.034)	(0.052)	(0.050)	(0.046)
Residual Price	0.195***	0.181***	0.182***	0.178***	0.307***
	(0.048)	(0.021)	(0.063)	(0.023)	(0.056)
Car kms ^{2}	-0.026	0.306***	0.133	0.224**	0.390
	(0.143)	(0.113)	(0.386)	(0.108)	(0.299)
Network length ^{2}	-0.046	0.303**	0.058	0.289**	0.322
	(0.174)	(0.148)	(0.371)	(0.131)	(0.288)
Load Factor^2	0.145	0.164	0.278	0.172	0.376
	(0.187)	(0.132)	(0.323)	(0.205)	(0.278)
Labour Price ^{2}	0.134***	0.177***	0.176	0.152***	0.140*
	(0.052)	(0.036)	(0.117)	(0.051)	(0.083)
Energy Price ^{2}	0.121***	0.100***	0.134**	0.092***	0.162***
	(0.039)	(0.031)	(0.054)	(0.035)	(0.060)
Residual Price ^{2}	0.007	0.008	-0.009	0.000	0.011
	(0.037)	(0.013)	(0.036)	(0.016)	(0.032)
Car kms x Network length	0.152	-0.559**	-0.109	-0.444**	-0.640
	(0.316)	(0.255)	(0.744)	(0.221)	(0.579)
Car kms x Load Factor	0.839***	-0.690***	0.676*	-0.727***	0.244
	(0.252)	(0.180)	(0.395)	(0.282)	(0.410)
Car kms x Labour Price	0.224**	-0.244***	0.202*	-0.184	0.028
	(0.095)	(0.090)	(0.121)	(0.117)	(0.159)
Car kms x Energy Price	-0.297***	0.227***	-0.267*	0.189*	-0.200
	(0.095)	(0.080)	(0.140)	(0.106)	(0.158)
Car kms x Residual Price	0.072	0.017	0.065	-0.005	0.172**
	(0.087)	(0.039)	(0.105)	(0.041)	(0.076)
Network length x Load Factor	-1.312***	0.713***	-1.180***	0.723**	-0.628
	(0.312)	(0.216)	(0.416)	(0.354)	(0.448)

Table A.1 Continued from previous page.

Explanatory variables (logarithms	Static Panel Models			Dynamic Panel Models	
except for dummy variables)	POLS	FE	IV: Sys GMM	AR(1): Diff GMM	AR(1): Sys GMM
Network length x Labour Price	-0.403***	0.178*	-0.396***	0.143	-0.210
	(0.123)	(0.103)	(0.111)	(0.132)	(0.157)
Network length x Energy Price	0.438***	-0.188**	0.421***	-0.184*	0.340**
	(0.112)	(0.085)	(0.136)	(0.102)	(0.147)
Network length x Residual Price	-0.035	0.010	-0.024	0.040	-0.130
	(0.095)	(0.043)	(0.100)	(0.044)	(0.103)
Load Factor x Labour Price	0.350**	0.236**	0.466**	0.237	0.506***
	(0.138)	(0.097)	(0.229)	(0.164)	(0.188)
Load Factor x Energy Price	-0.321***	-0.245**	-0.401**	-0.247	-0.459***
	(0.108)	(0.099)	(0.180)	(0.155)	(0.156)
Load Factor x Residual Price	-0.029	0.009	-0.065	0.010	-0.047
	(0.080)	(0.039)	(0.125)	(0.040)	(0.092)
Labour Price x Energy Price	-0.124***	-0.135***	-0.159**	-0.122***	-0.145**
	(0.041)	(0.029)	(0.080)	(0.039)	(0.066)
Labour Price x Residual Price	-0.010	-0.043**	-0.017	-0.030	0.005
	(0.038)	(0.020)	(0.060)	(0.023)	(0.049)
Energy Price x Residual Price	0.003	0.035**	0.025	0.030	-0.016
	(0.040)	(0.018)	(0.064)	(0.019)	(0.046)
Lag (Dependent Variable)	-	-	-	0.020	0.196***
				(0.016)	(0.064)
Year Effects Included	YES	YES	YES	YES	YES
No. of Observations	165	165	140	119	119
Adjusted R - square	0.974	0.998	-	-	-
Arellano-Bond test for AR(1)	-	-	z = -1.59	z = -1.93	z = -1.86
			Pr > z = 0.112	Pr > z = 0.053	Pr > z = 0.052
Arellano-Bond test for AR(2)	-	-	z = 0.14	z = -1.91	z = -1.21
			Pr > z = 0.891	Pr > z = 0.056	Pr > z = 0.226
Sargan Test	-	-	$\chi^2(126) = 1181.96$	$\chi^2(84) = 180.50$	$\chi^2(103) = 412.14$
			Pr > χ^2 = 0.000	Pr > χ^2 = 0.000	Pr > χ^2 = 0.000
No. of Instruments	-	-	161	119	131
Returns to Density (RTD)	1.227***	1.978***	1.193***	2.119***	1.562***
	(0.096)	(0.306)	(0.201)	(0.492)	(0.283)
Returns to Scale (RTS)	1.045***	1.300***	1.023***	1.207***	1.223***
	(0.028)	(0.079)	(0.058)	(0.081)	(0.081)

(i) Figures in brackets denote the standard errors associated with the estimates.

(ii) Significance: (***) 99 percent, (**) 95 percent, (*) 90 percent.

Table A.1 Continued from previous page.

Explanatory variables (logarithms	Static Panel Models			Dynamic Panel Models	
except for dummy variables)	POLS	FE	IV: Sys GMM	AR(1): Diff GMM	AR(1): Sys GMM

(iii) Estimation Methods from left to right increase in terms of flexibility and provide more control for endogeneity.

Table A.2: Summary of RTD and RTS estimates obtained using different methodologies.

Estimation	Returns to Density (RTD)				Returns to Scale (RTS)			
Methodology	Coef.	Std. Err.	95% C.I.		Coef.	Std. Err.	95% C.I.	
POLS	1.227	0.0956	1.040	1.414	1.045	0.028	0.989	1.101
FE	1.978	0.306	1.378	2.578	1.300	0.079	1.145	1.455
IV (Sys. GMM)	1.193	0.201	0.799	1.587	1.023	0.058	0.909	1.137
AR(1) (Diff. GMM)	2.119	0.492	1.155	3.082	1.207	0.081	1.048	1.366
AR(1) (Sys. GMM)	1.562	0.283	1.006	2.117	1.223	0.081	1.065	1.381

*Coef. stands for estimated Coefficient; Std. Err for associated Standard Error

**C.I. denotes Confidence Interval

A.3 Robustness check against exogeneity of factor prices

In this section, we carry out robustness checks to test the sensitivity of our results to treatment of factor prices, that is, labour price, energy price and residual prices, as exogenous instead of endogenous as in our proposed model. For the purpose of demonstration, we use a Cobb Douglas cost function as using this functional specification allows for direct comparison of parameter estimates from the cost model. We apply the System GMM estimation.

Table A.3: Robustness check against treatment of factor prices as endogenous.

Explanatory variable	Estimate with factor prices treated as endogenous	Estimate with factor prices treated as exogenous
Car kms	0.571 (0.111)***	0.557 (0.102)***
Network length	0.347 (0.127)***	0.361 (0.112)***
Load Factor	0.643 (0.125)***	0.658 (0.127)***
Labour Price	0.691 (0.044)***	0.701 (0.044)***
Energy Price	0.094 (0.062)	0.090 (0.066)
Residual Price	0.214 (0.040)***	0.209 (0.047)***
Lag (Dependent Variable)	0.076 (0.042)*	0.080 (0.048)*
Year Effects Included	YES	YES
No. of Observations	119	119
No. of Instruments	134	131
Arellano-Bond test for AR(1)	z = -1.98, Pr > z = 0.048	z = -1.84, Pr > z = 0.066
Arellano-Bond test for AR(2)	z = -0.96, Pr > z = 0.339	z = -0.91, Pr > z = 0.365
Sargan Test of over-identifying restrictions:	$\chi^2(113) = 417.85$, Pr > $\chi^2 = 0.00$	$\chi^2(110) = 391.30$, Pr > $\chi^2 = 0.00$
Returns to Scale (RTS)	1.088 (0.039)***	1.089 (0.039)***
Returns to Density (RTD)	1.750 (0.340)***	1.796 (0.328)***

Notes:

(1) All explanatory variables are in their logarithmic form except for dummy variables.

(2) Significance: (***) 99 percent, (**) 95 percent, (*) 90 percent.

(3) Figures in bracket indicate the associated robust standard errors.

We find that the parameter estimates of the cost model, and thus the scale economy estimates are not substantively different in the two cases.[1]

[1]Similar tests have been carried out for a translog specification and the resulting RTD and RTS estimates have been found to be substantively the same. Results can be produced upon request.

A.4 Robustness check against inclusion of residual prices

In this section, we carry out robustness checks to test the sensitivity of our results to exclusion of residual price as one of the factor prices from our proposed cost model. For the purpose of demonstration, we use a Cobb Douglas cost function as using this functional specification allows for direct comparison of parameter estimates from the cost model. We apply the System GMM estimation.

Table A.4: Robustness check against inclusion of residual prices in the cost model.

Explanatory variable	Estimate with residual price included in the model	Estimate with residual price excluded from the model
Car kms	0.571 (0.111)***	0.217 (0.109)**
Network length	0.347 (0.127)***	0.259 (0.079)***
Load Factor	0.643 (0.125)***	0.412 (0.079)***
Labour Price	0.691 (0.044)***	0.385 (0.064)***
Energy Price	0.094 (0.062)	0.615 (0.064)***
Residual Price	0.214 (0.040)***	-
Lag (Dependent Variable)	0.076 (0.042)*	0.560 (0.090)***
Year Effects Included	YES	YES
No. of Observations	119	119
No. of Instruments	134	131
Arellano-Bond test for AR(1)	z = -1.98, Pr > z = 0.048	z = -3.11, Pr > z = 0.002
Arellano-Bond test for AR(2)	z = -0.96, Pr > z = 0.339	z = -0.44, Pr > z = 0.662
Sargan Test of over-identifying restrictions:	$\chi^2(113) = 417.85$, Pr > $\chi^2 = 0.00$	$\chi^2(109) = 209.18$, Pr > $\chi^2 = 0.00$
Returns to Scale (RTS)	1.088 (0.039)***	2.098 (0.374)***
Returns to Density (RTD)	1.750 (0.340)***	4.604 (2.306)**

Notes:

(1) All explanatory variables are in their logarithmic form except for dummy variables.

(2) Significance: (***) 99 percent, (**) 95 percent, (*) 90 percent.

(3) Figures in bracket indicate the associated robust standard errors.

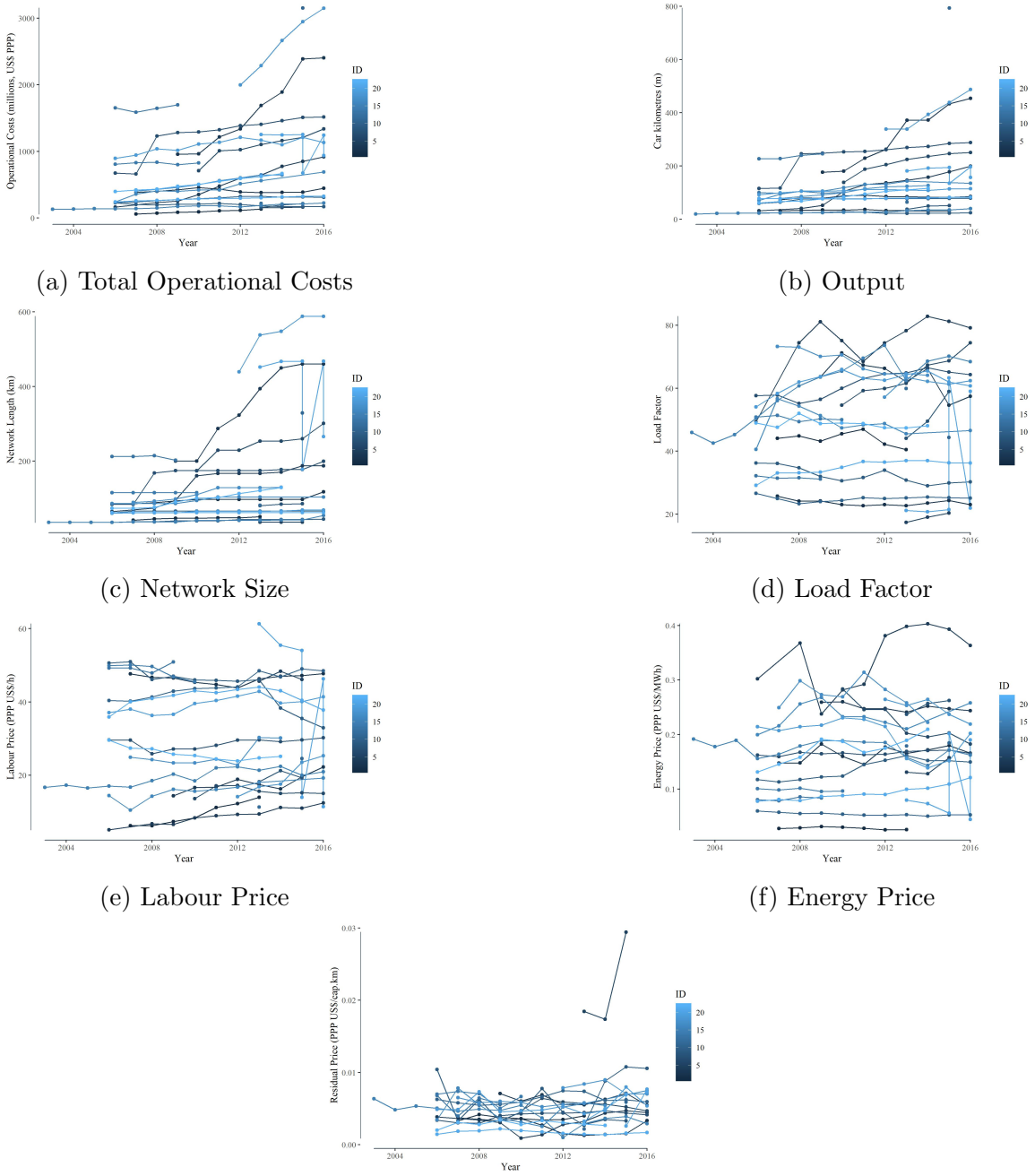
We find that the parameter estimates when residual prices are excluded from the cost

model are not very plausible as the scale economy estimates are unreasonably high.[2]

[2]Similar tests have been carried out for a translog specification and the resulting RTD and RTS estimates for the cost model without residual costs have been found to be unreasonably high. Results can be produced upon request.

A.5 Annual variation in variables

In this section, we present the annual variation in the variables used in this analysis for different metro systems.



(a) Total Operational Costs

(b) Output

(c) Network Size

(d) Load Factor

(e) Labour Price

(f) Energy Price

(g) Residual Price

Figure A.3: Annual variation in total operational costs and its descriptors for different metro systems.

Due to the sensitive commercial nature of the data, we have presented the figures in an anonymised form.

Appendix B

Supplementary Material: Chapter 4

In this appendix, we demonstrate the potential sources of confounding discussed in Section 4.4.2 in the fundamental relationship of traffic flow in mathematical terms.

B.1 Omitted Variable Bias

To illustrate the endogeneity bias due to omitted covariates, we simplify equation 4.1, where we consider that $S(.)$ has a linear specification, that is $S(\mathbf{o}) = \mathbf{o}\beta$. We suppose that $\delta = \mathbf{w}\alpha$, where $\mathbf{w}$ represents, say, driving and vehicular characteristics. For the notational simplicity, we drop time-day subscripts and superscripts. We, thus, have a data generating process given by:

$$\mathbf{q} = \mathbf{o}\beta + \mathbf{w}\alpha + \xi, \tag{B.1}$$

where $\mathbf{q}$ is an $N \times 1$ vector of dependent variables, $\mathbf{o}$ and $\mathbf{w}$ are $N \times 1$ and $N \times K$ matrices and ξ is an $N \times 1$ error vector that is assumed to be uncorrelated with $\mathbf{o}$ and $\mathbf{w}$. Application of a standard regression technique such as an ordinary least squares (OLS) estimation of $\mathbf{q}$ on $\mathbf{o}$ and $\mathbf{w}$ yields consistent parameter estimates of α and β [1].

[1] Note that an estimator $\hat{\beta}$ is said to be consistent for β if it converges in probability to the true value β, that is, $\text{plim}(\hat{\beta}) \to \beta$.

Suppose instead that $\mathbf{w}$ is omitted from the equation and $\mathbf{q}$ is regressed on $\mathbf{o}$ alone. Then $\mathbf{w}\alpha$ becomes a part of the error term and the estimated model becomes:

$$\mathbf{q} = \mathbf{o}\beta + (\mathbf{w}\alpha + \xi),$$

where $\mathbf{w}\alpha + \xi$ is the new error term. The OLS estimator of β equals:

$$\begin{aligned}
\beta_{OLS} &= (\mathbf{o}'\mathbf{o})^{-1}\mathbf{o}'\mathbf{q} \\
&= (\mathbf{o}'\mathbf{o})^{-1}\mathbf{o}'(\mathbf{o}\beta + \mathbf{w}\alpha + \xi) \\
&= (\mathbf{o}'\mathbf{o})^{-1}\mathbf{o}'\mathbf{o}\beta + (\mathbf{o}'\mathbf{o})^{-1}\mathbf{o}'\mathbf{w}\alpha + (\mathbf{o}'\mathbf{o})^{-1}\mathbf{o}'\xi \\
&= \beta + (N^{-1}\mathbf{o}'\mathbf{o}^{-1})(N^{-1}\mathbf{o}'\mathbf{w})\alpha + (N^{-1}\mathbf{o}'\mathbf{o})^{-1}(N^{-1}\mathbf{o}'\xi)
\end{aligned}$$

Under the assumption that $\mathbf{o}$ is uncorrelated with ξ, the final term has probability limit zero. However, because, $\mathbf{o}$ is correlated with $\mathbf{w}$,

$$\text{plim}[\beta_{OLS}] = \beta + \delta\alpha$$

where, $\delta = \text{plim}[(N^{-1}\mathbf{o}'\mathbf{o}^{-1})(N^{-1}\mathbf{o}'\mathbf{w})]$ is the probability limit of the OLS estimator in the regression of the omitted regressor ($\mathbf{w}$) on the included regressors ($\mathbf{o}$). This inconsistency is called omitted variable bias, which exists as long as the omitted regressor is correlated with the included regressors. In general the inconsistency could be positive or negative. A positive bias exists if the correlation between $\mathbf{o}$ and $\mathbf{w}$, that is, δ and that between $\mathbf{q}$ and $\mathbf{w}$, that is, α are both either positive or negative, that is, $\alpha\delta > 0$. If these correlations are of opposite sign, that is, $\alpha\delta < 0$, the bias is negative. For instance, if $\mathbf{w}$ represents the risk-taking ability of drivers, we may expect a positive correlation between $\mathbf{o}$ and $\mathbf{w}$ as well as $\mathbf{q}$ and $\mathbf{w}$, resulting in positive bias due to omission of drivers' risk taking abilities. This is because we may expect an average population of risk taking drivers to drive at smaller headways or higher densities even at very high speeds, thus resulting into larger

flows.

In Table B.1, we enlist various confounders for the fundamental relationship based on the literature (refer Section 4.2.1) and their expected correlations with occupancy and flow.

Table B.1: Various sources of confounding in the fundamental relationship.

Confounder	Expected correlation with flow	Expected correlation with occupancy
Risk-taking behaviour of drivers	+	+
Risk-averse behaviour of drivers	-	-
Vehicle accelerations	-	+
Vehicle decelerations	+	-
Lane change manoeuvres	+/-	+/-
Vehicle lengths	+/-	+/-
Detector-level (measurement) errors	+/-	+/-
Weather conditions	+/-	+/-
Other characteristics of demand	+/-	+/-

B.2 Reverse Causality

To illustrate bias due to reverse causality, we further simplify the data generating process in equation B.1 as follows:

$$\mathbf{q} = \mathbf{o}\beta + \xi, \tag{B.2}$$

To obtain an unbiased estimate of β via OLS, the Gauss Markov condition of zero conditional mean of errors, that is, $E[\xi|o] = 0$, or in other terms, $\text{Cov}[\xi, o] = 0$,must be satisfied. In case of reverse causality, there exists another data generating process given by:

$$\mathbf{o} = \mathbf{q}\gamma + \psi, \tag{B.3}$$

Consequently, we have,

$$\begin{aligned}
\text{Cov}[\xi, o] &= \text{Cov}[\xi, (q\gamma + \psi)] \\
&= \gamma \text{Cov}[\xi, q] \quad \text{assuming that} \quad \xi \perp \psi \\
&= \gamma \text{Cov}[\xi, (o\beta + \xi)] \\
&= \gamma \text{Cov}[\xi, o\beta] + \text{Var}(\xi) \\
&\neq 0
\end{aligned}$$

Thus, the zero conditional mean assumption of errors is violated and OLS may result into a biased estimate of β.

Appendix C

Supplementary Material: Chapter 5

C.1 Estimated Reservoir-level MFDs

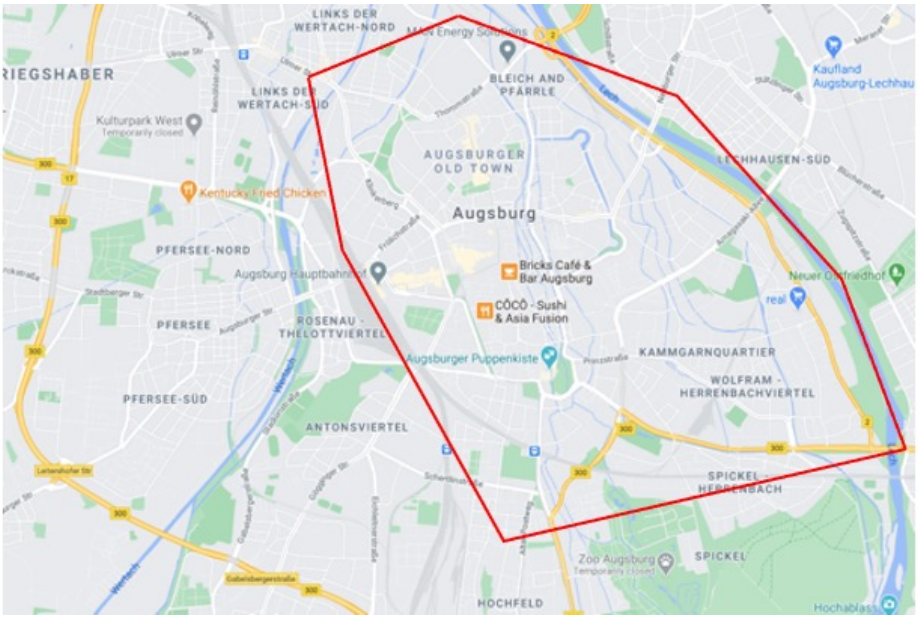
Figures C.1-C.34 illustrate the estimated MFDs for the thirty-five reservoirs studied in Chapter 5.

C.2 Distribution of Errors

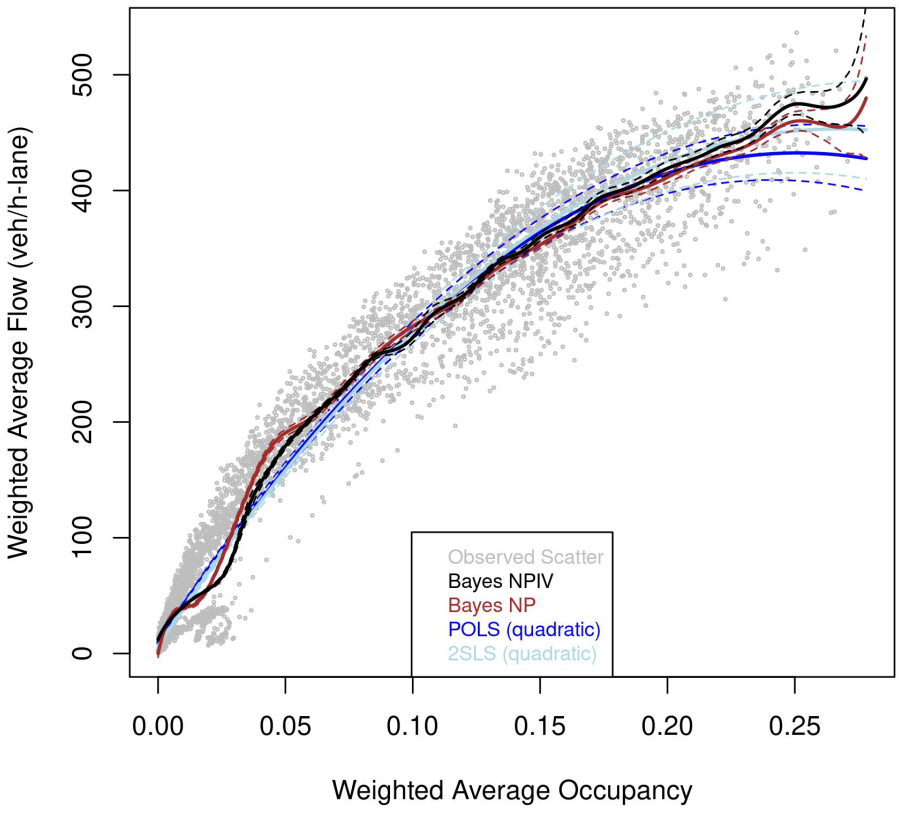
Figure C.35 shows the contour plot of the joint distribution of errors from the first stage (ϵ_1) and the second stage (ϵ_2).

C.3 Relevance of Instruments

Figure C.36 illustrates the results (that is, the estimated $h(.)$) from regression of the endogenous covariate on the instrument for the three highway sections.

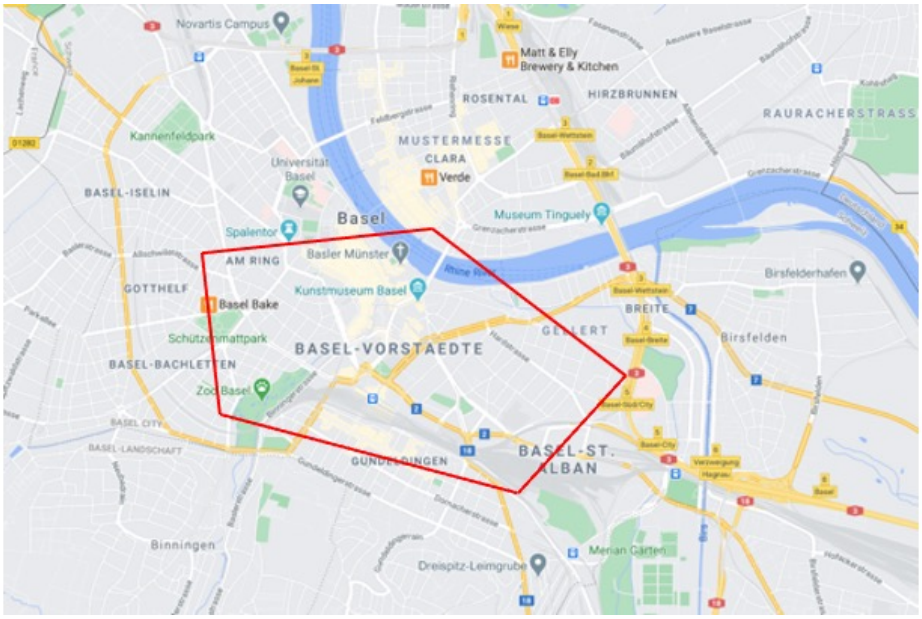


(a) Network exhibit used for the MFD estimation.

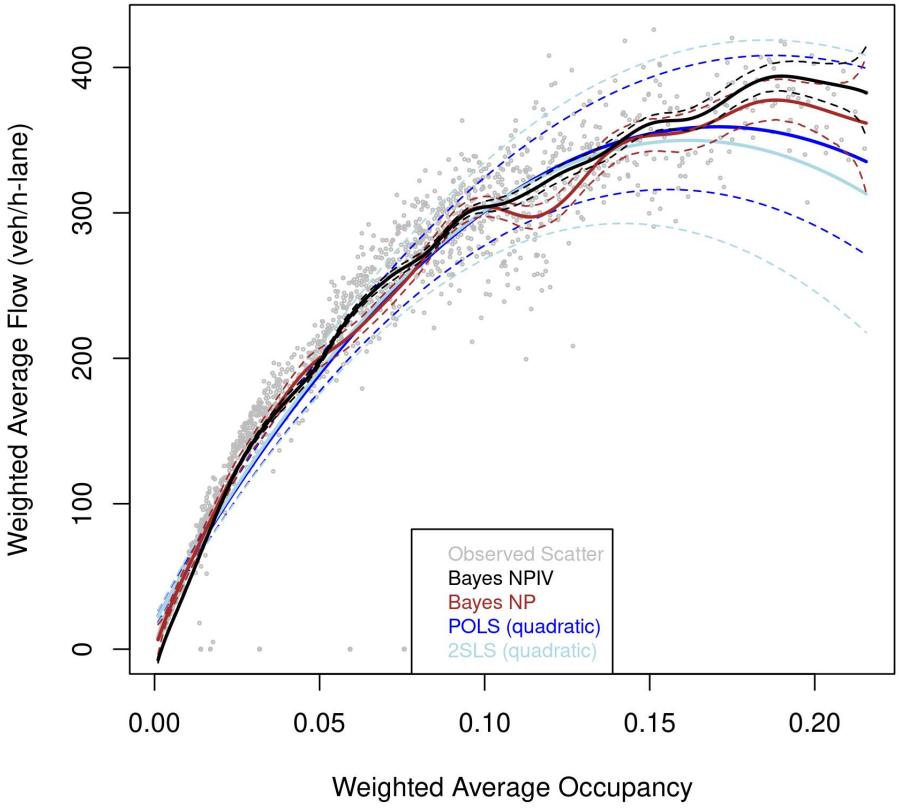


(b) Comparison of different estimators.

Figure C.1: Estimated MFD for Ausburg

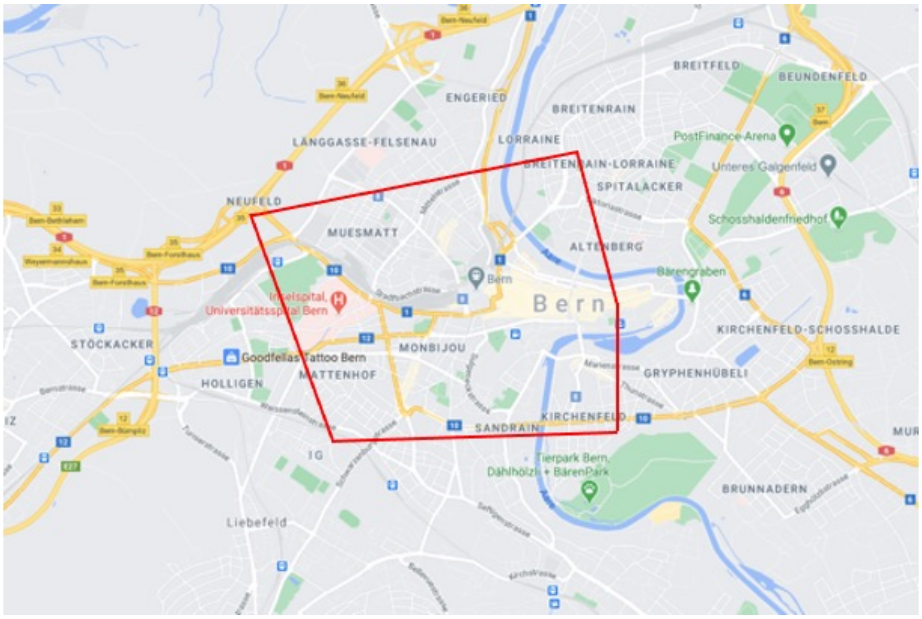


(a) Network exhibit used for the MFD estimation.

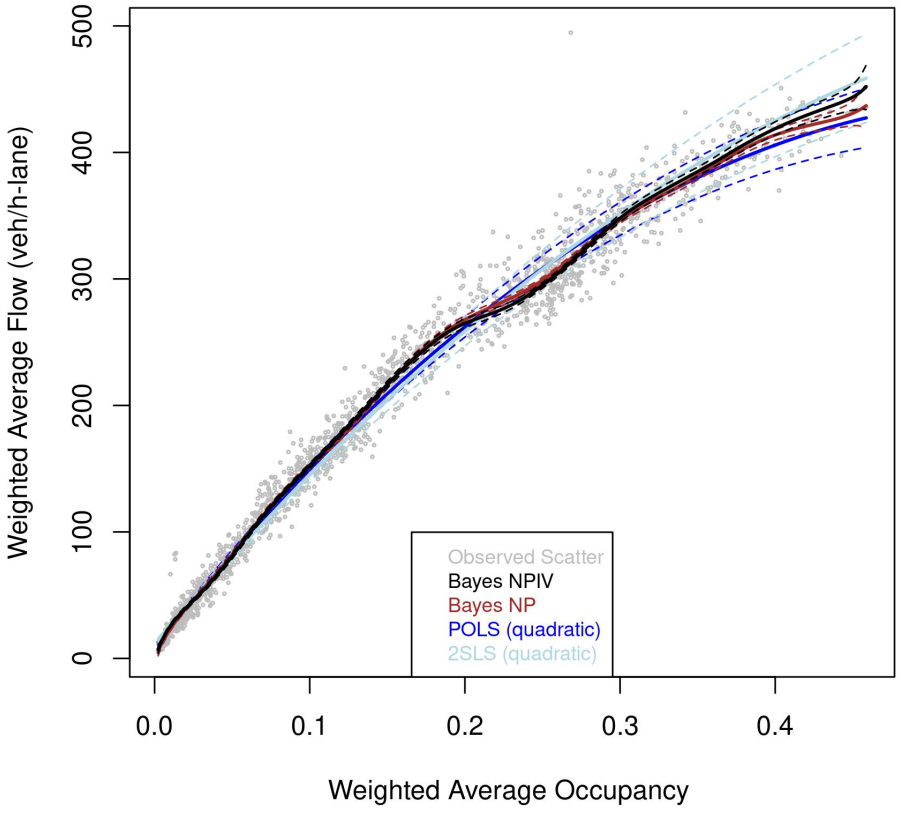


(b) Comparison of different estimators.

Figure C.2: Estimated MFD for Basel

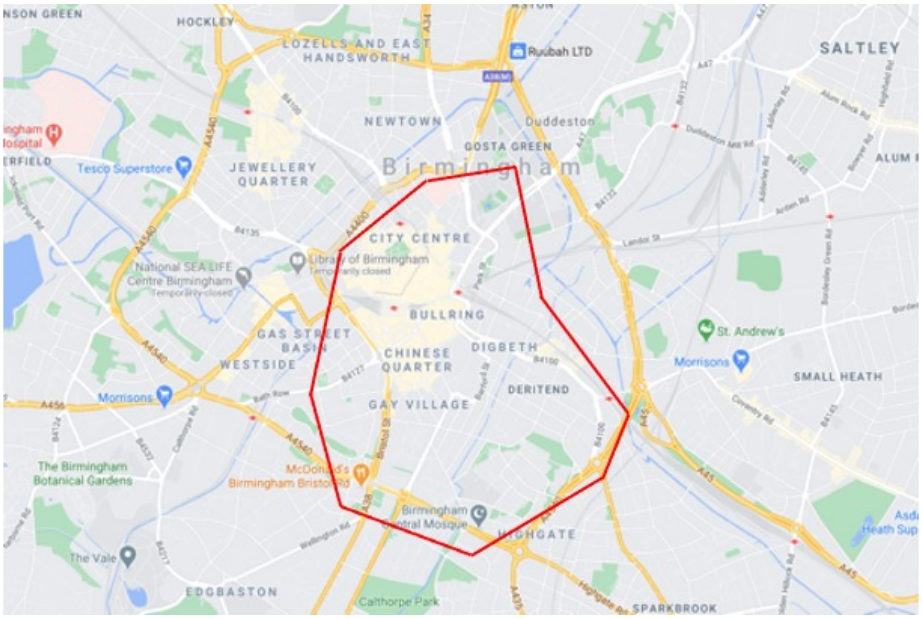


(a) Network exhibit used for the MFD estimation.

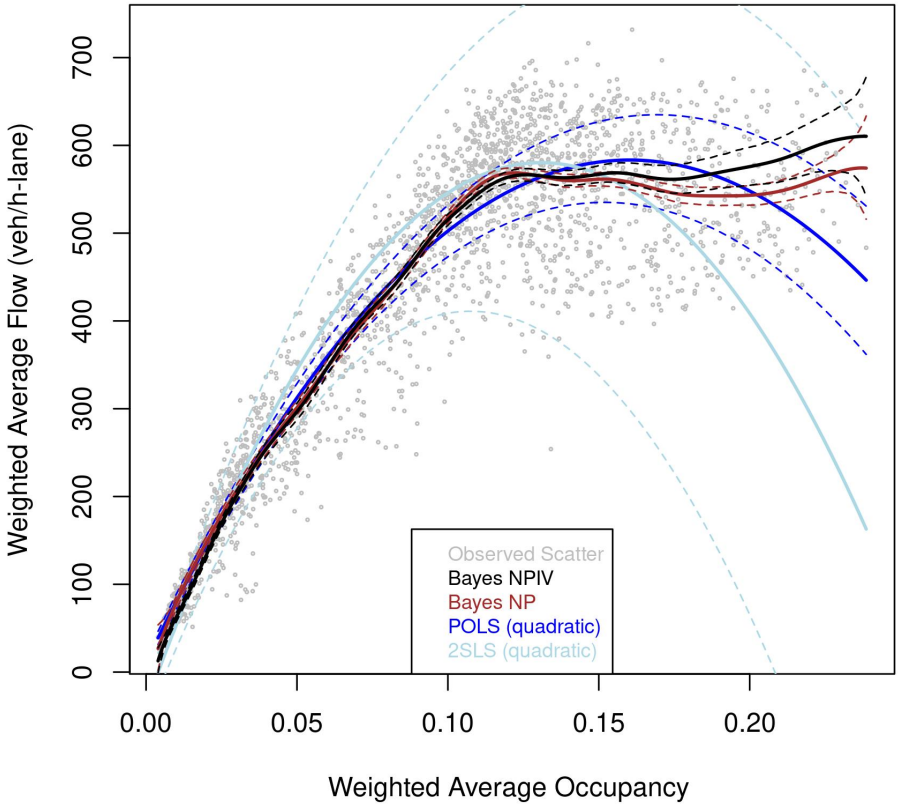


(b) Comparison of different estimators.

Figure C.3: Estimated MFD for Bern

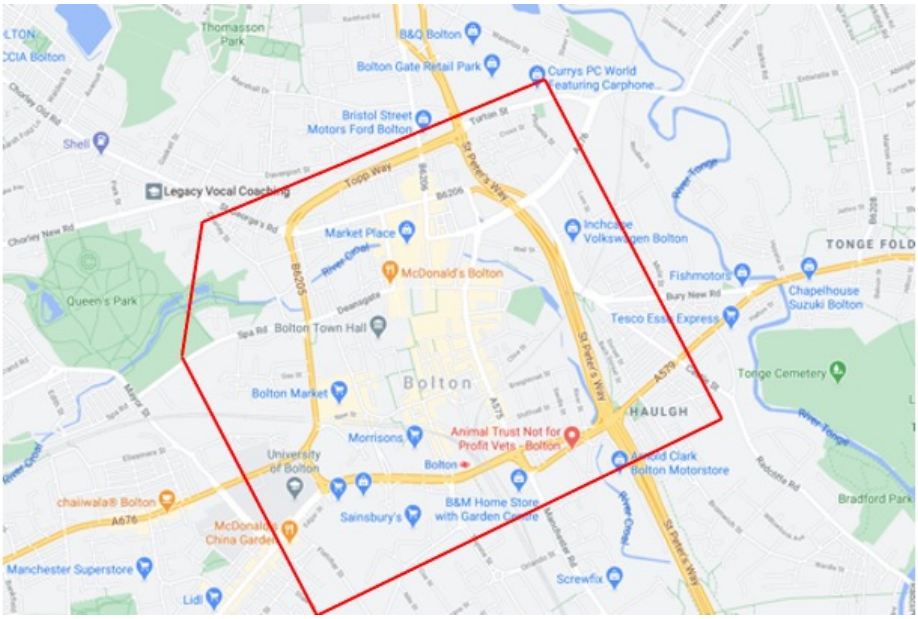


(a) Network exhibit used for the MFD estimation.

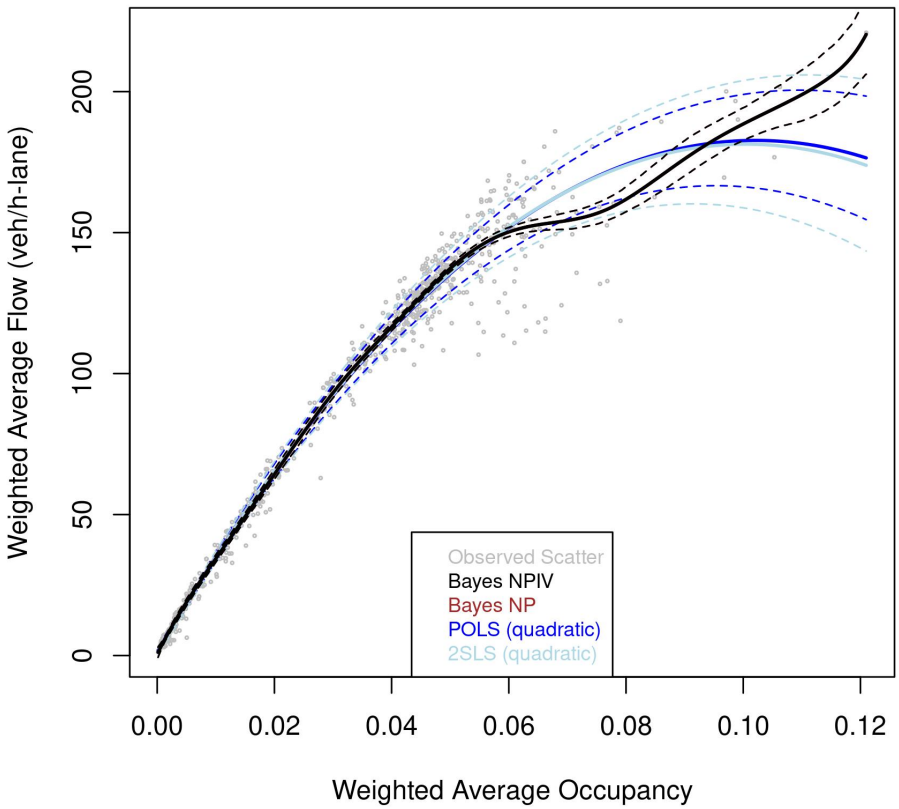


(b) Comparison of different estimators.

Figure C.4: Estimated MFD for Birmingham

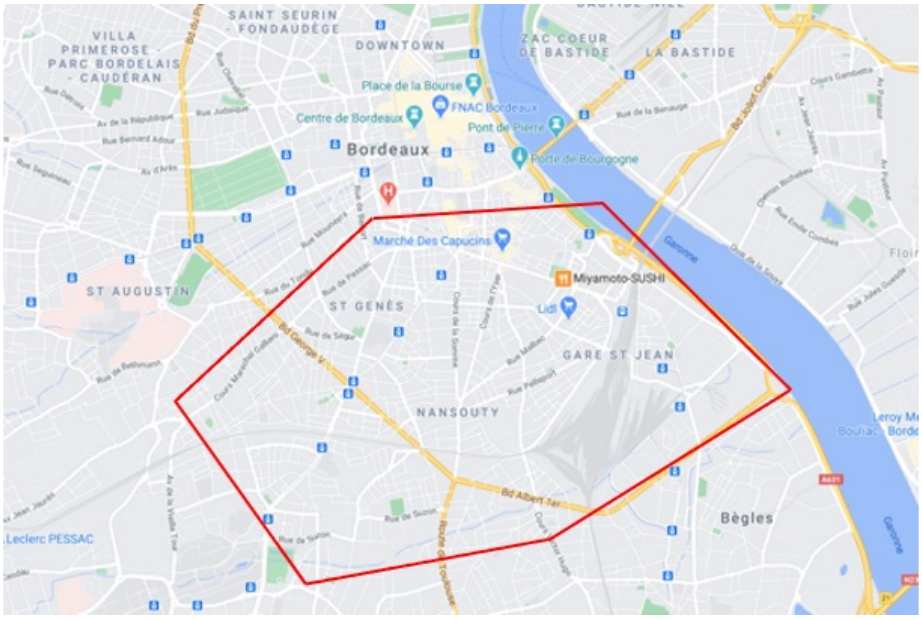


(a) Network exhibit used for the MFD estimation.

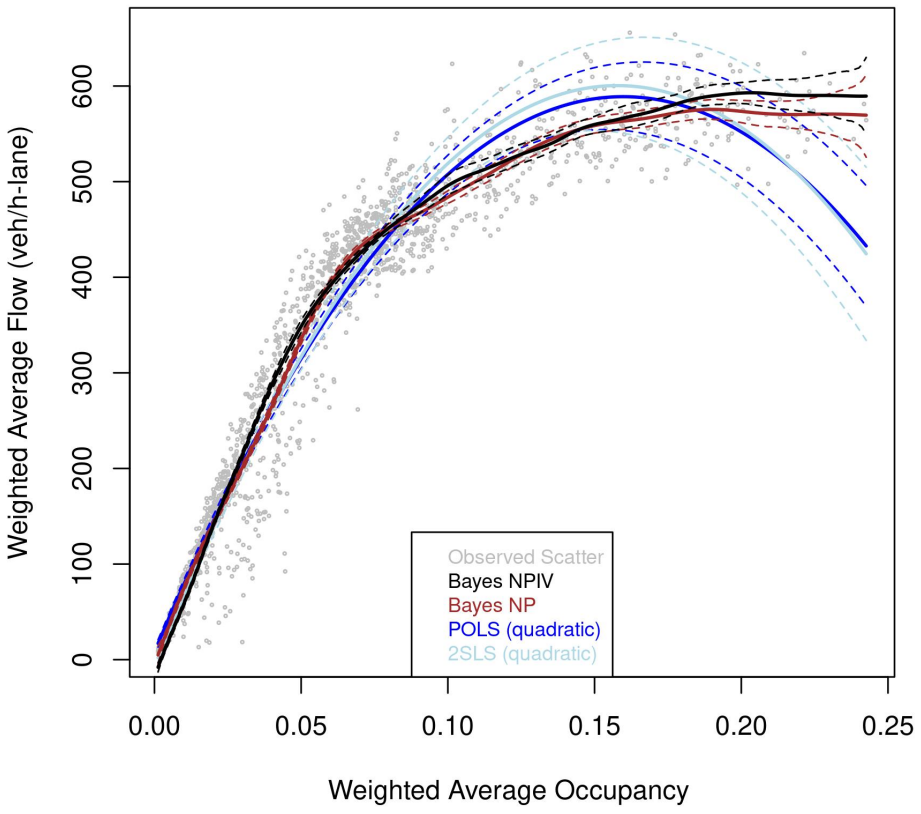


(b) Comparison of different estimators.

Figure C.5: Estimated MFD for Bolton

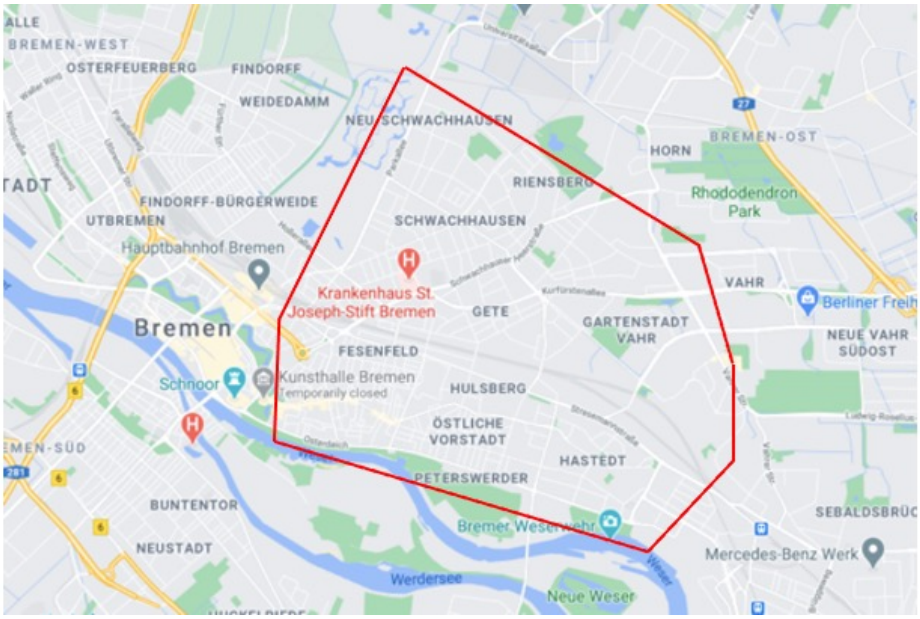


(a) Network exhibit used for the MFD estimation.

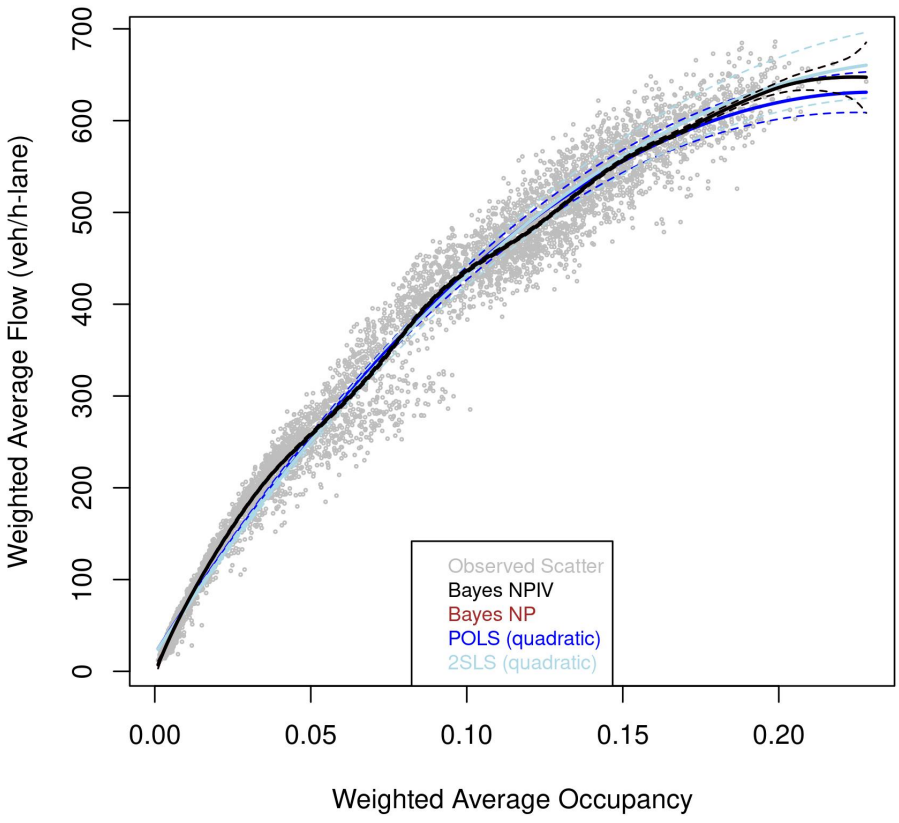


(b) Comparison of different estimators.

Figure C.6: Estimated MFD for Bordeaux

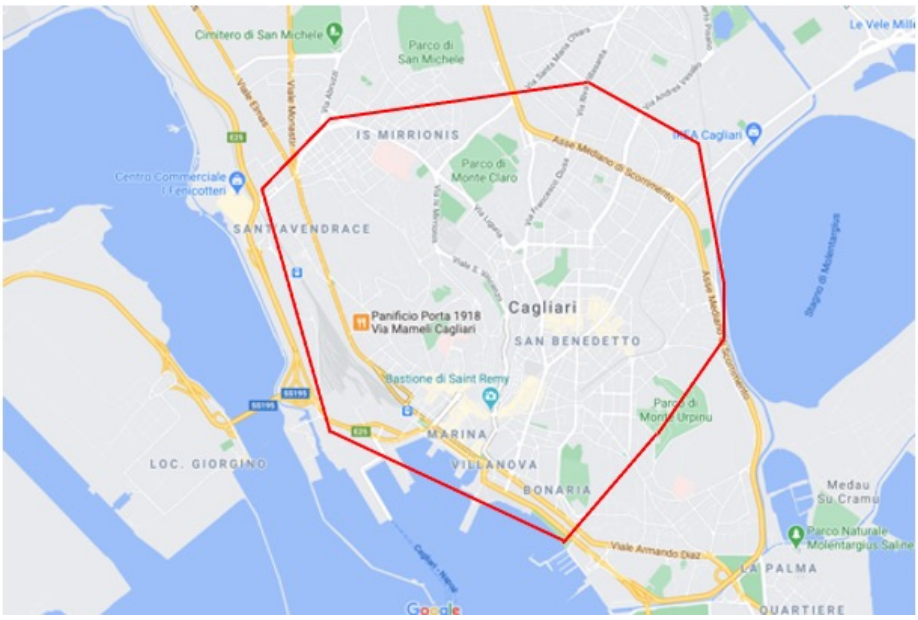


(a) Network exhibit used for the MFD estimation.

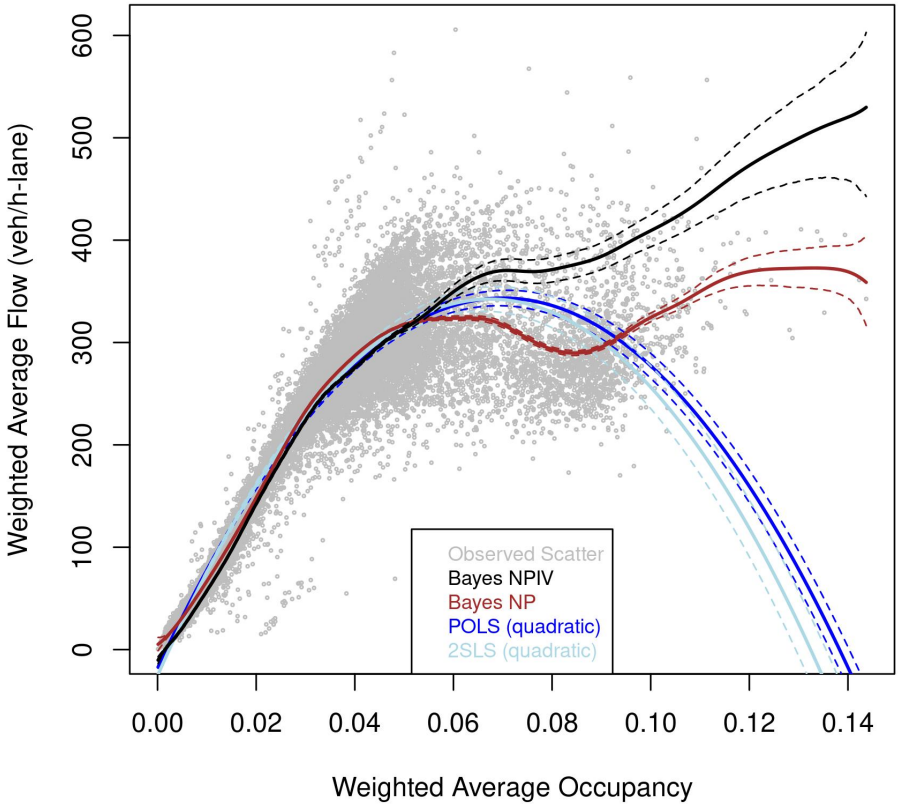


(b) Comparison of different estimators.

Figure C.7: Estimated MFD for Bremen

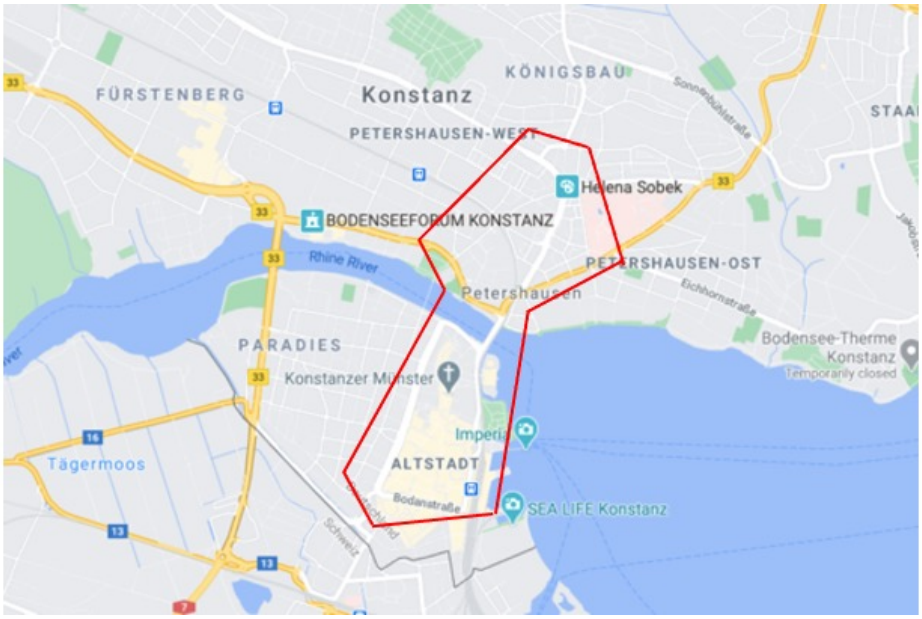


(a) Network exhibit used for the MFD estimation.

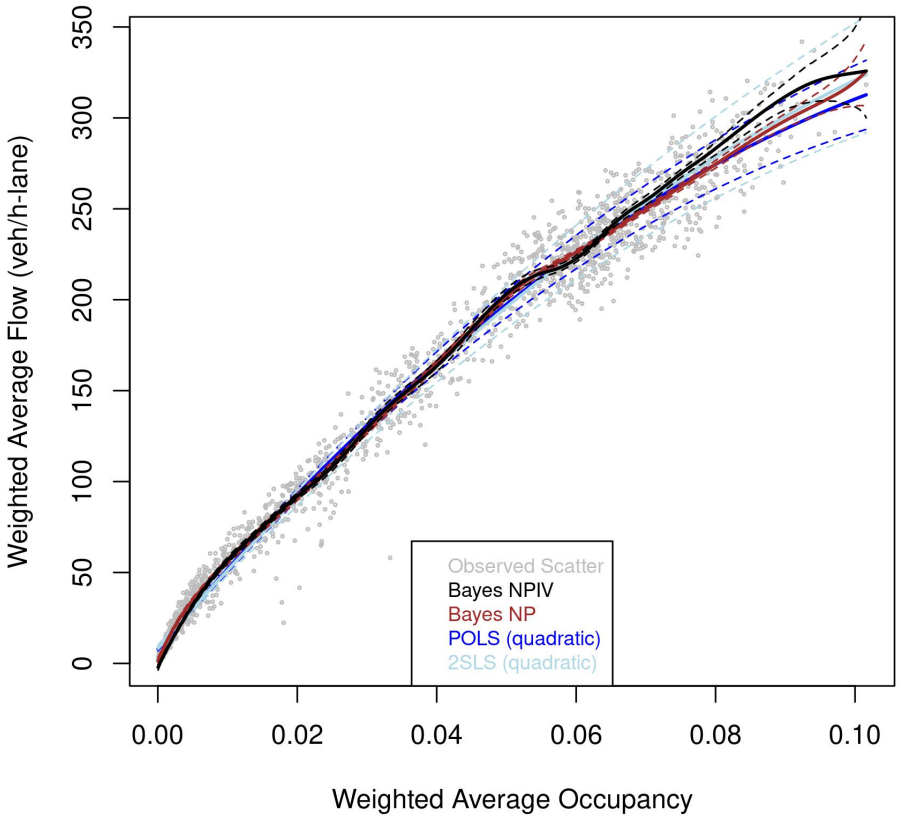


(b) Comparison of different estimators.

Figure C.8: Estimated MFD for Cagliari



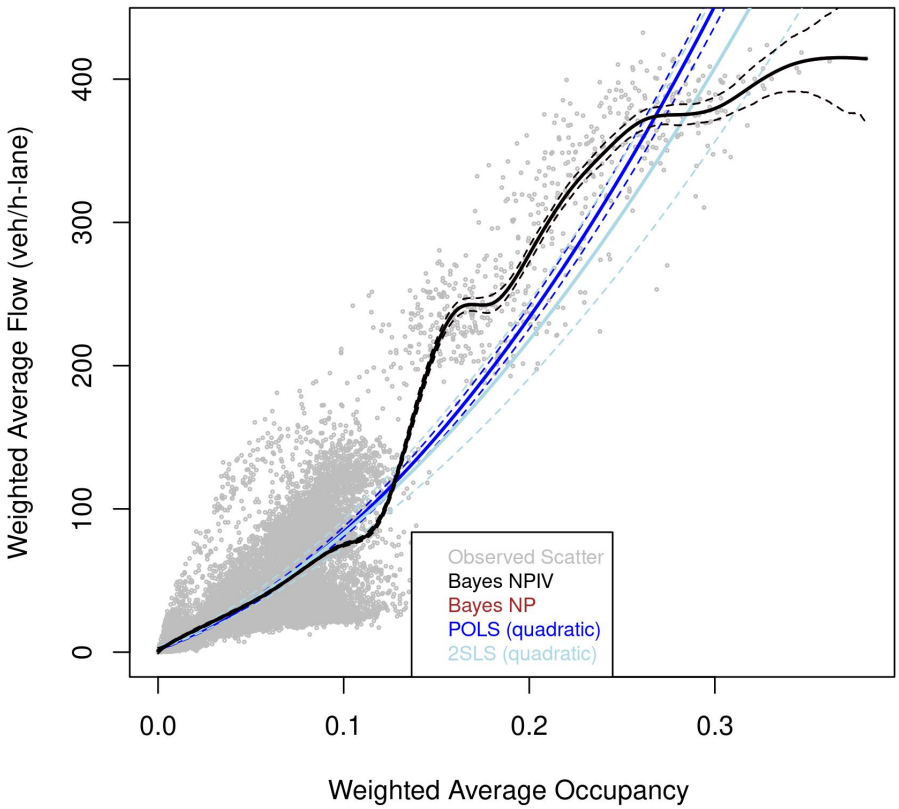
(a) Network exhibit used for the MFD estimation.



(b) Comparison of different estimators.

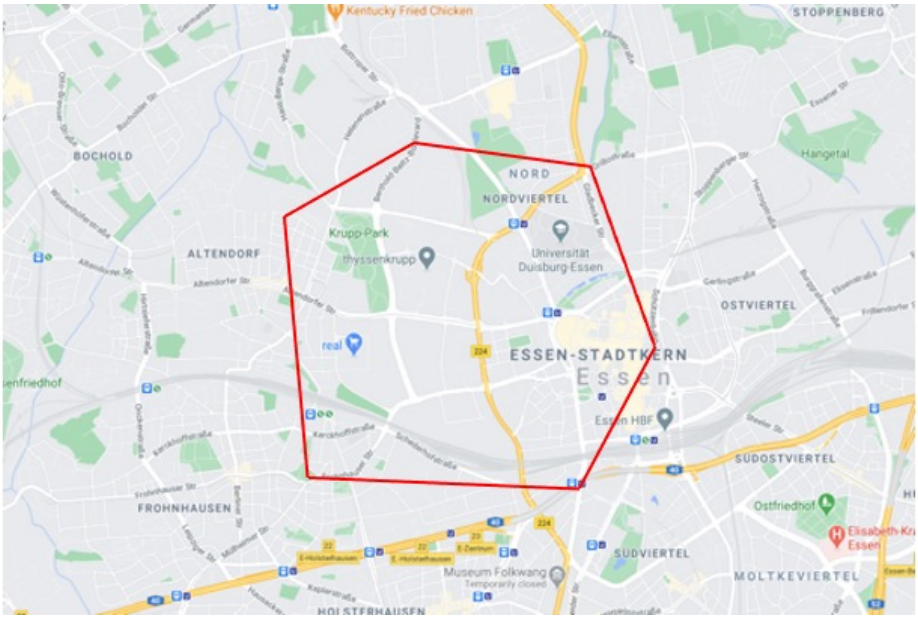
Figure C.9: Estimated MFD for Constance

(a) Network exhibit used for the MFD estimation.

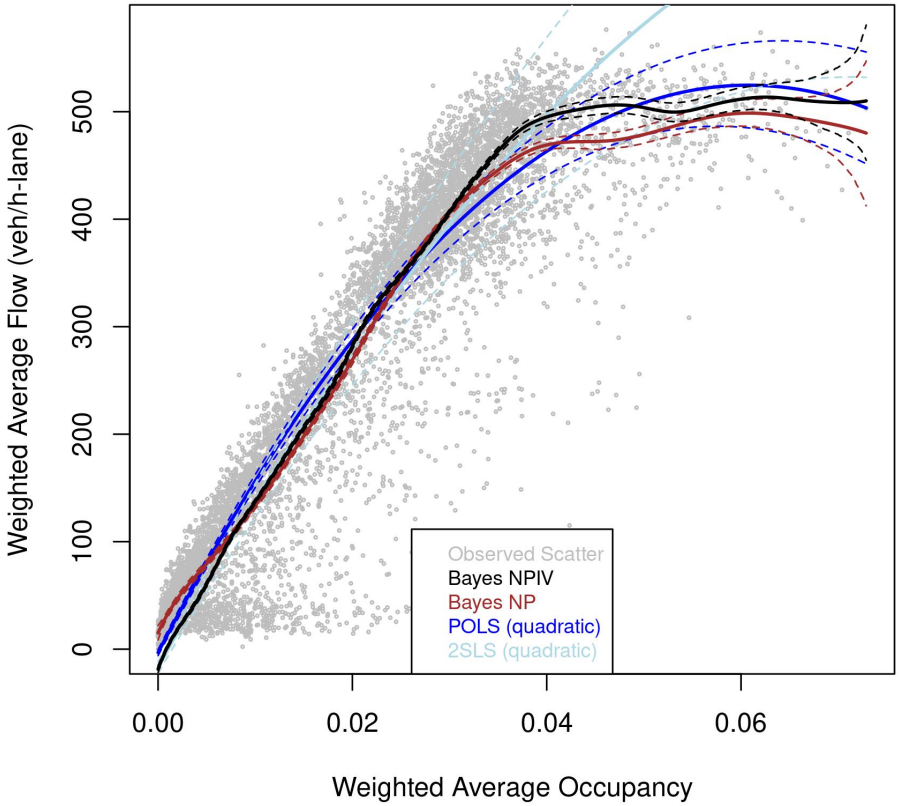


(b) Comparison of different estimators.

Figure C.10: Estimated MFD for Darmstadt



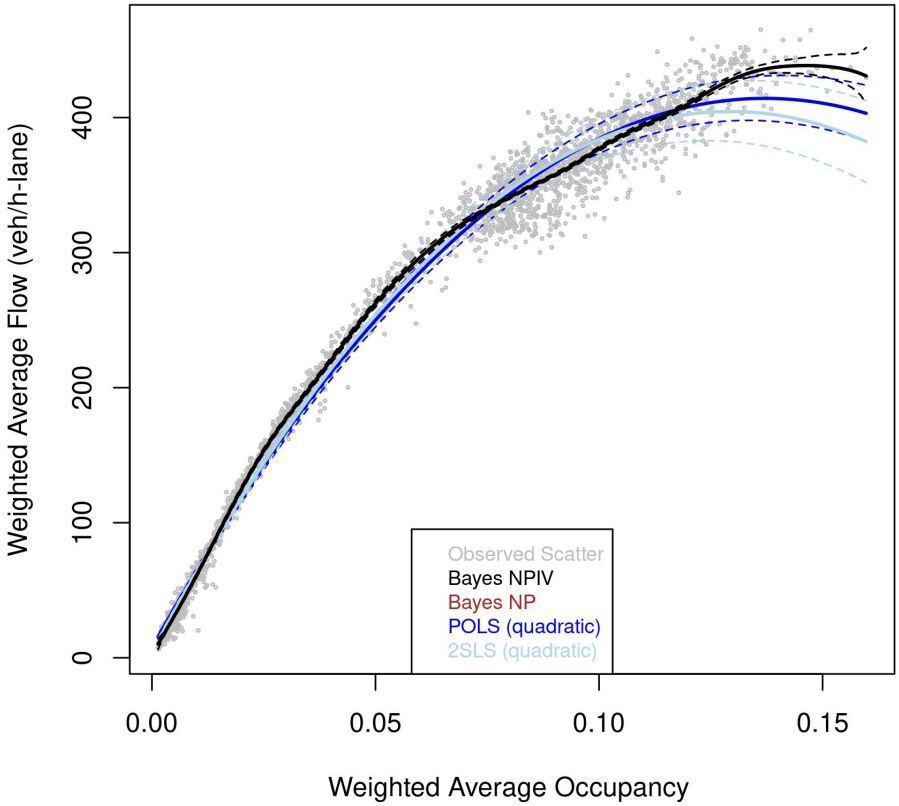
(a) Network exhibit used for the MFD estimation.



(b) Comparison of different estimators.

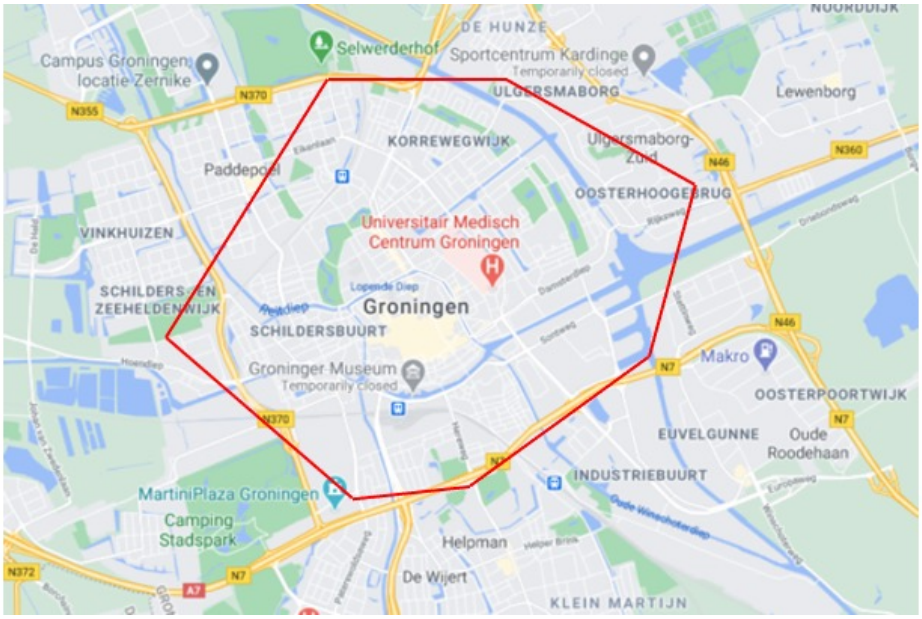
Figure C.11: Estimated MFD for Essen

(a) Network exhibit used for the MFD estimation.

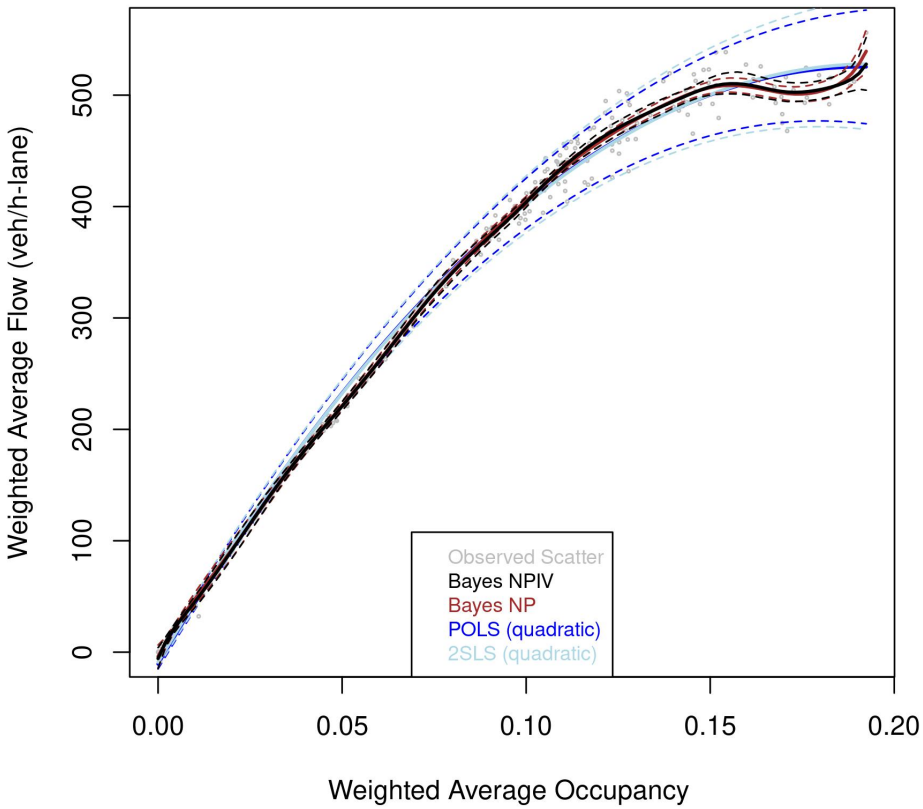


(b) Comparison of different estimators.

Figure C.12: Estimated MFD for Graz

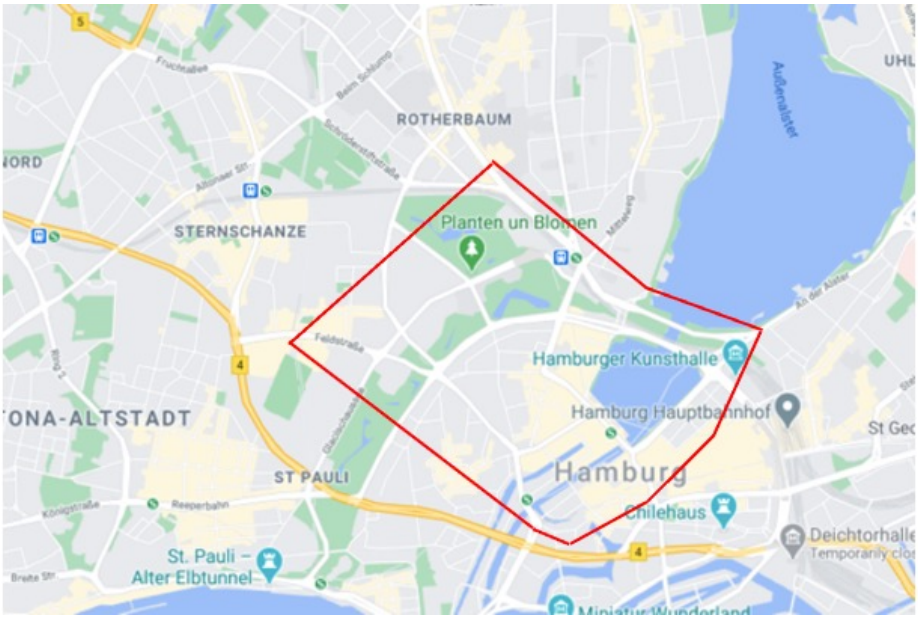


(a) Network exhibit used for the MFD estimation.

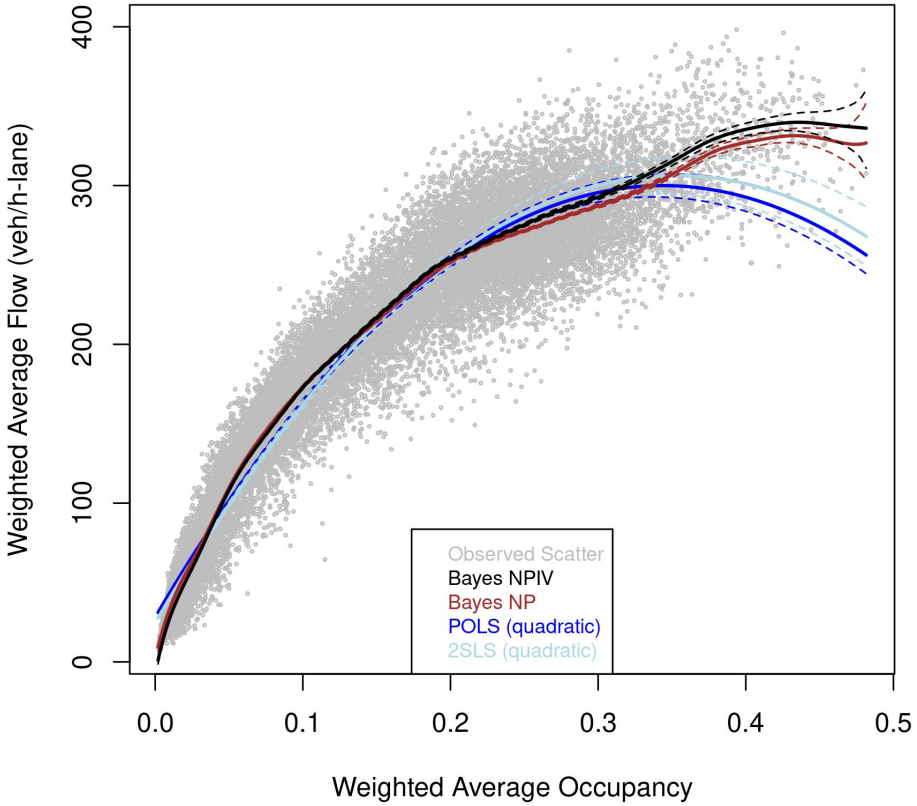


(b) Comparison of different estimators.

Figure C.13: Estimated MFD for Groningen



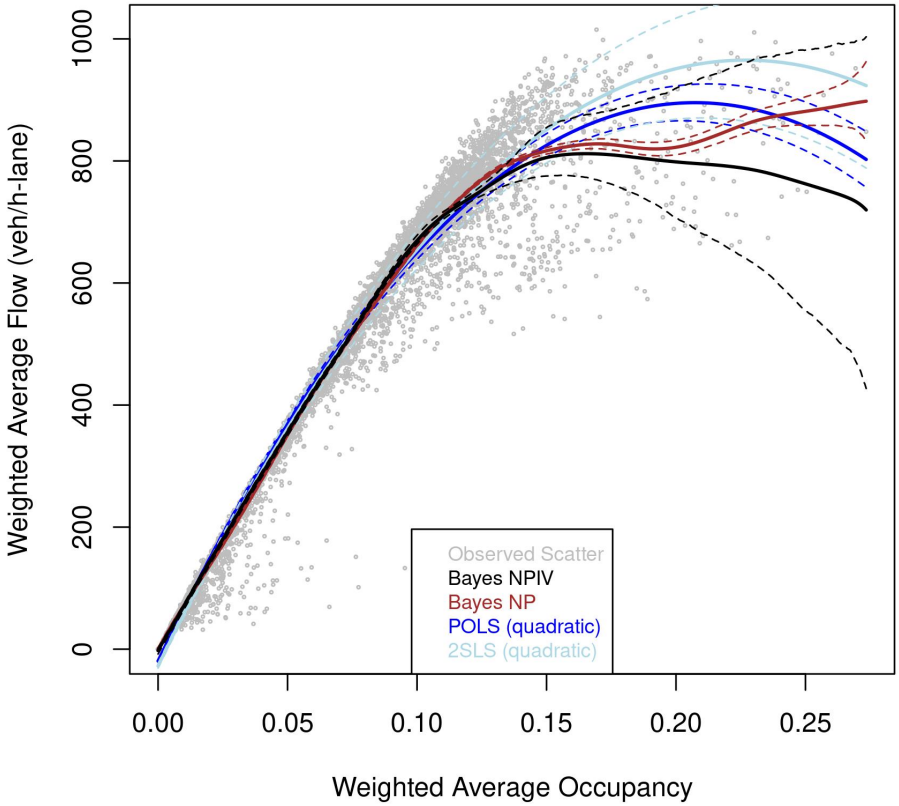
(a) Network exhibit used for the MFD estimation.



(b) Comparison of different estimators.

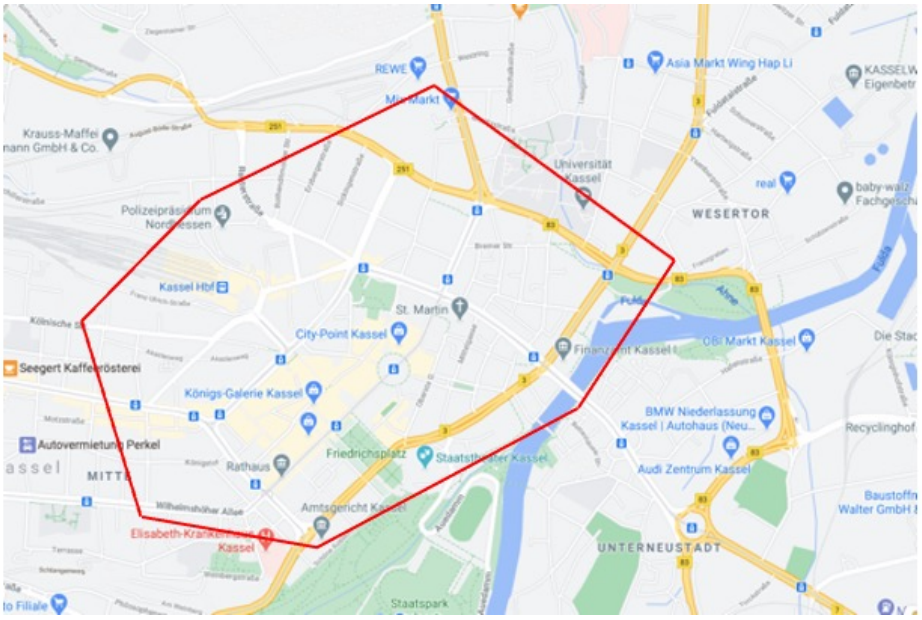
Figure C.14: Estimated MFD for Hamburg

(a) Network exhibit used for the MFD estimation.

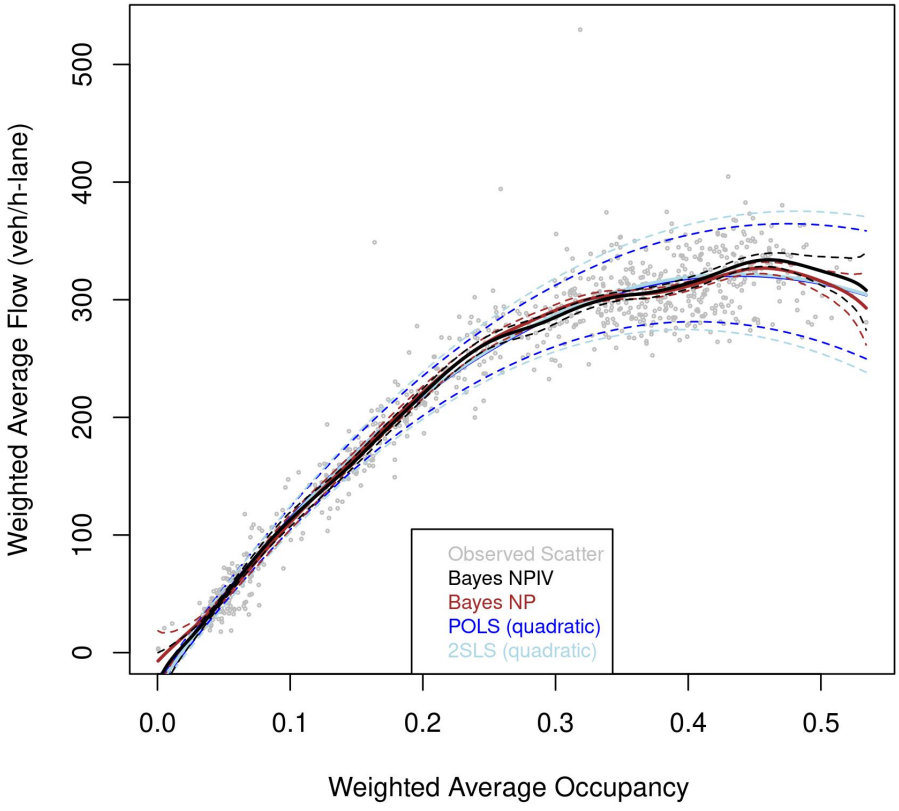


(b) Comparison of different estimators.

Figure C.15: Estimated MFD for Innsbruck

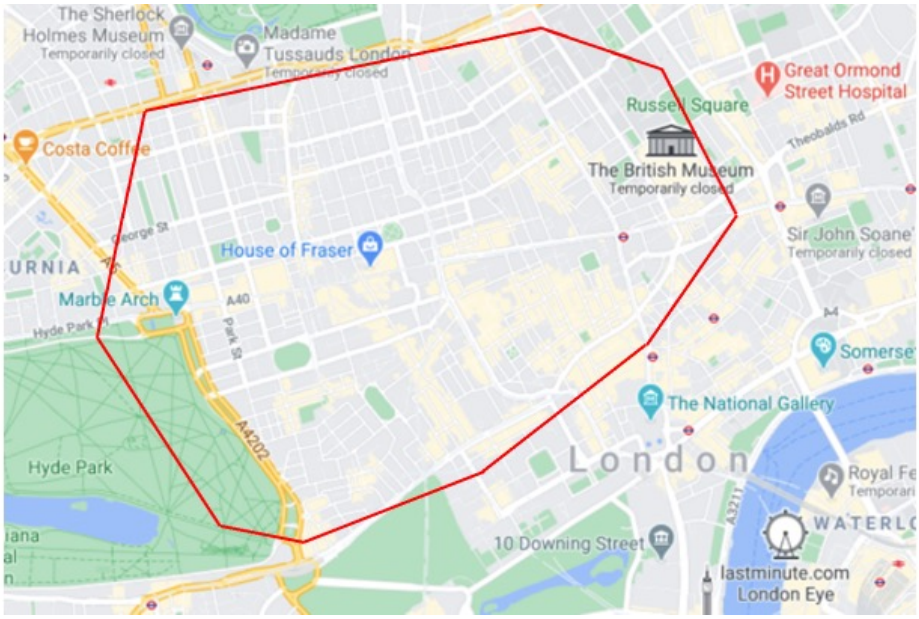


(a) Network exhibit used for the MFD estimation.

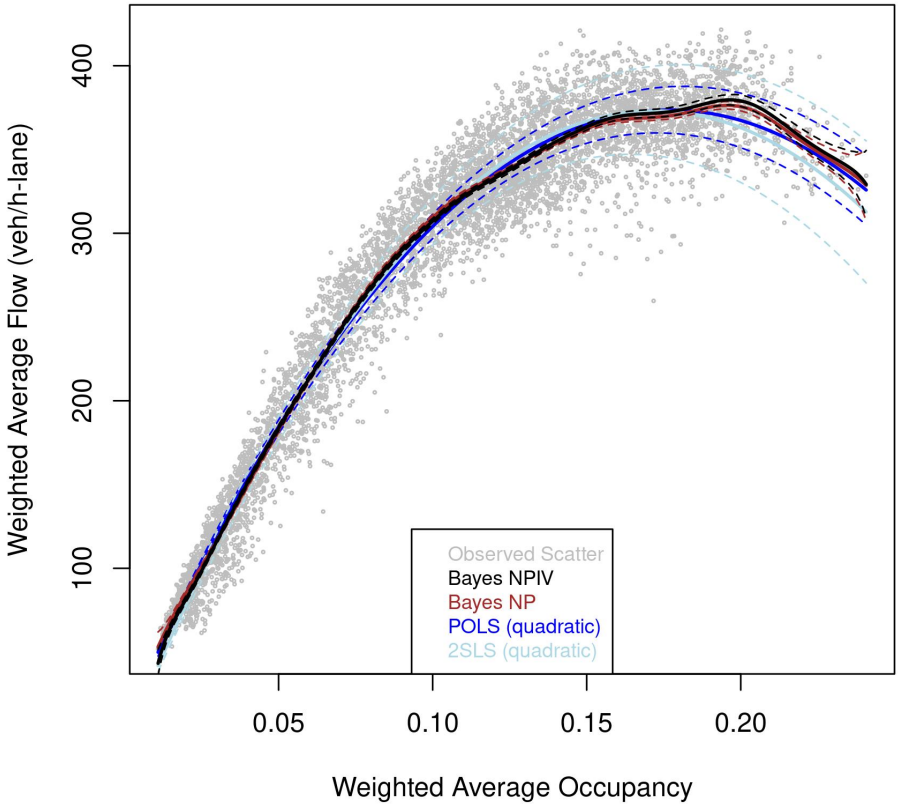


(b) Comparison of different estimators.

Figure C.16: Estimated MFD for Kassel

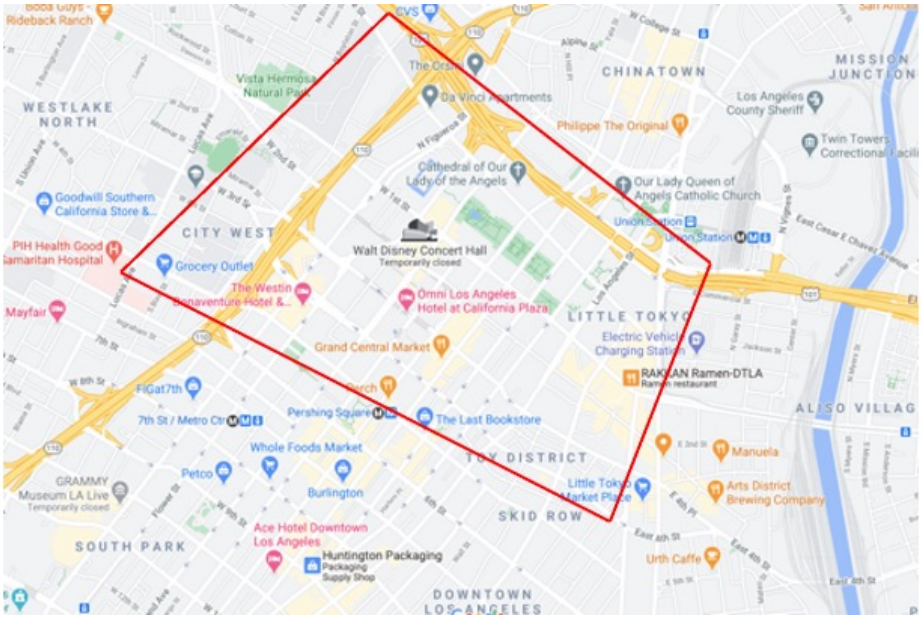


(a) Network exhibit used for the MFD estimation.

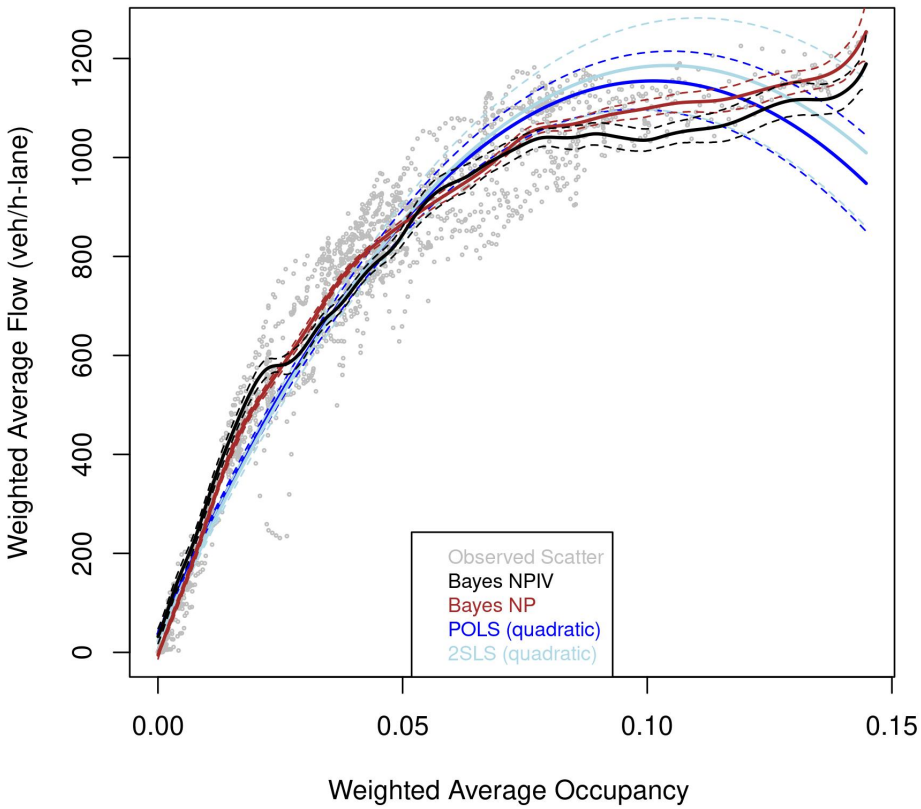


(b) Comparison of different estimators.

Figure C.17: Estimated MFD for London

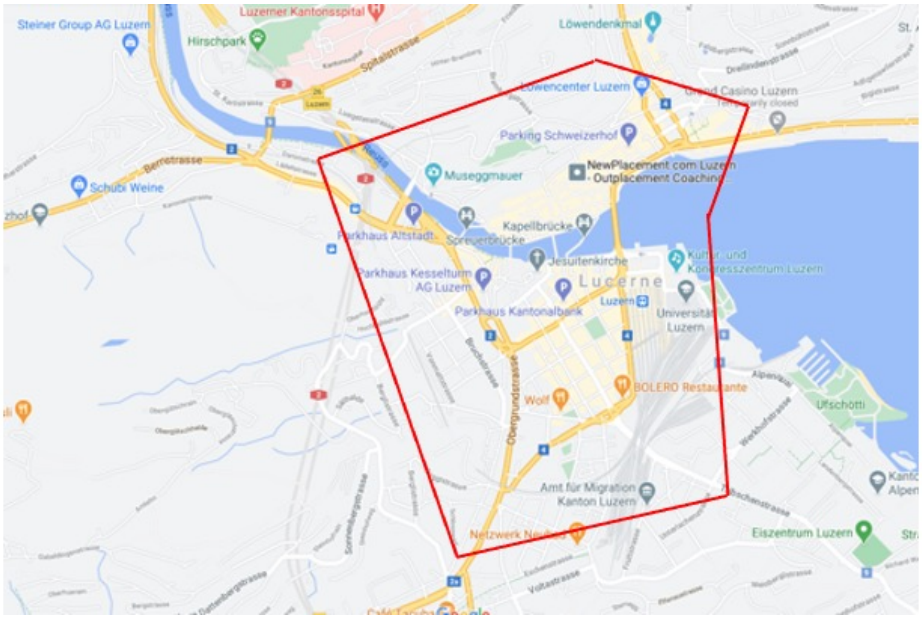


(a) Network exhibit used for the MFD estimation.

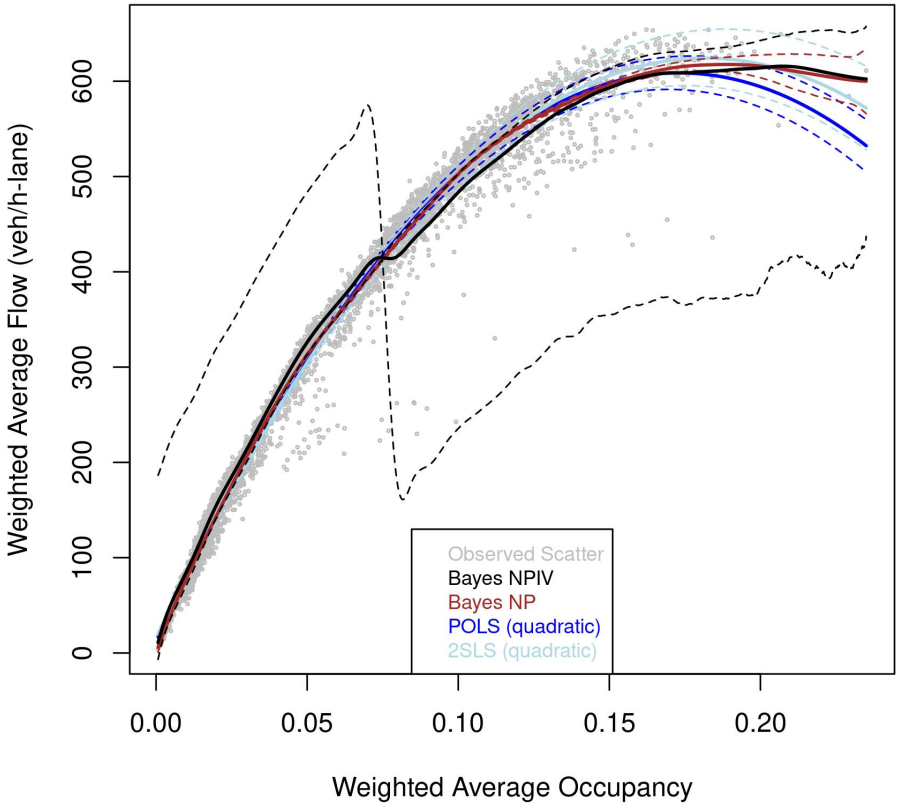


(b) Comparison of different estimators.

Figure C.18: Estimated MFD for Los Angeles

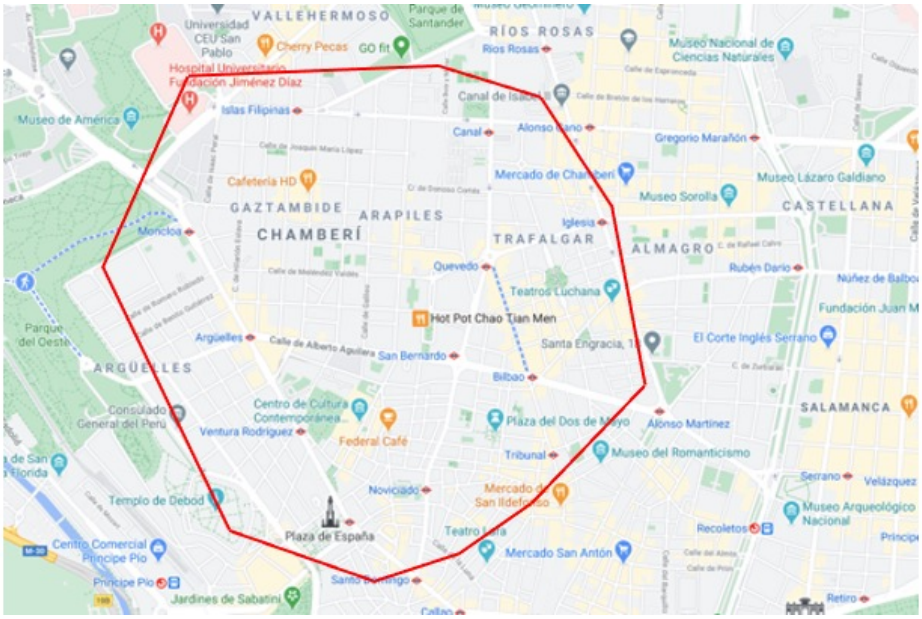


(a) Network exhibit used for the MFD estimation.

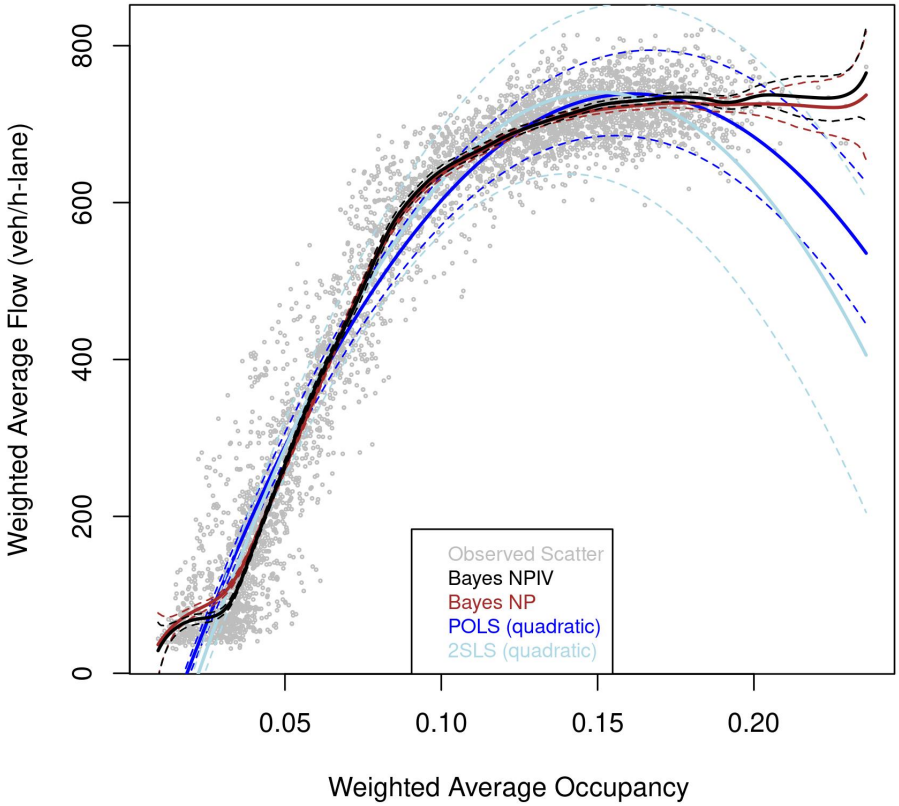


(b) Comparison of different estimators.

Figure C.19: Estimated MFD for Luzern



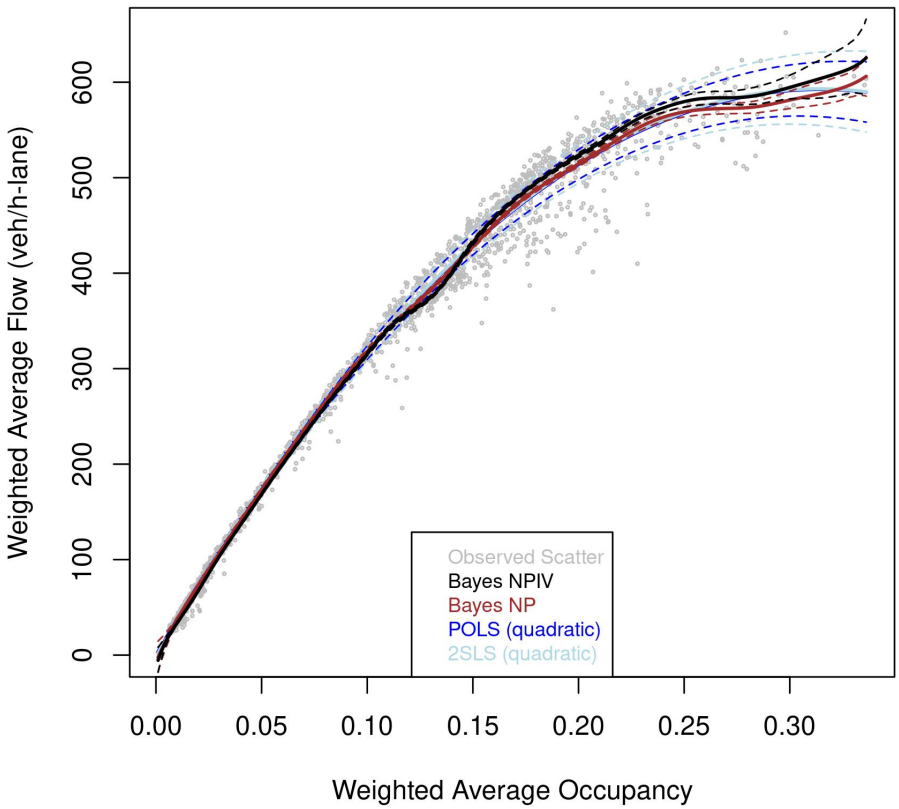
(a) Network exhibit used for the MFD estimation.



(b) Comparison of different estimators.

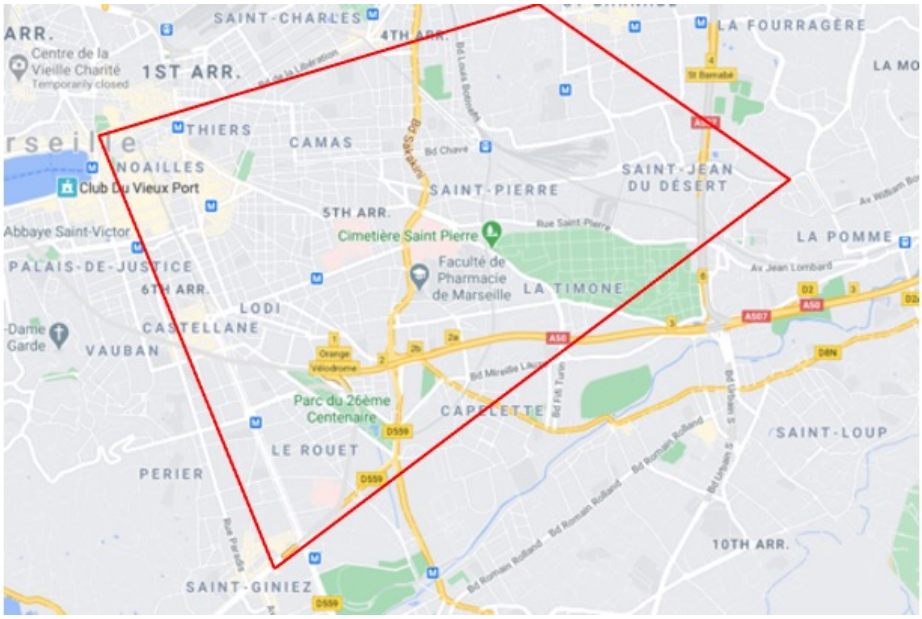
Figure C.20: Estimated MFD for Madrid

(a) Network exhibit used for the MFD estimation.

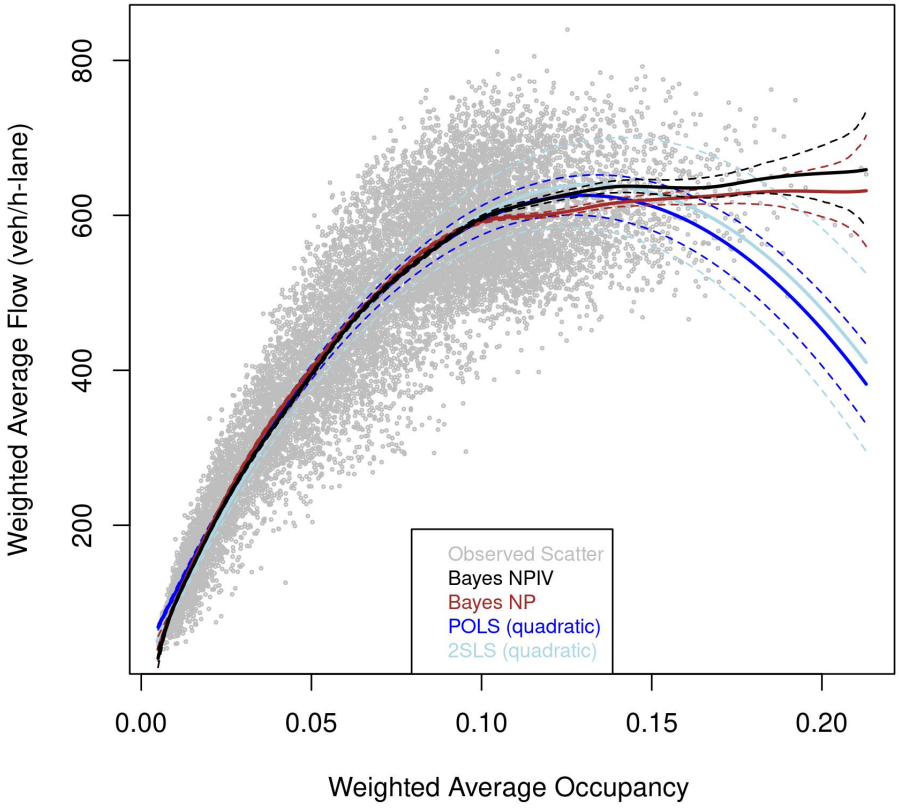


(b) Comparison of different estimators.

Figure C.21: Estimated MFD for Manchester

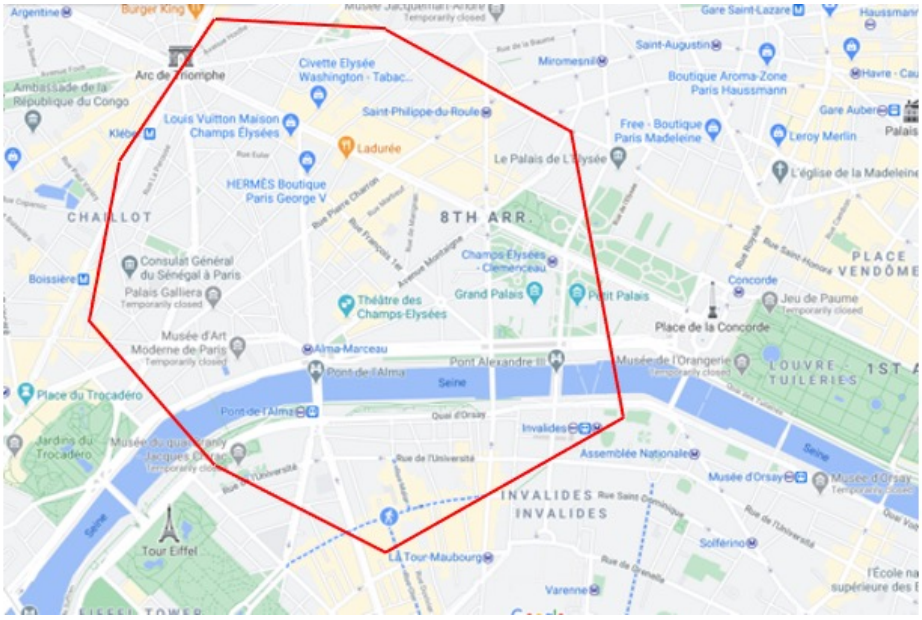


(a) Network exhibit used for the MFD estimation.

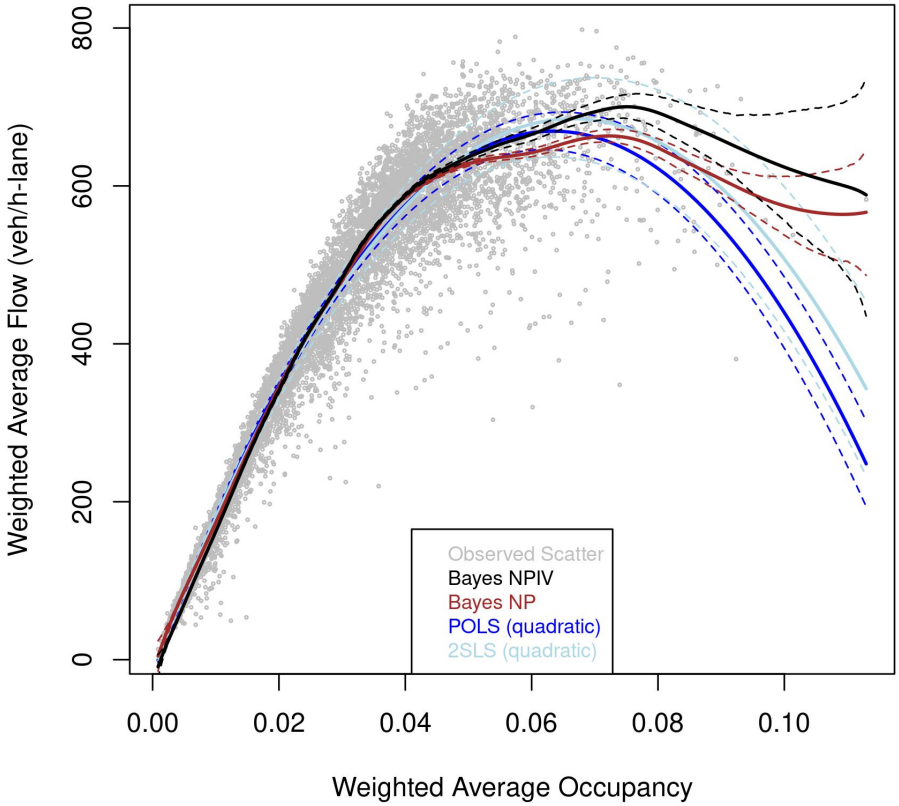


(b) Comparison of different estimators.

Figure C.22: Estimated MFD for Marseille

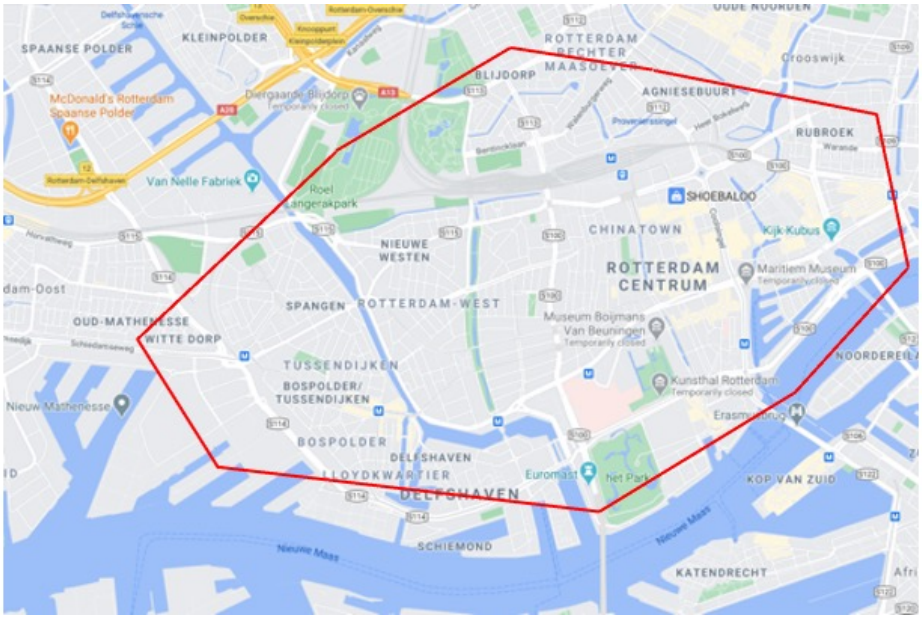


(a) Network exhibit used for the MFD estimation.

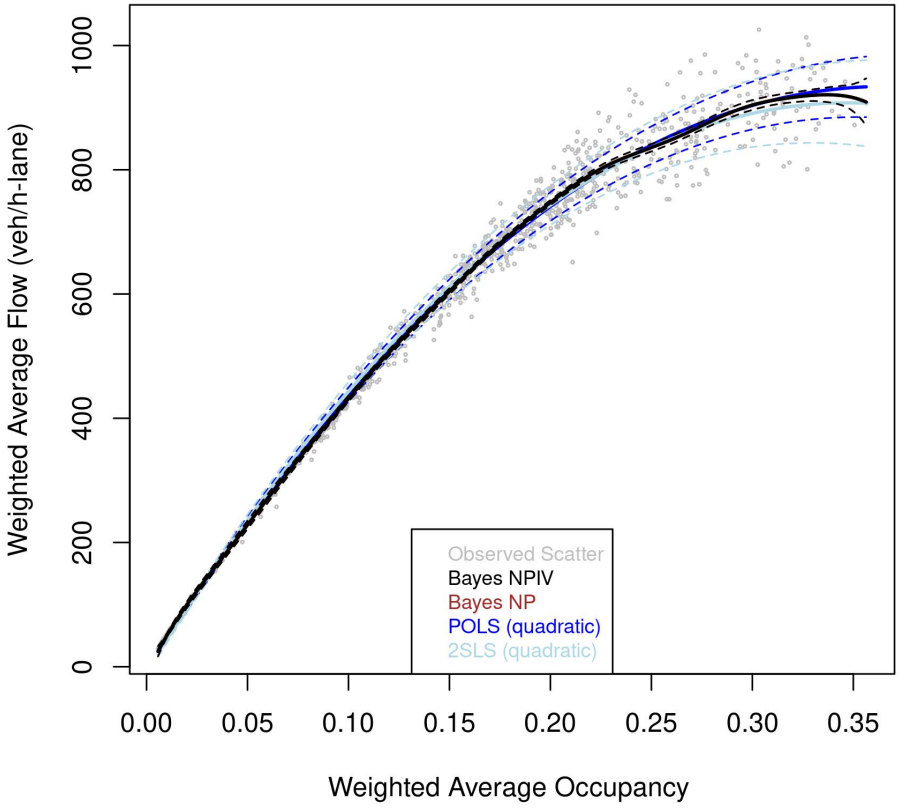


(b) Comparison of different estimators.

Figure C.23: Estimated MFD for Paris

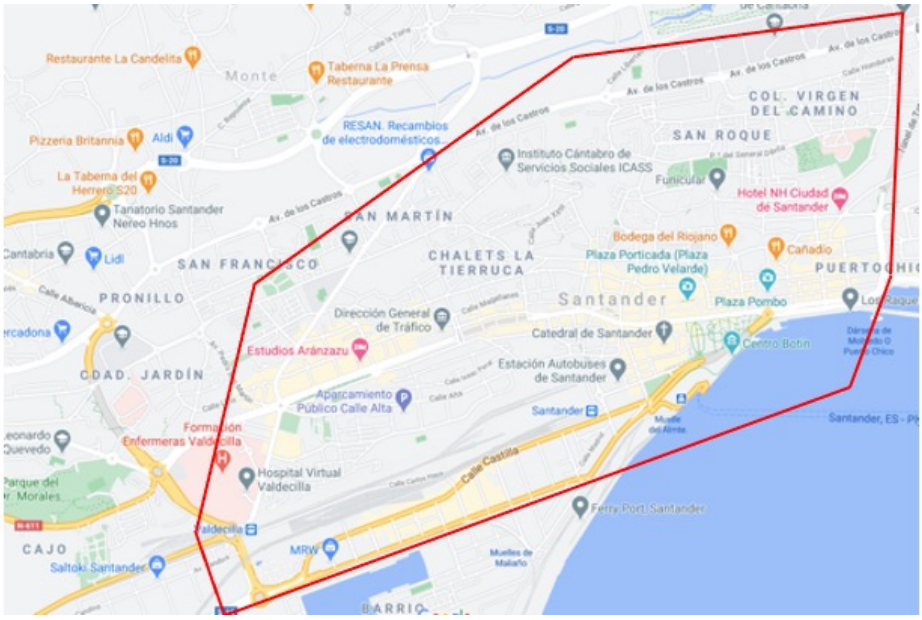


(a) Network exhibit used for the MFD estimation.

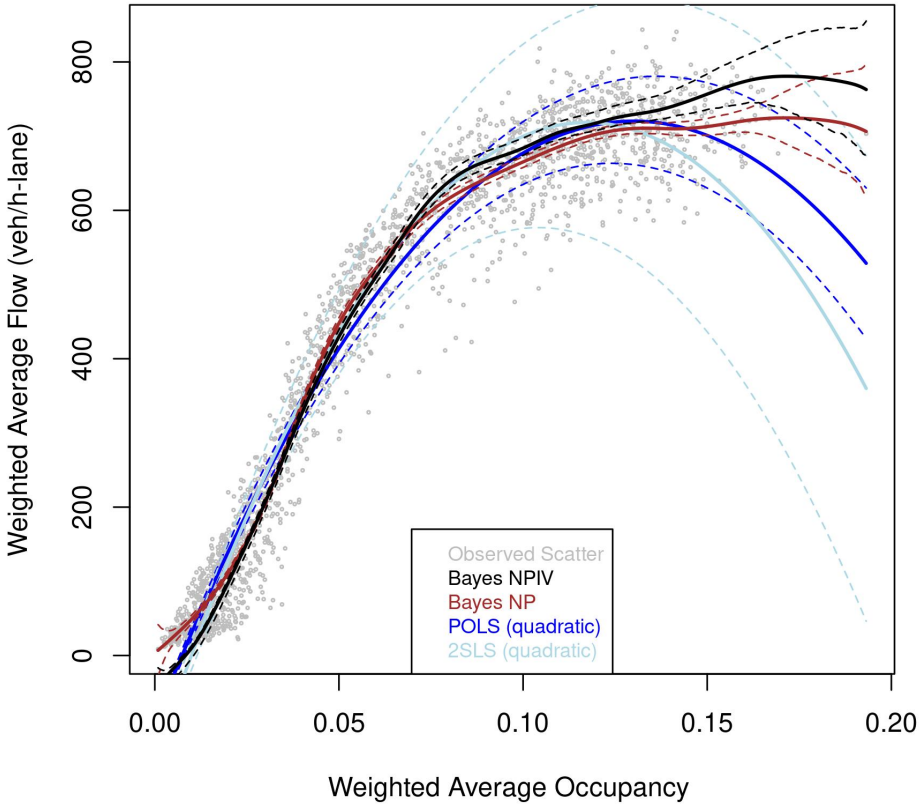


(b) Comparison of different estimators.

Figure C.24: Estimated MFD for Rotterdam

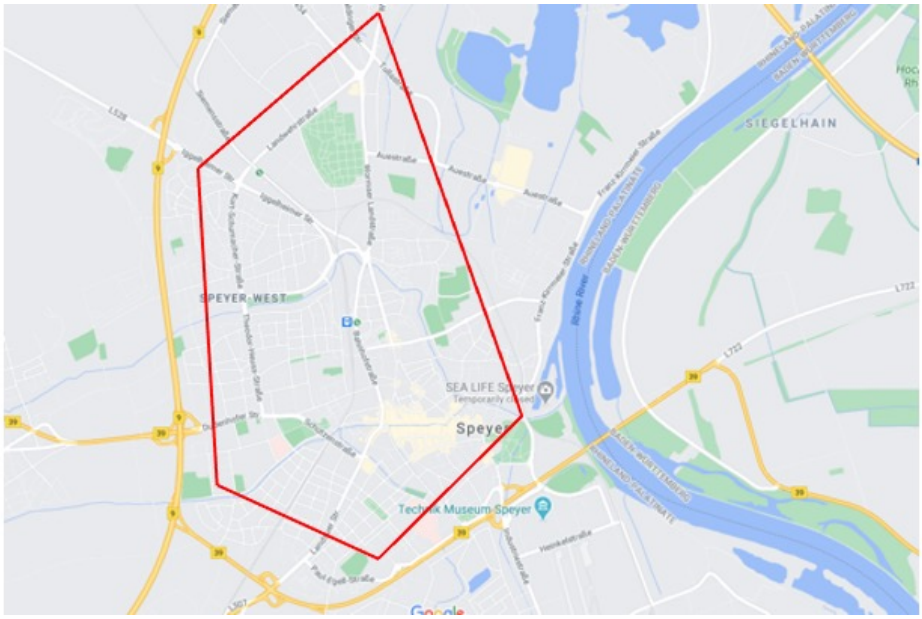


(a) Network exhibit used for the MFD estimation.

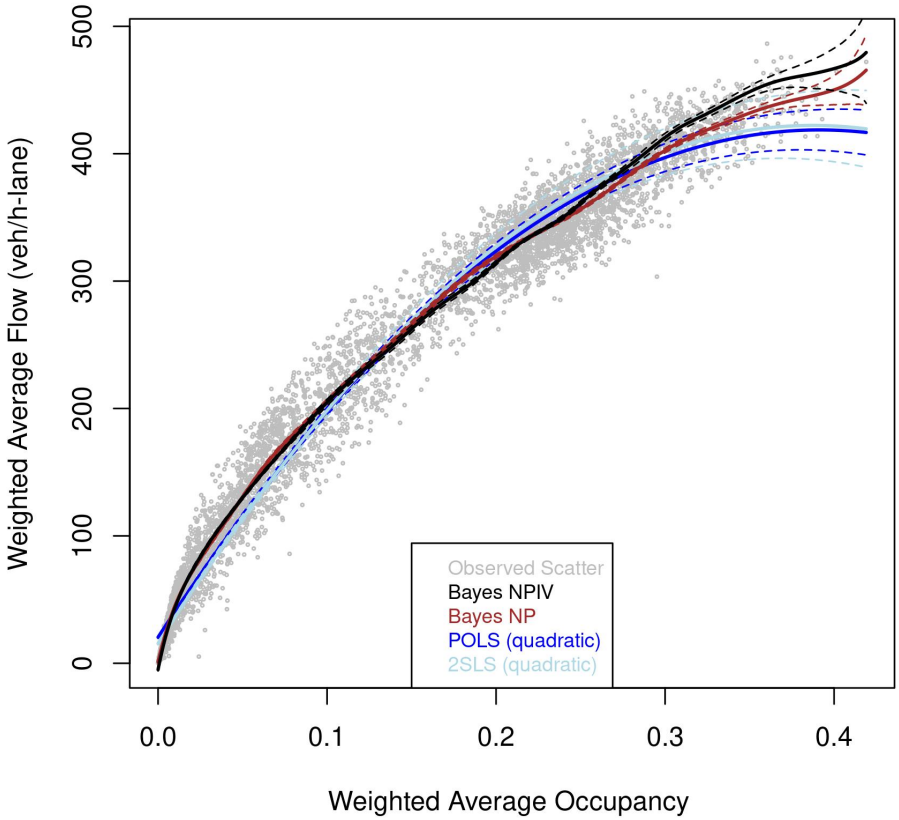


(b) Comparison of different estimators.

Figure C.25: Estimated MFD for Santander

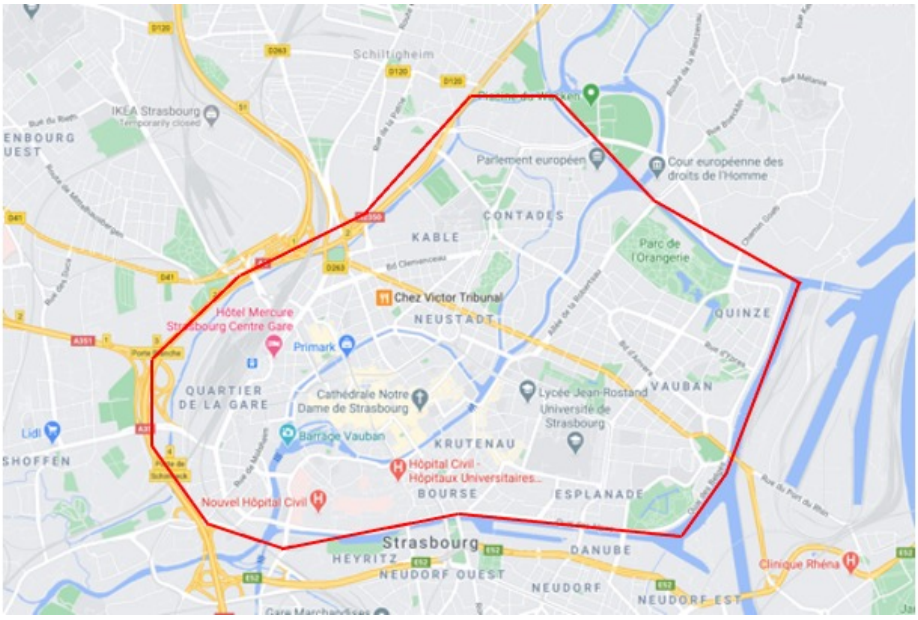


(a) Network exhibit used for the MFD estimation.

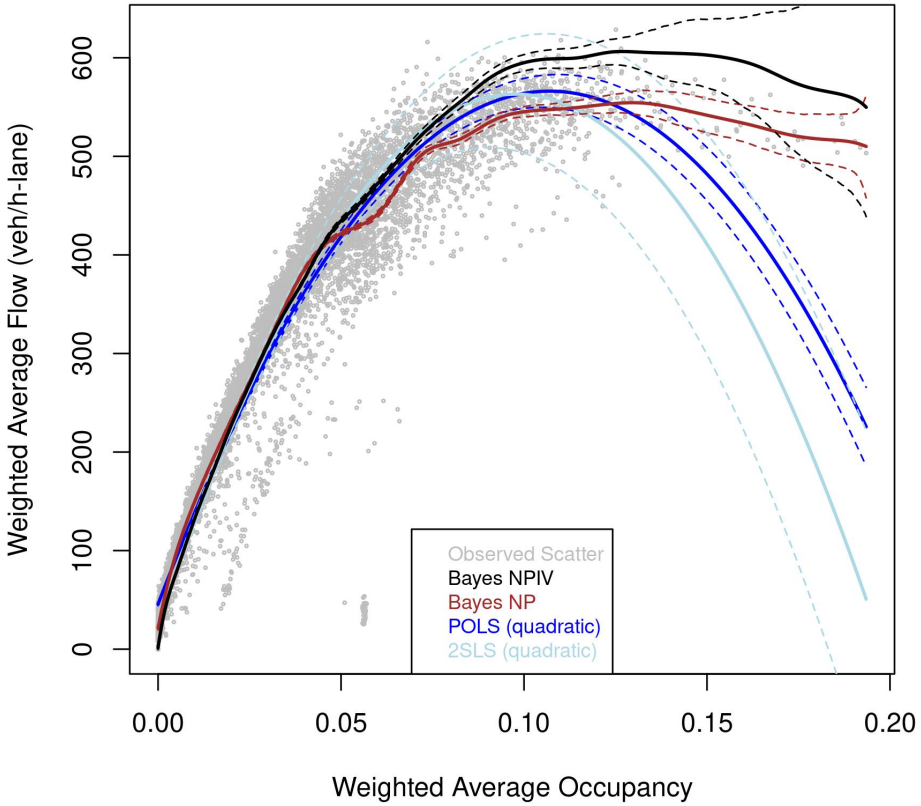


(b) Comparison of different estimators.

Figure C.26: Estimated MFD for Speyer



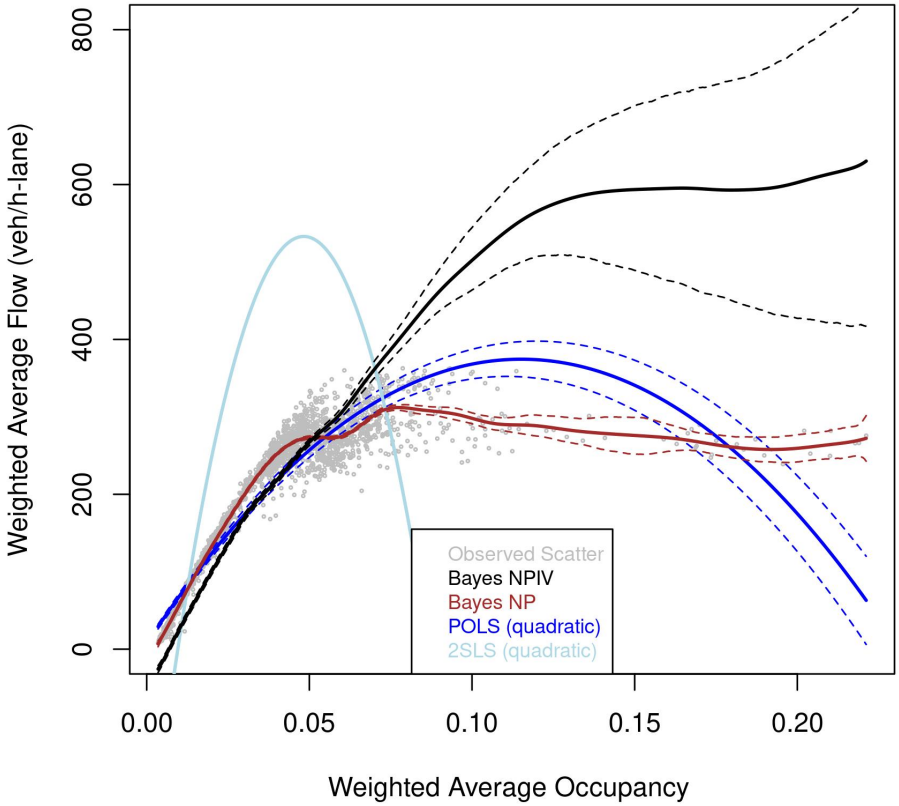
(a) Network exhibit used for the MFD estimation.



(b) Comparison of different estimators.

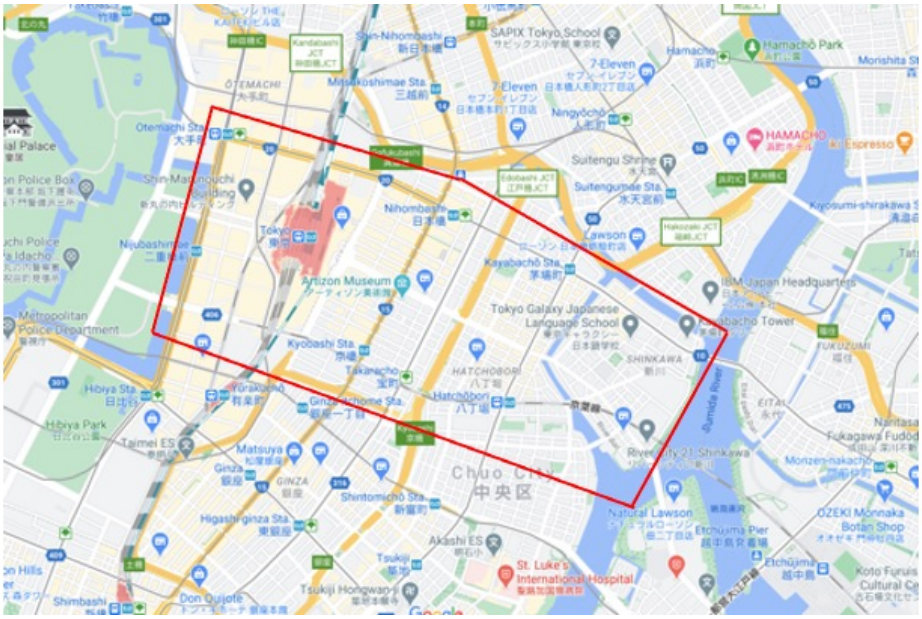
Figure C.27: Estimated MFD for Strasbourg

(a) Network exhibit used for the MFD estimation.

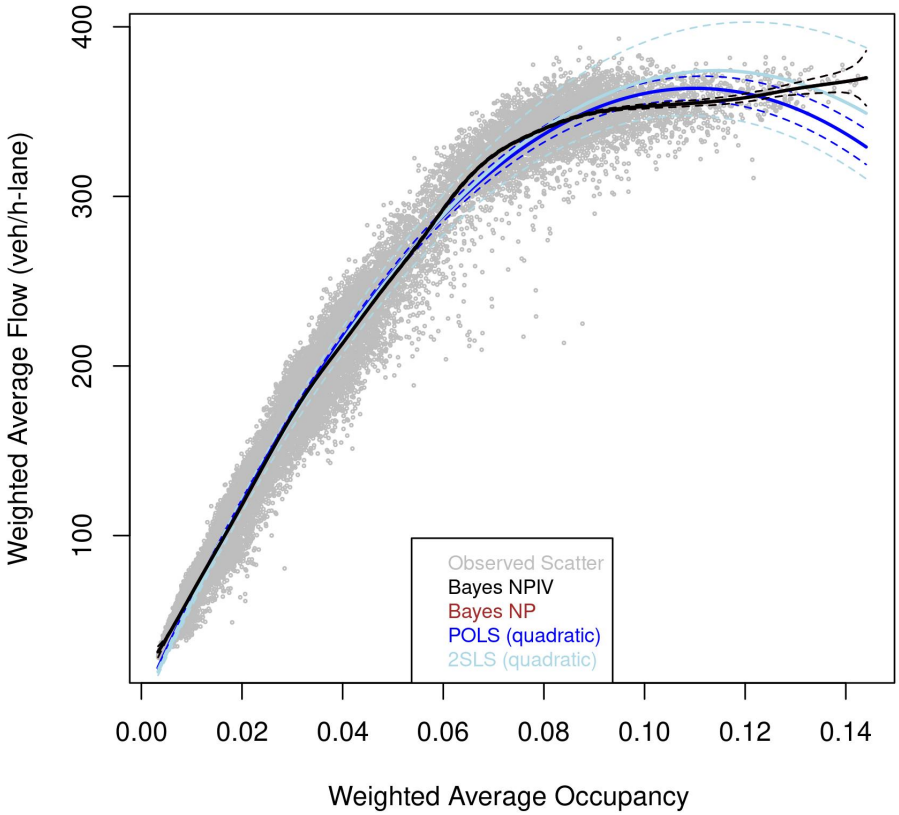


(b) Comparison of different estimators.

Figure C.28: Estimated MFD for Stuttgart

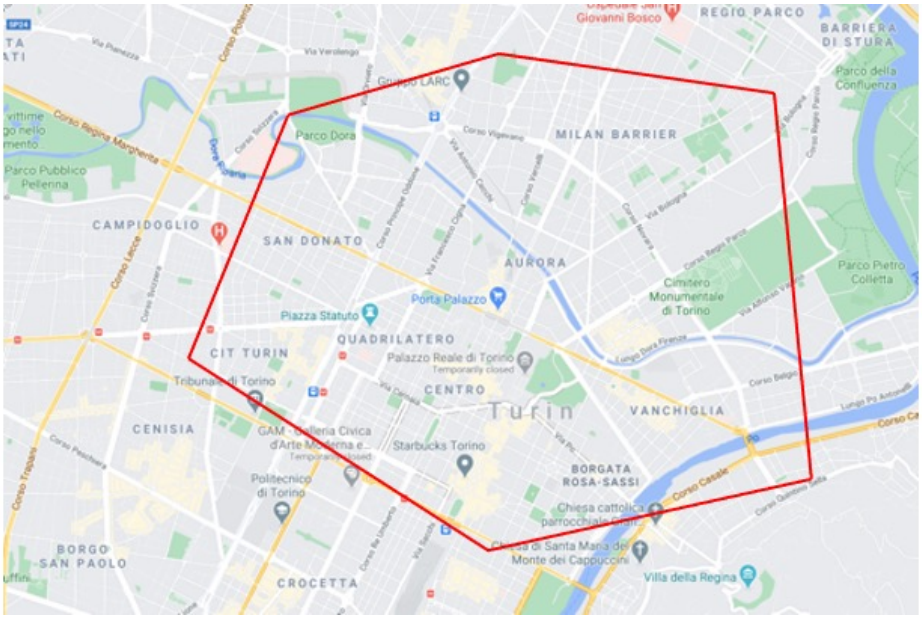


(a) Network exhibit used for the MFD estimation.

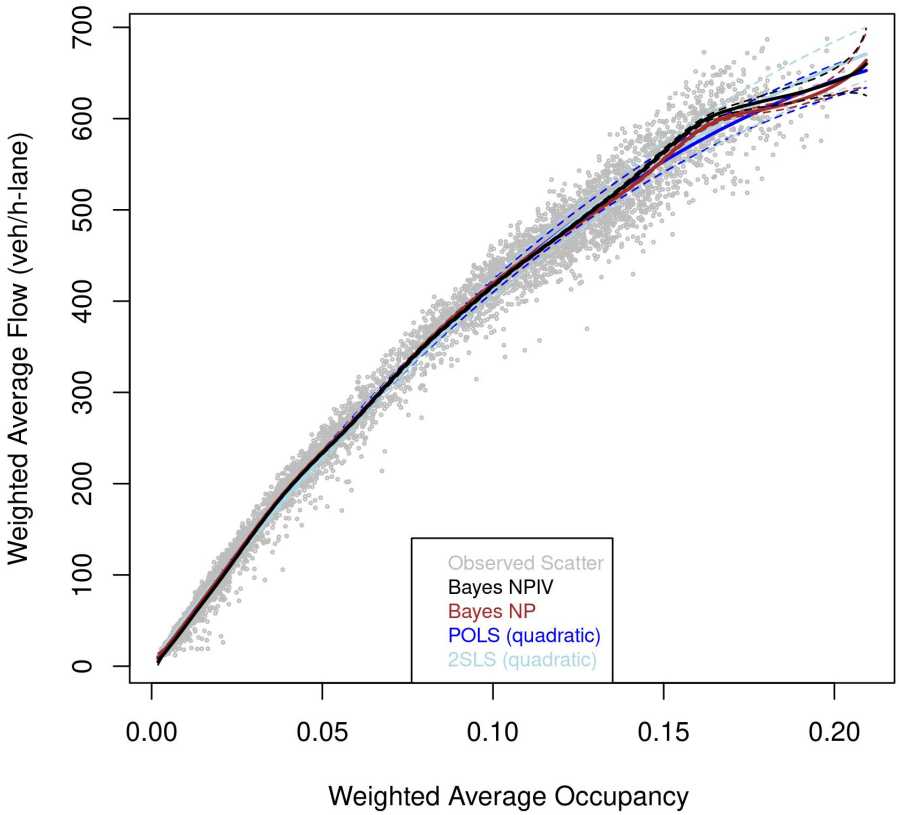


(b) Comparison of different estimators.

Figure C.29: Estimated MFD for Tokyo

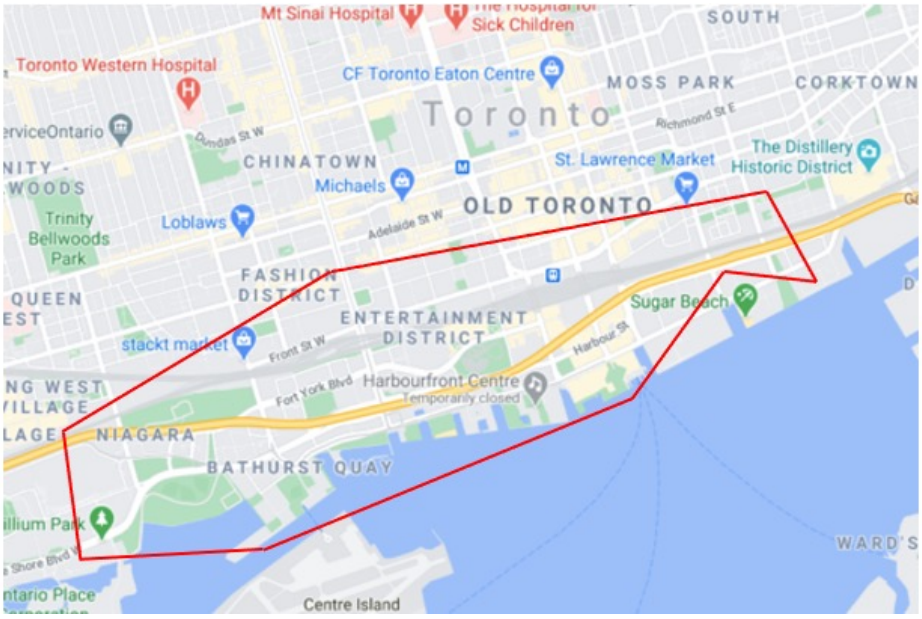


(a) Network exhibit used for the MFD estimation.

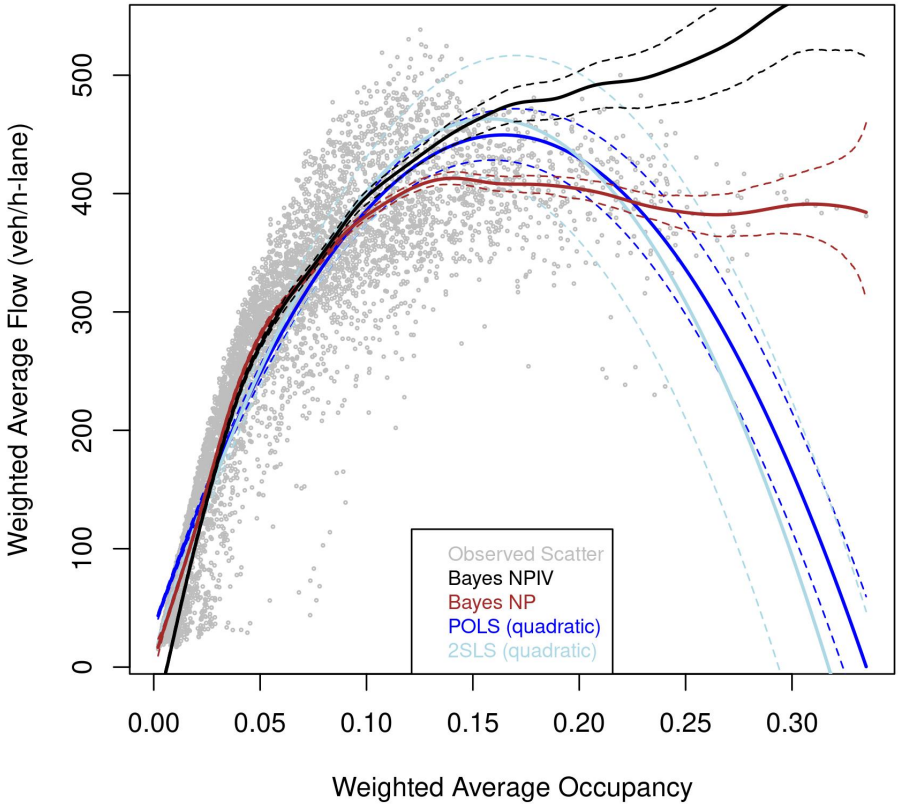


(b) Comparison of different estimators.

Figure C.30: Estimated MFD for Torino

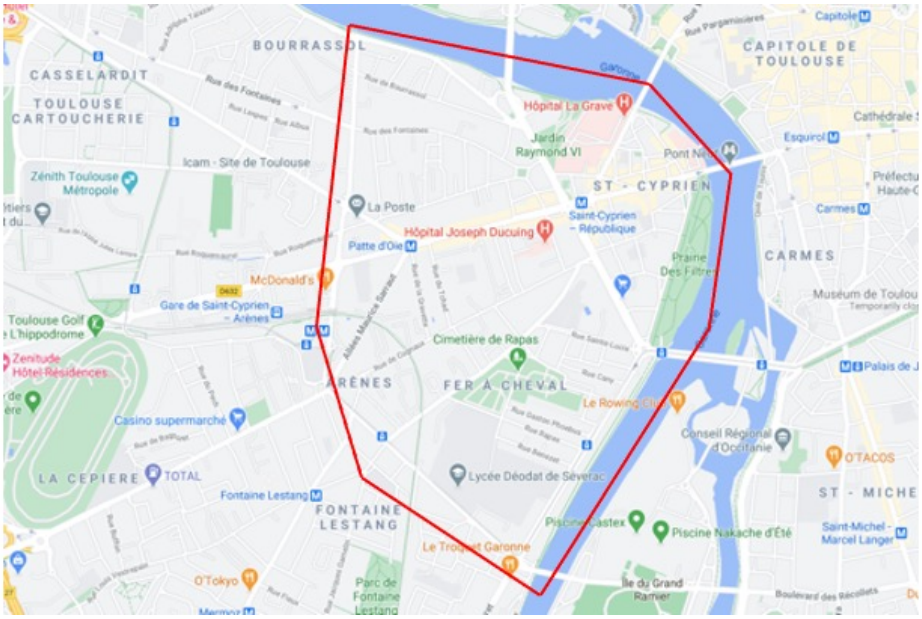


(a) Network exhibit used for the MFD estimation.

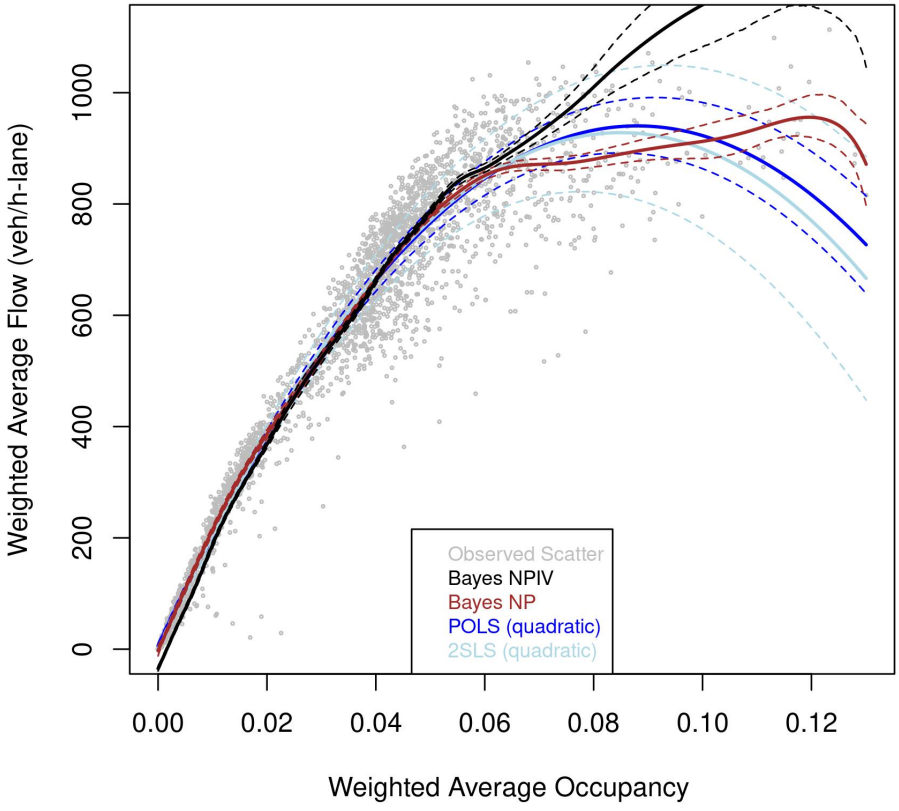


(b) Comparison of different estimators.

Figure C.31: Estimated MFD for Toronto

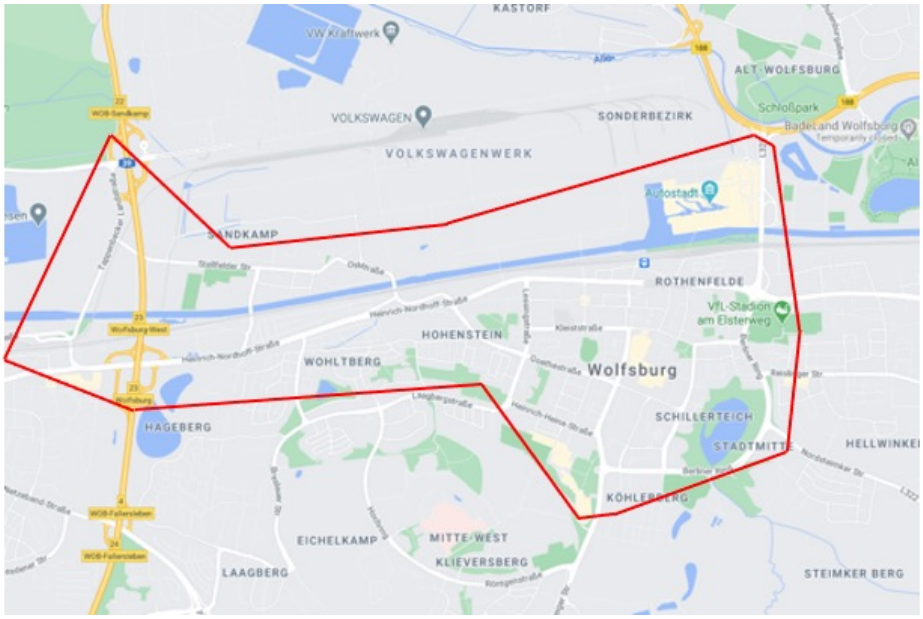


(a) Network exhibit used for the MFD estimation.

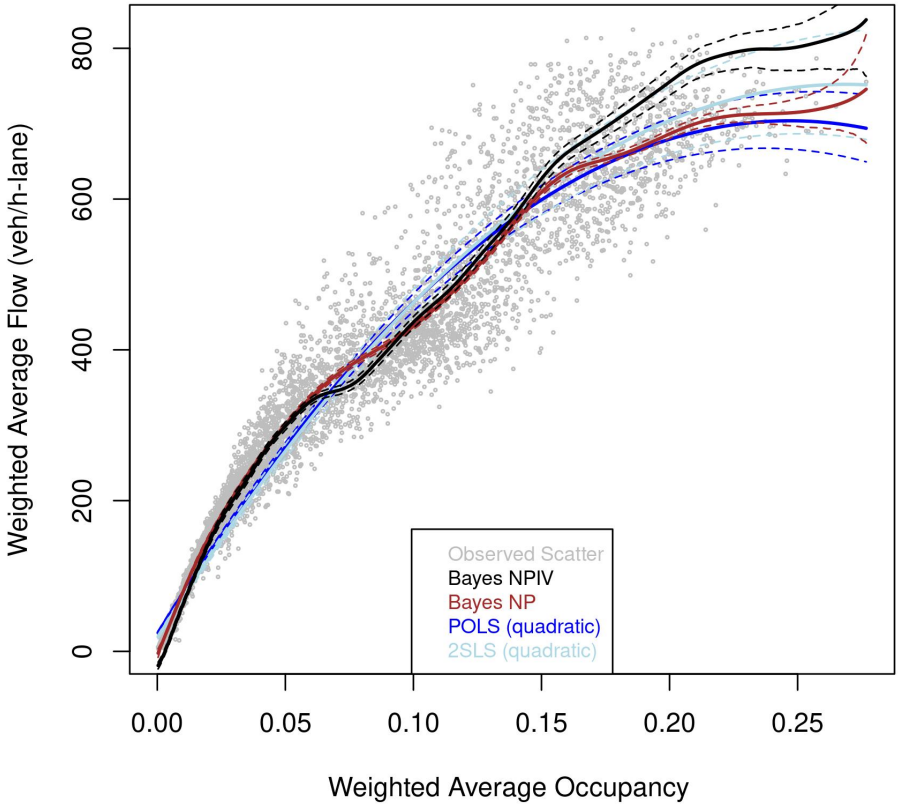


(b) Comparison of different estimators.

Figure C.32: Estimated MFD for Toulouse

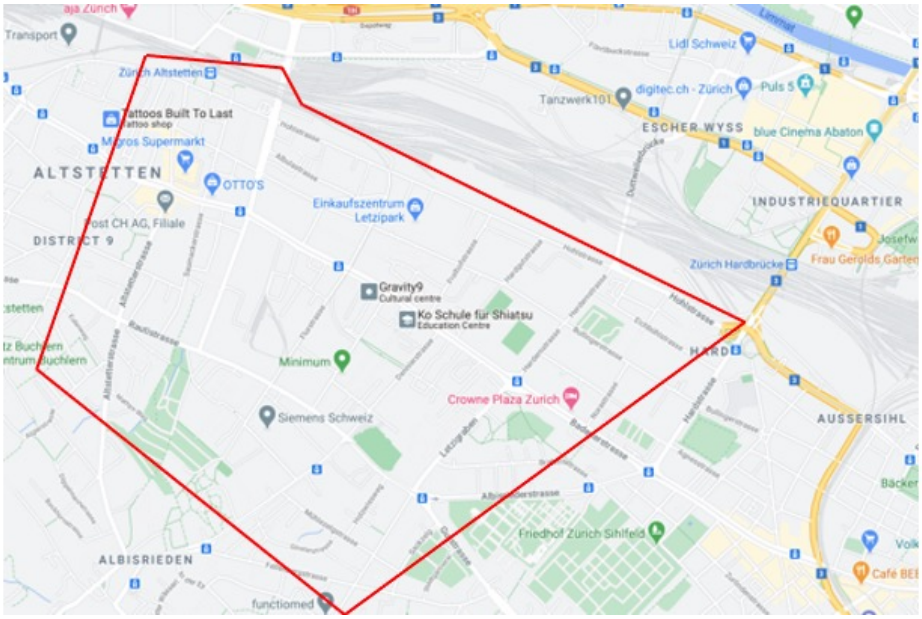


(a) Network exhibit used for the MFD estimation.

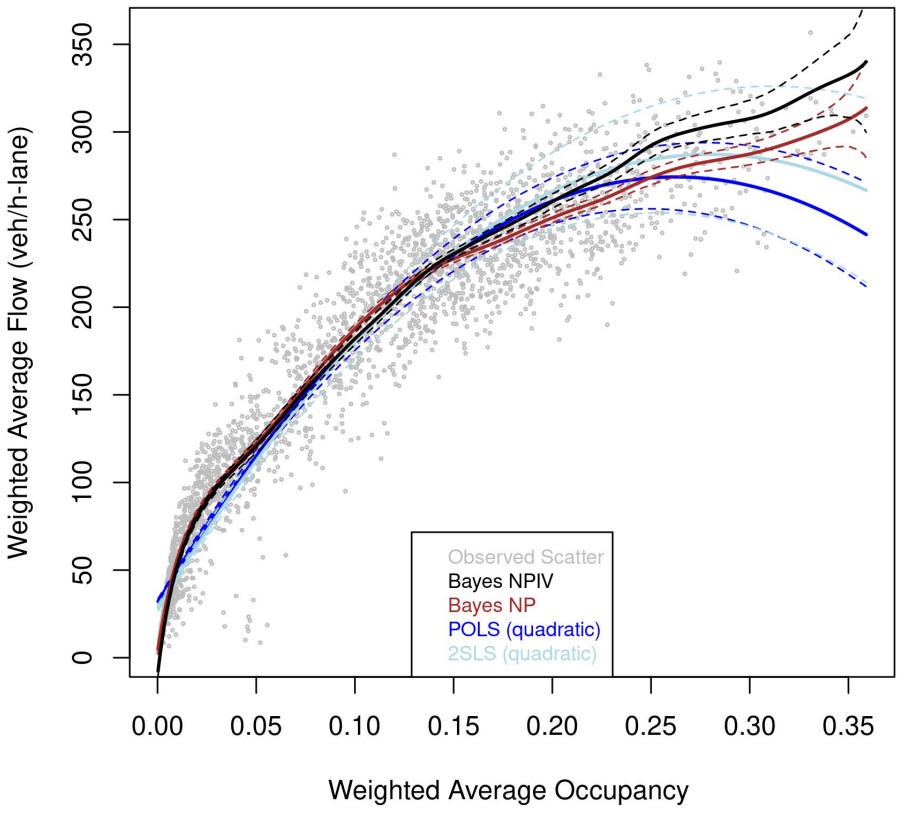


(b) Comparison of different estimators.

Figure C.33: Estimated MFD for Wolfsburg

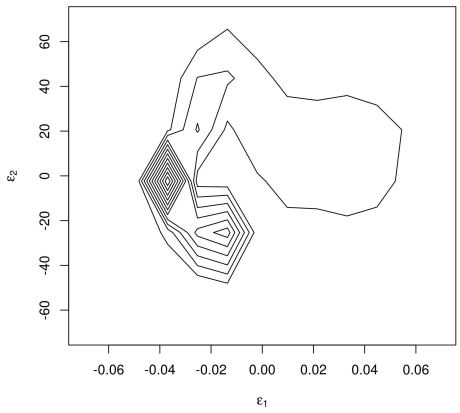


(a) Network exhibit used for the MFD estimation.

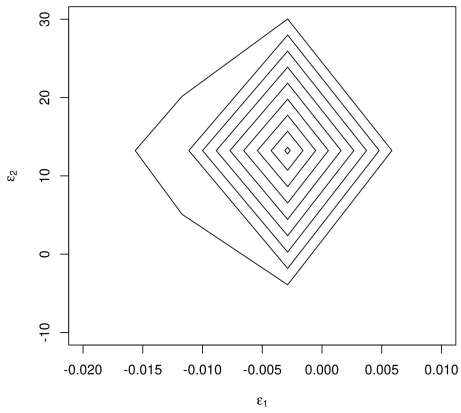


(b) Comparison of different estimators.

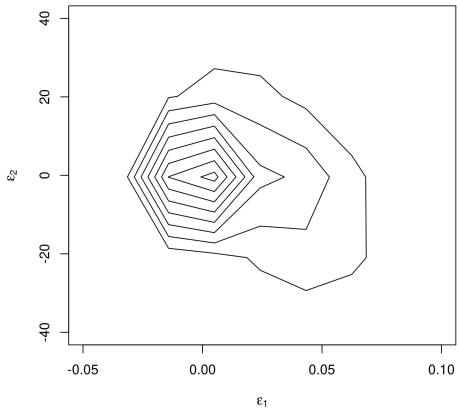
Figure C.34: Estimated MFD for Zurich



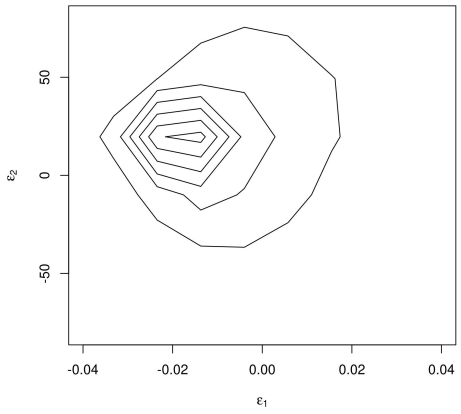
(a) Augsburg



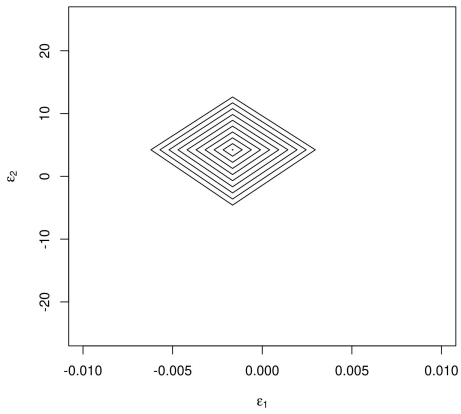
(b) Basel



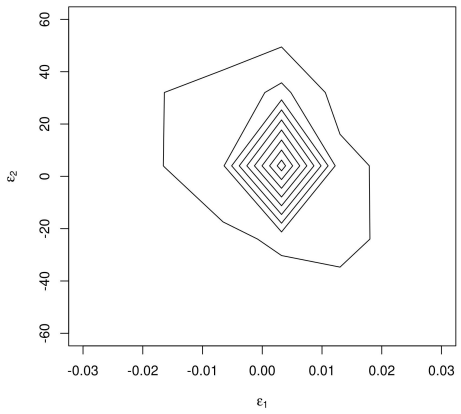
(c) Bern



(d) Birmingham

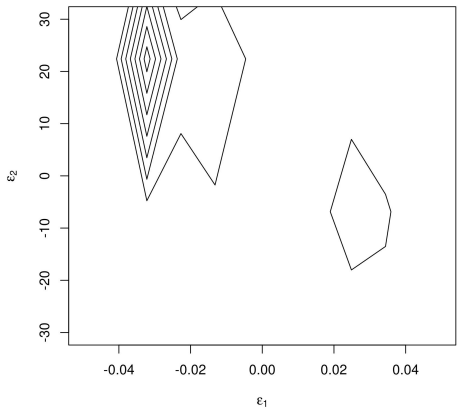


(e) Bolton

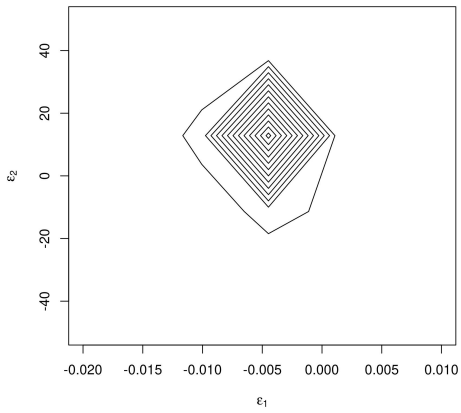


(f) Bordeaux

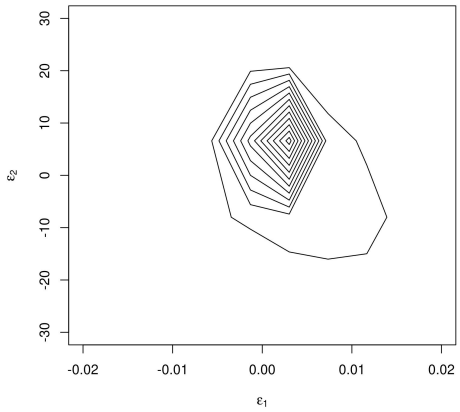
Figure C.35: Distribution of Errors.



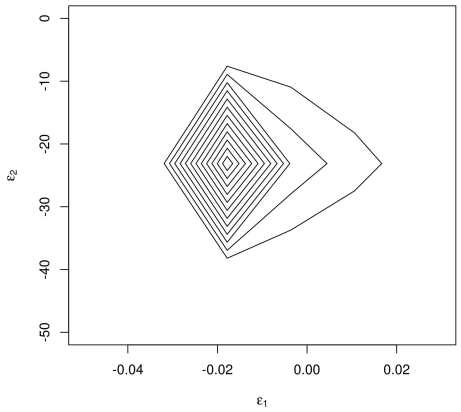
(g) Bremen



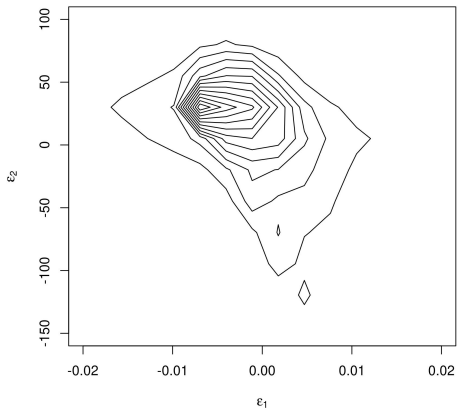
(h) Cagliari



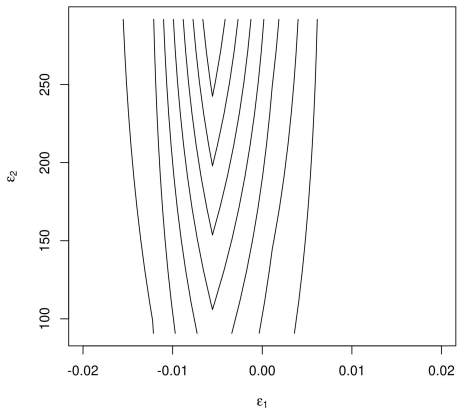
(i) Constance



(j) Darmstadt

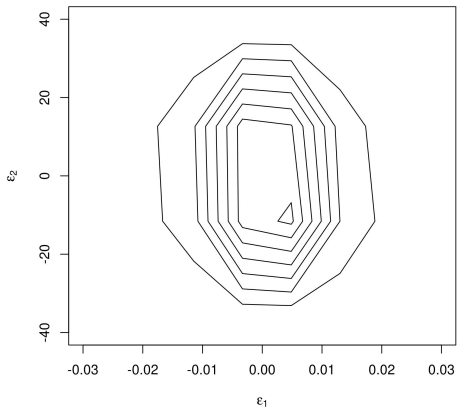


(k) Essen

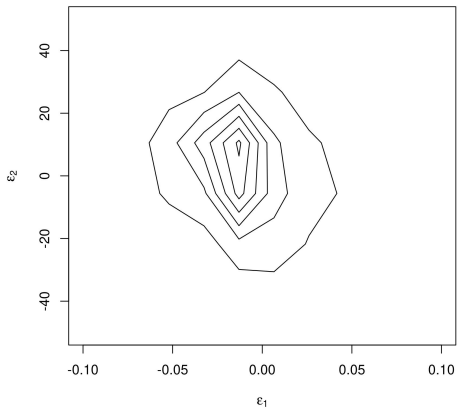


(l) Graz

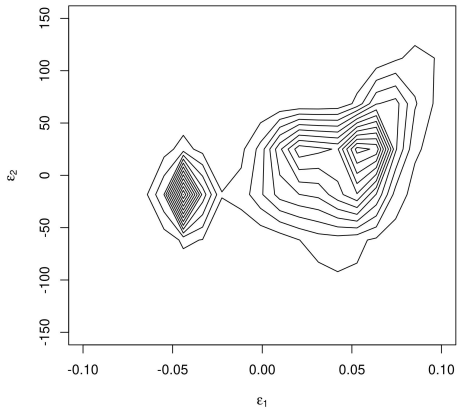
Figure C.35: Distribution of Errors.



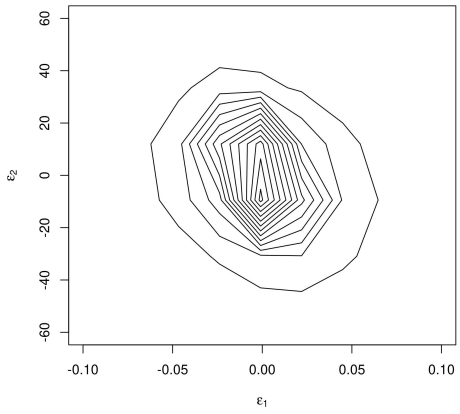
(m) Groningen



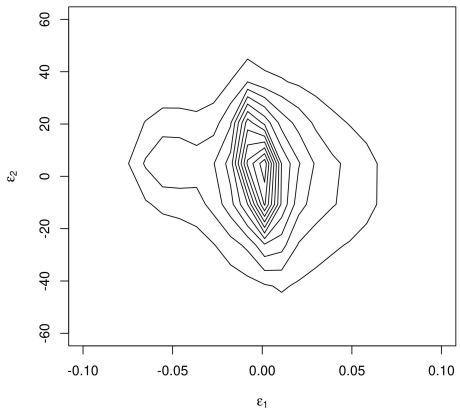
(n) Hamburg



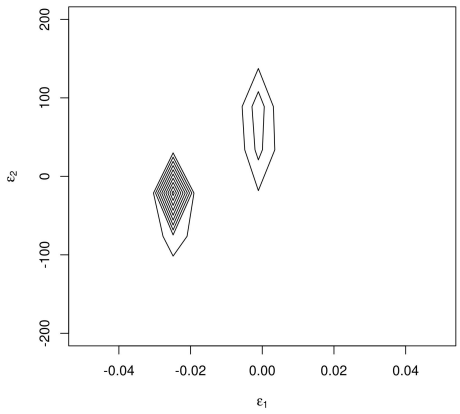
(o) Innsbruck



(p) Kassel

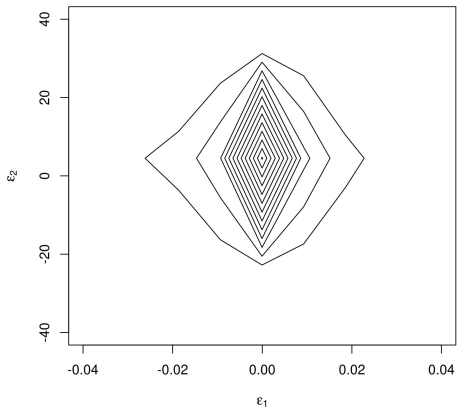


(q) London

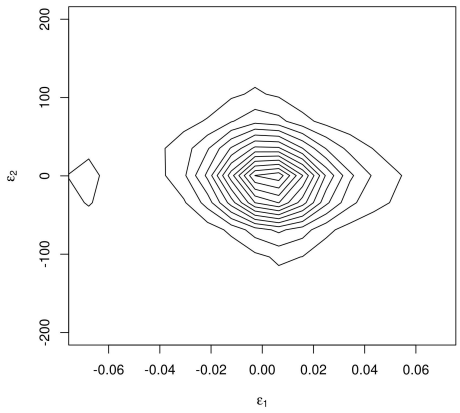


(r) Los Angeles

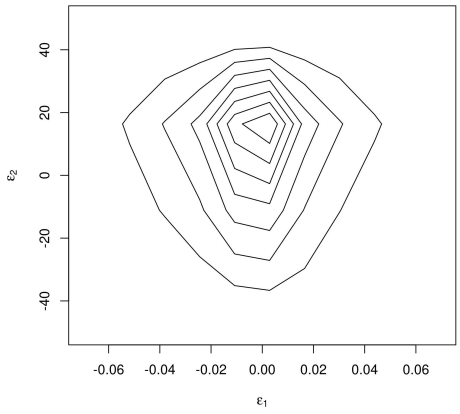
Figure C.35: Distribution of Errors.



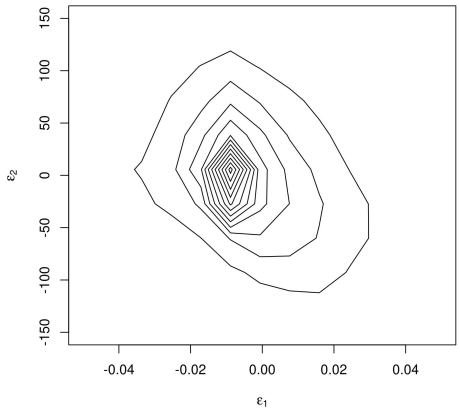
(s) Luzern



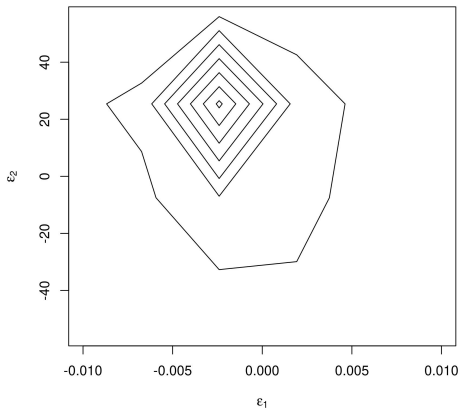
(t) Madrid



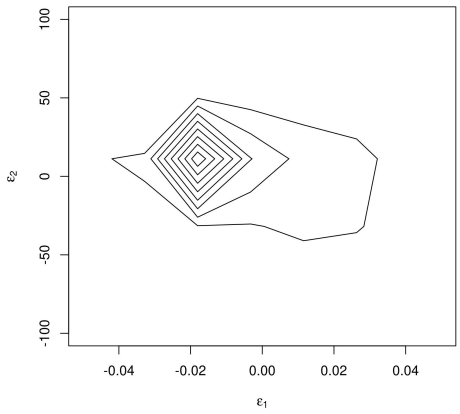
(u) Manchester



(v) Marseille

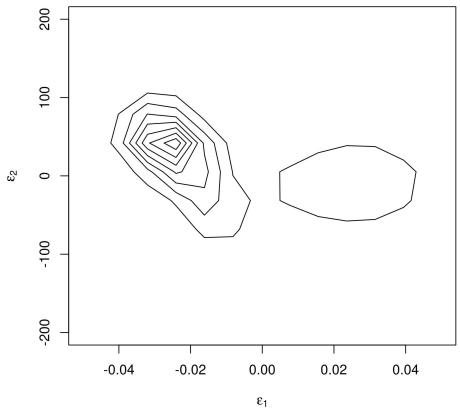


(w) Paris

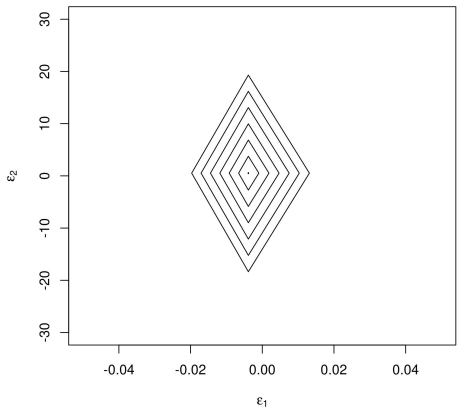


(x) Rotterdam

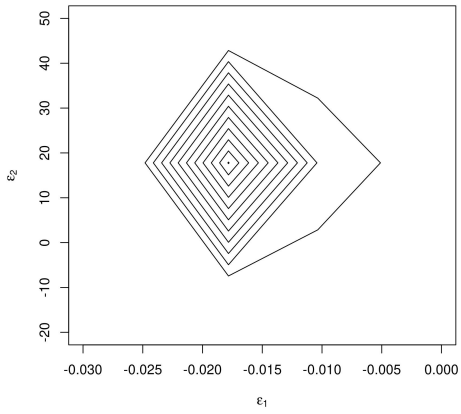
Figure C.35: Distribution of Errors.



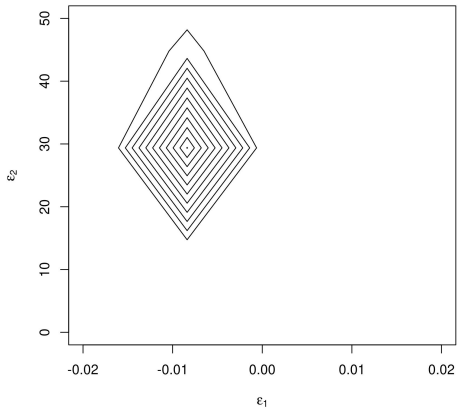
(y) Santander



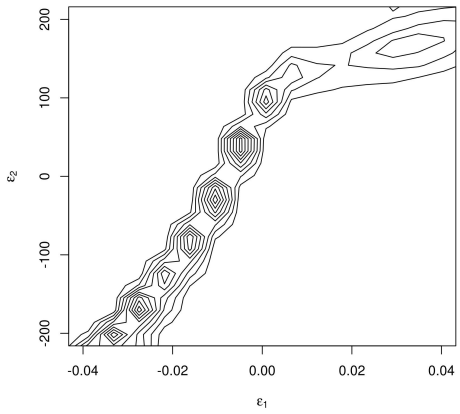
(z) Speyer



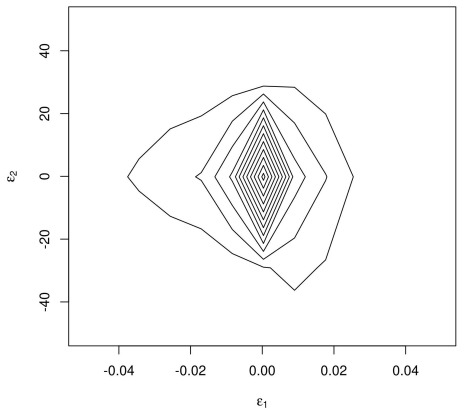
(aa) Strasbourg



(ab) Stuttgart

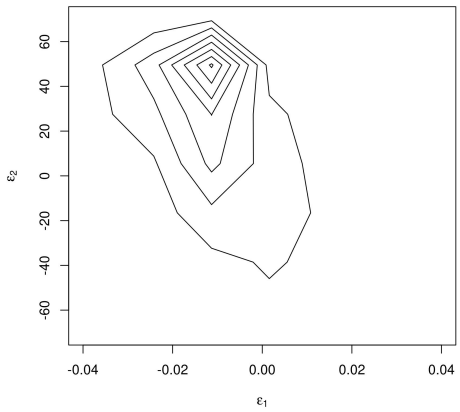


(ac) Tokyo

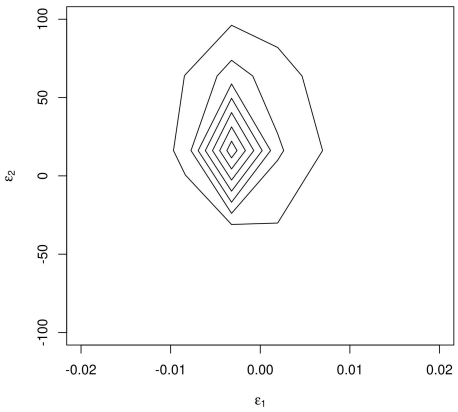


(ad) Torino

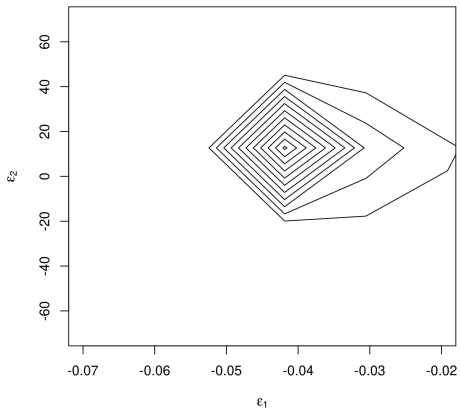
Figure C.35: Distribution of Errors.



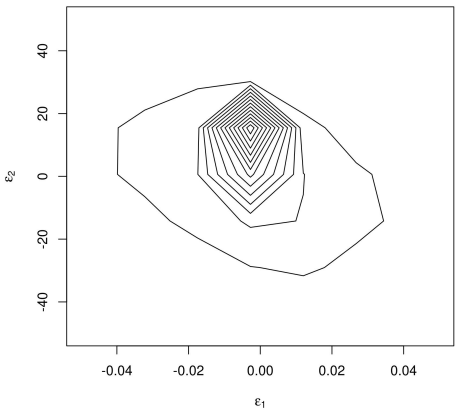
(ae) Toronto



(af) Toulouse

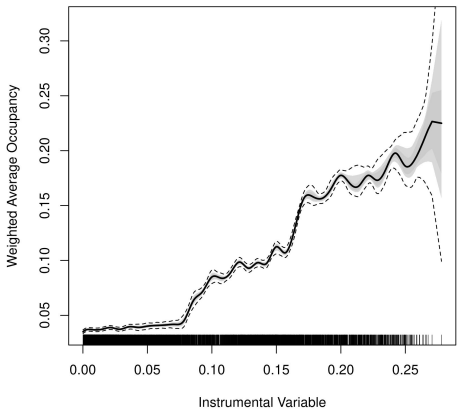


(ag) Wolfsburg

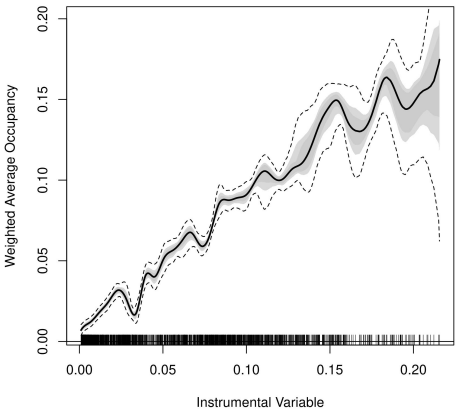


(ah) Zurich

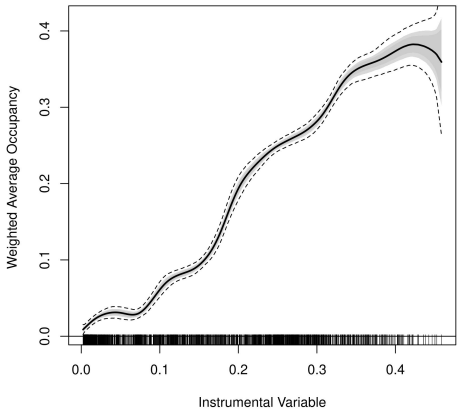
Figure C.35: Distribution of Errors.



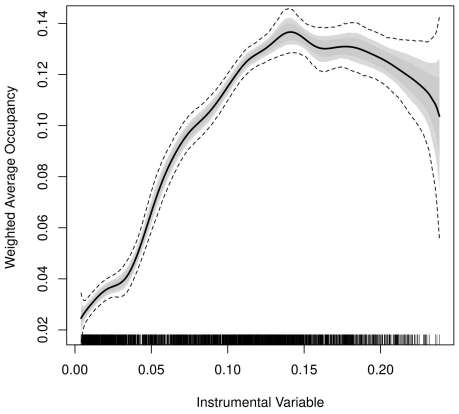
(a) Augsburg



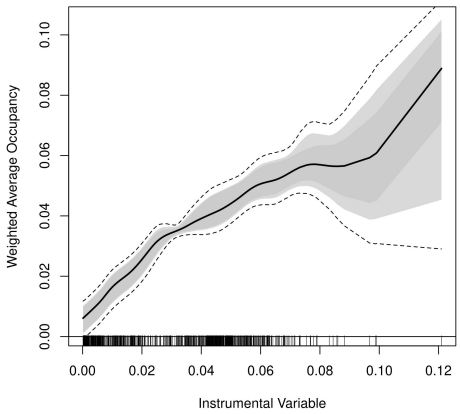
(b) Basel



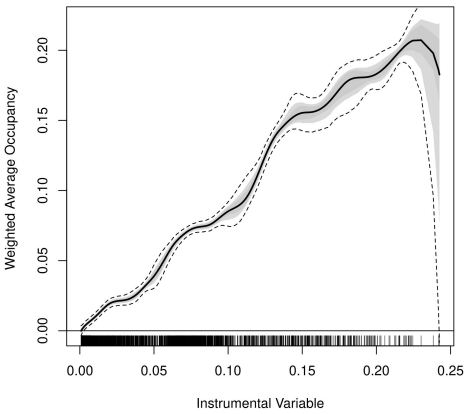
(c) Bern



(d) Birmingham

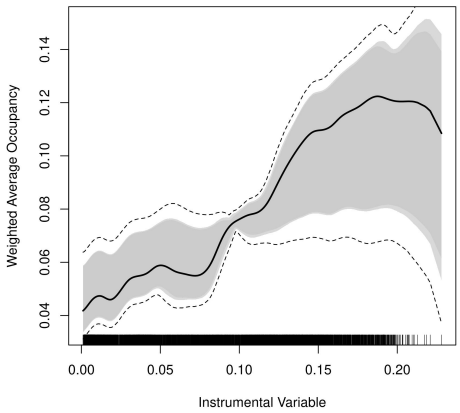


(e) Bolton

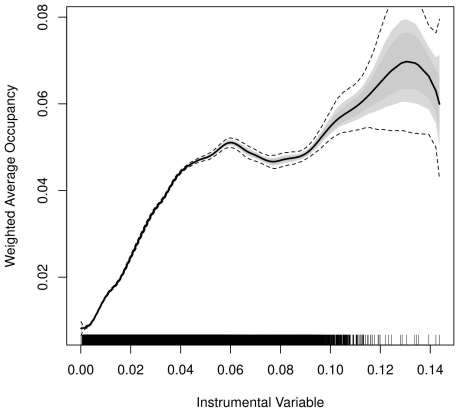


(f) Bordeaux

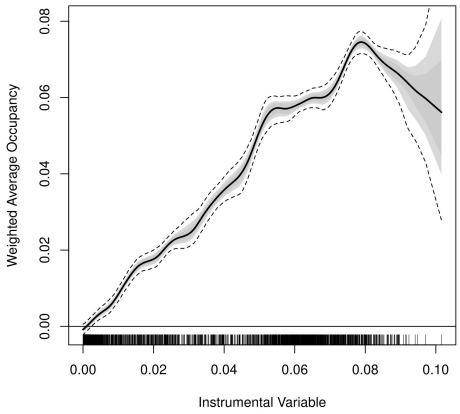
Figure C.36: Relevance of Instruments.



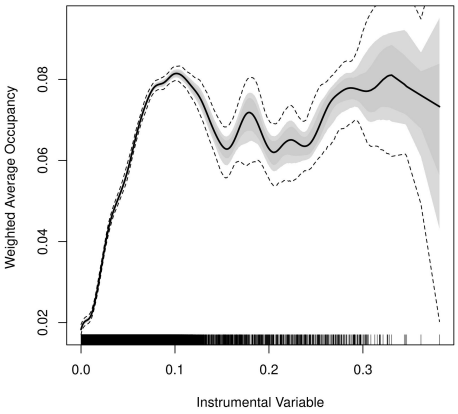
(g) Bremen



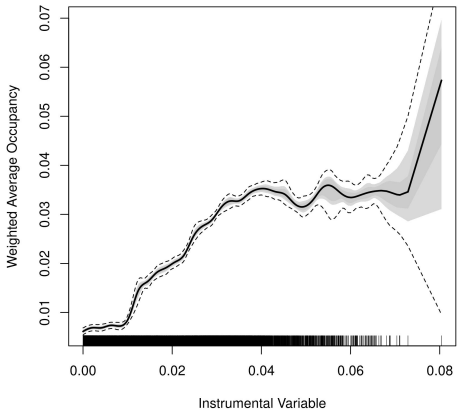
(h) Cagliari



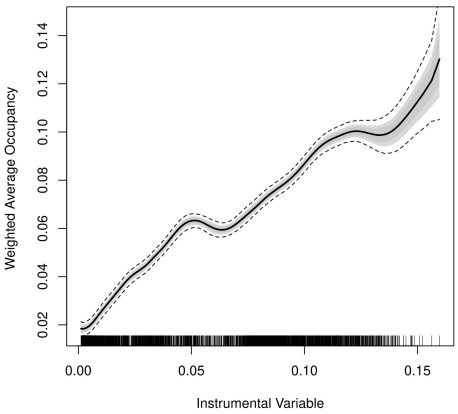
(i) Constance



(j) Darmstadt

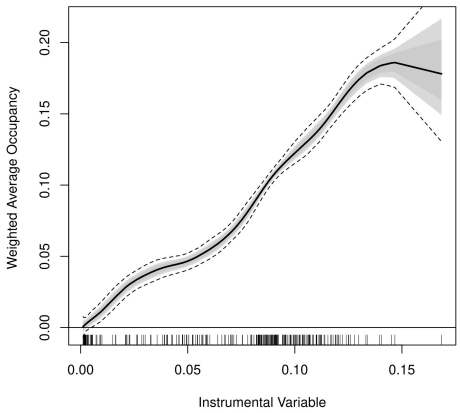


(k) Essen

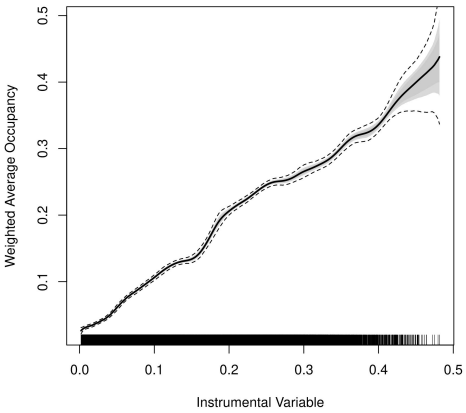


(l) Graz

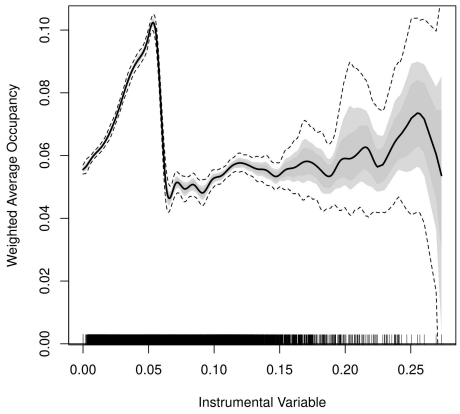
Figure C.36: Relevance of Instruments.



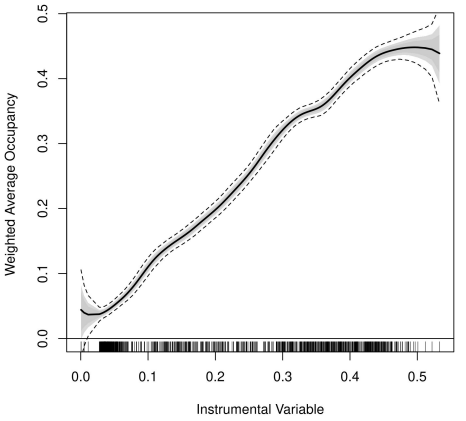
(m) Groningen



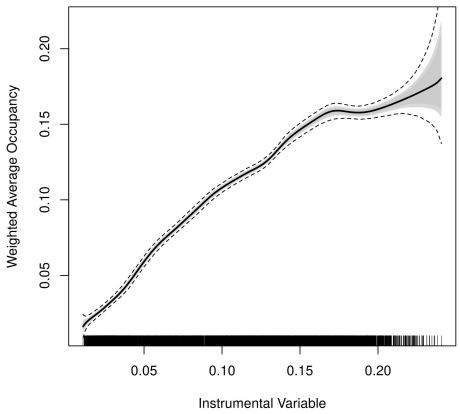
(n) Hamburg



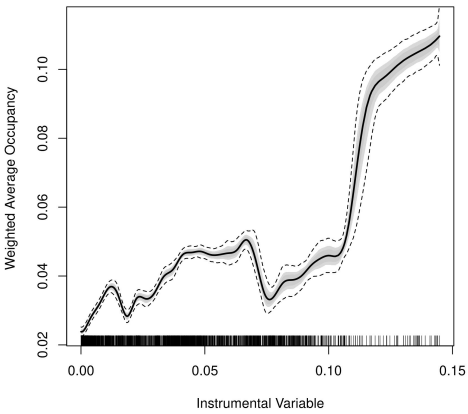
(o) Innsbruck



(p) Kassel

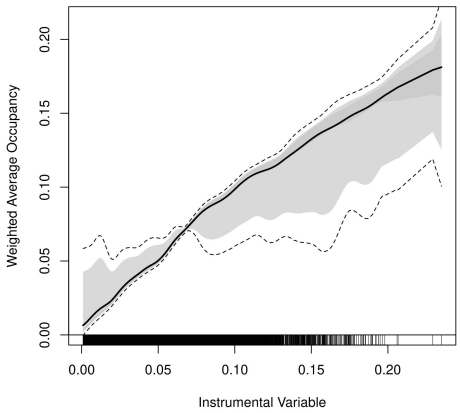


(q) London

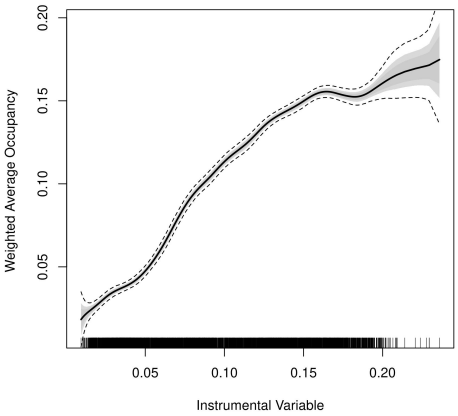


(r) Los Angeles

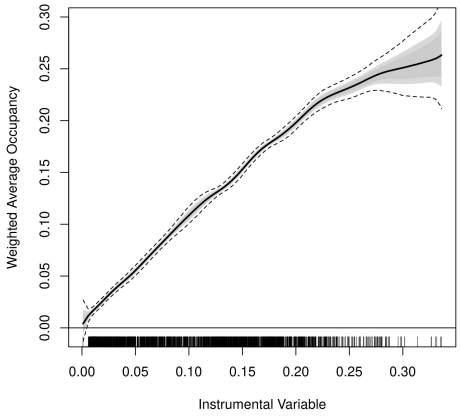
Figure C.36: Relevance of Instruments.



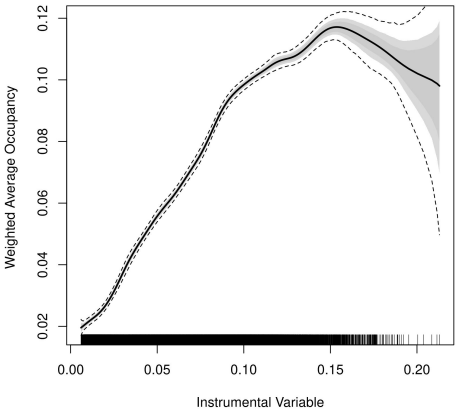
(s) Luzern



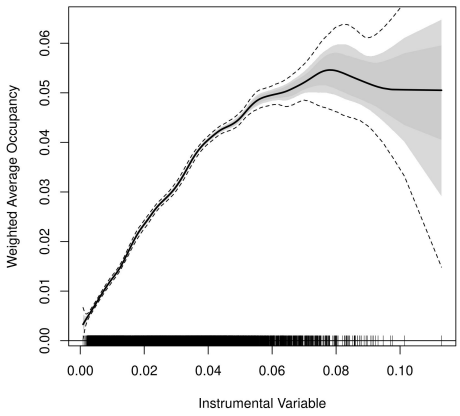
(t) Madrid



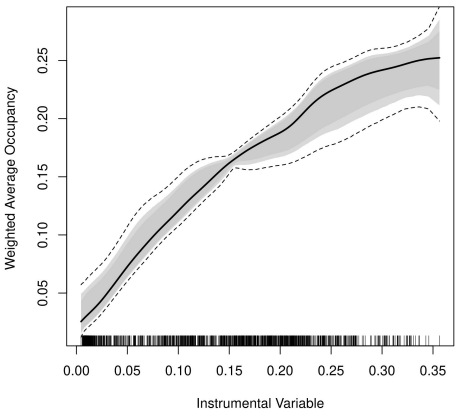
(u) Manchester



(v) Marseille

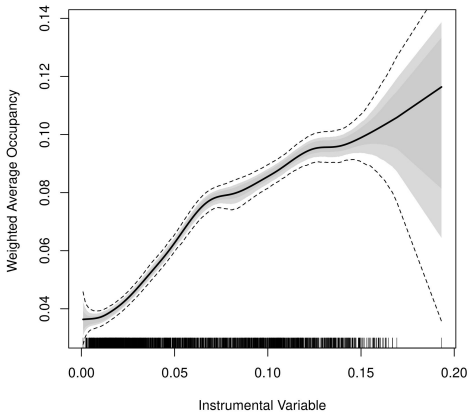


(w) Paris

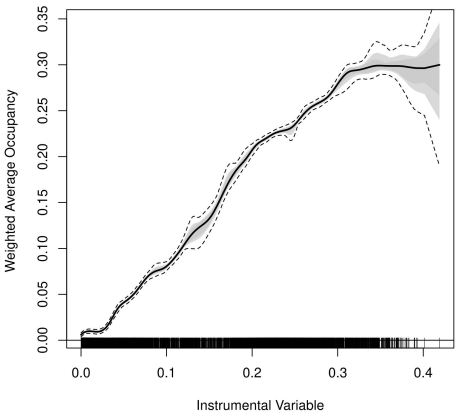


(x) Rotterdam

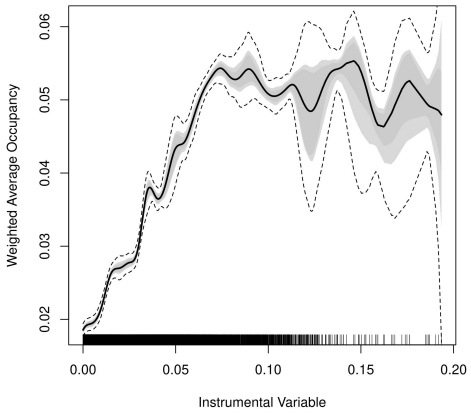
Figure C.36: Relevance of Instruments.



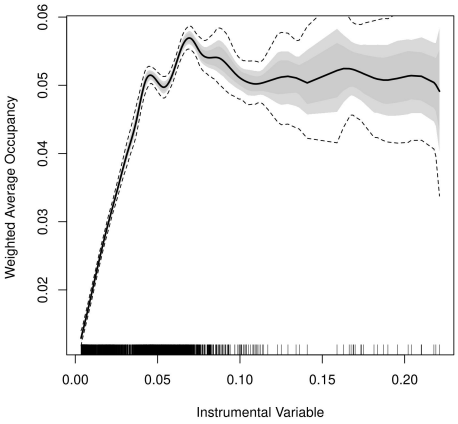
(y) Santander



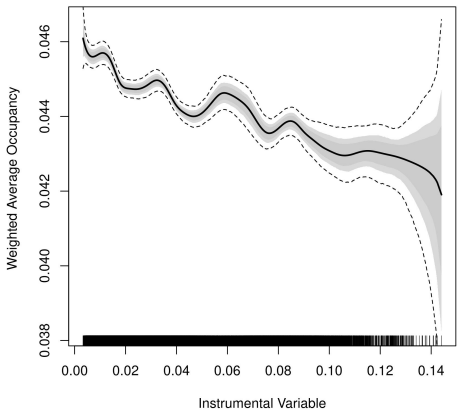
(z) Speyer



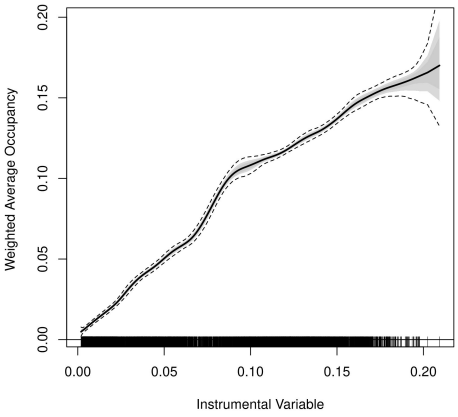
(aa) Strasbourg



(ab) Stuttgart

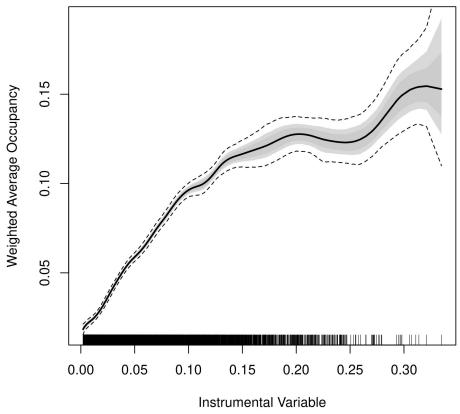


(ac) Tokyo

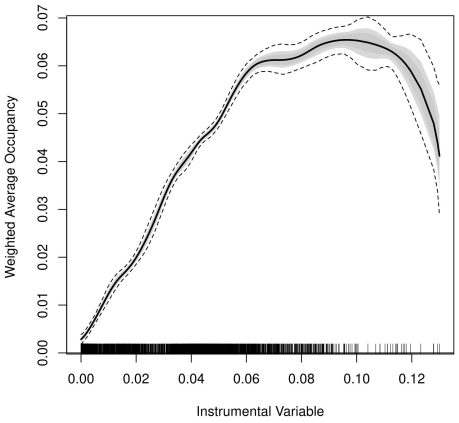


(ad) Torino

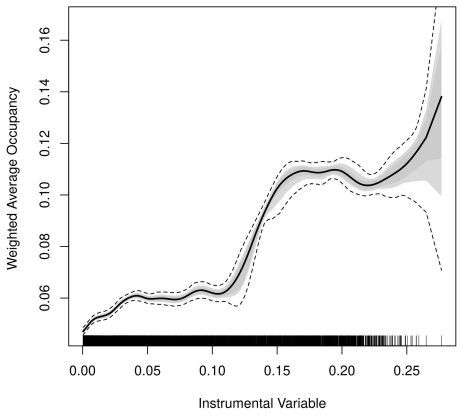
Figure C.36: Relevance of Instruments.



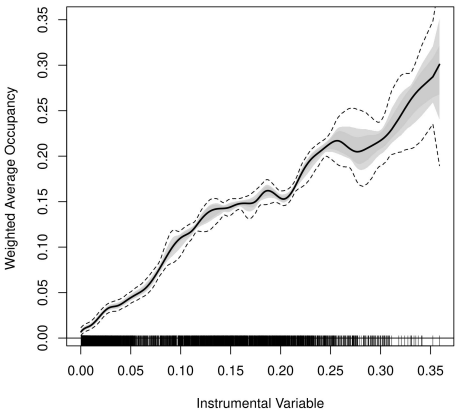
(ae) Toronto



(af) Toulouse



(ag) Wolfsburg



(ah) Zurich

Figure C.36: Relevance of Instruments.

Appendix D

Supplementary Material: Chapter 6

D.1 Map of Hong Kong Mass Transit Railway

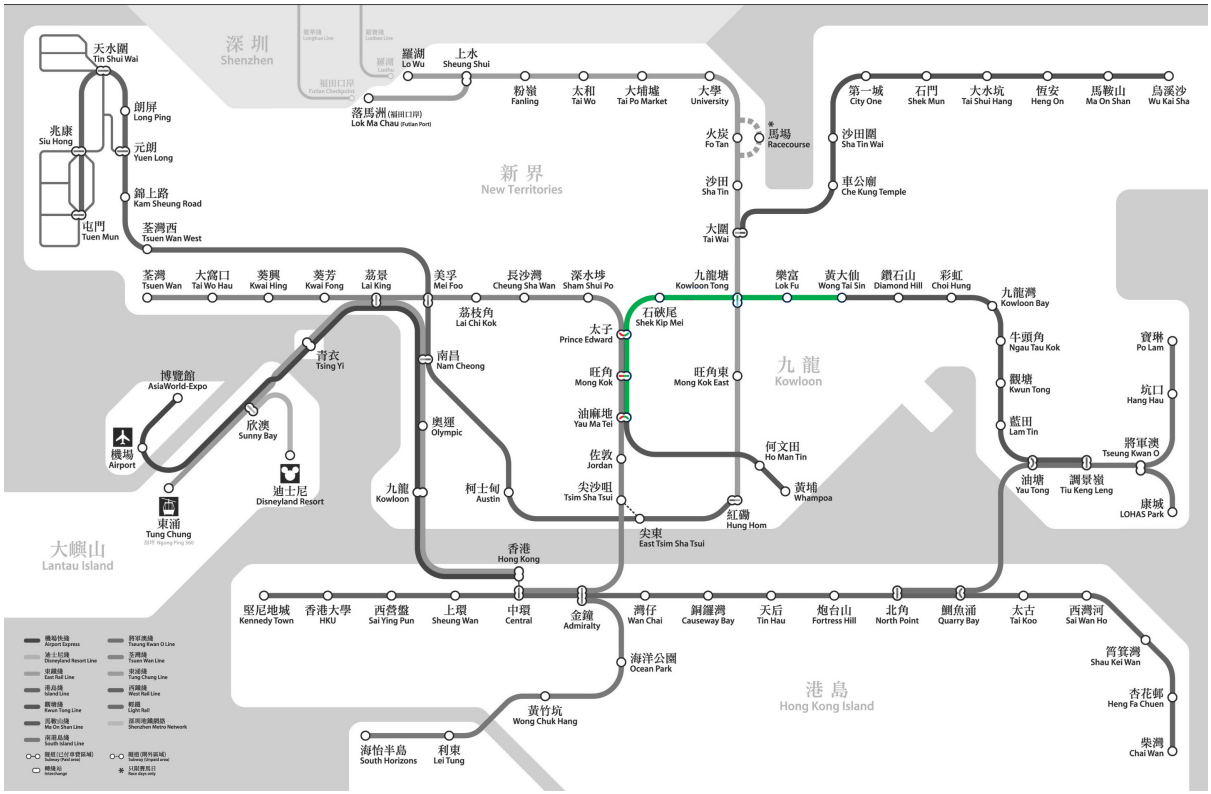
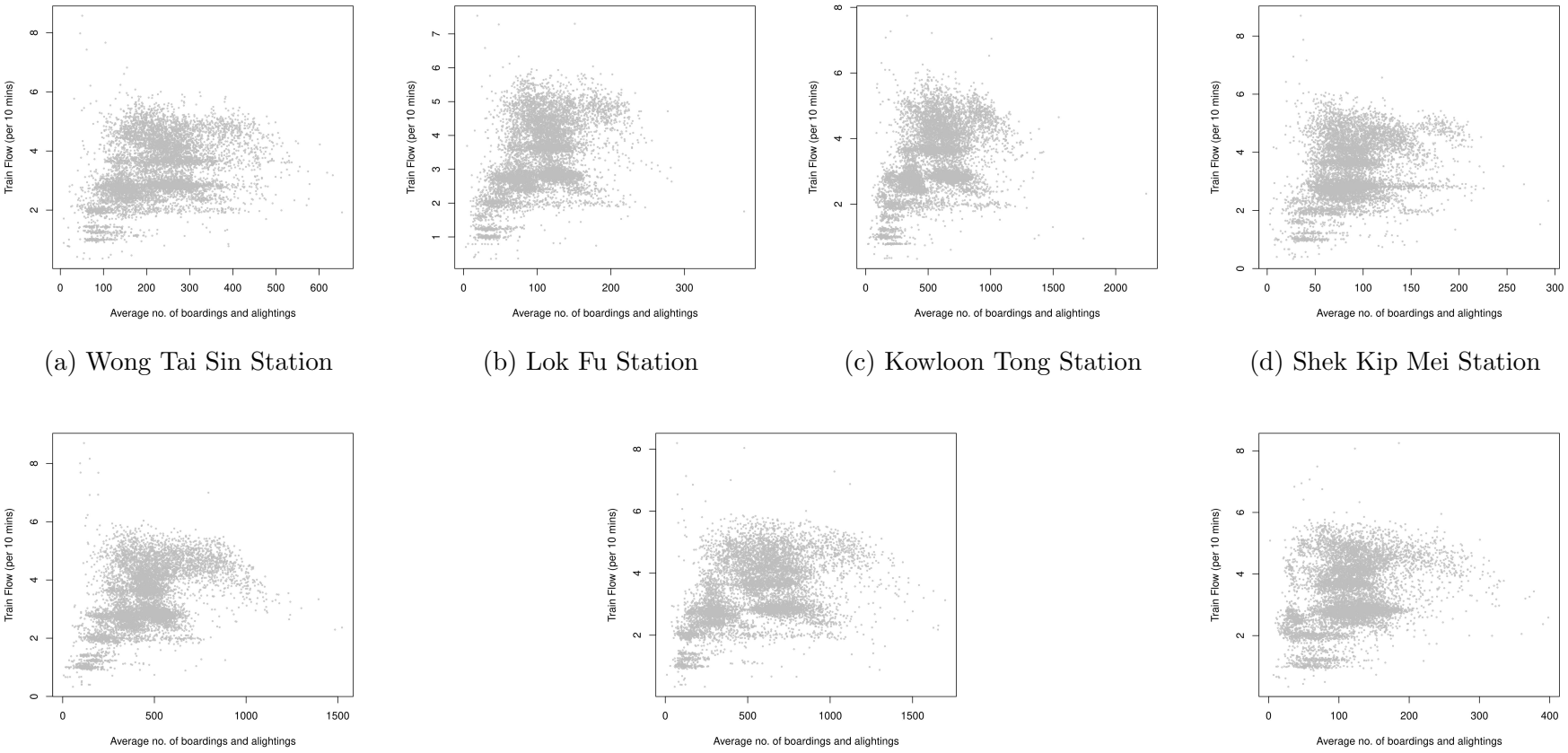


Figure D.1: Full map the MTR network where the line that we study is highlighted in green.

D.2 Observed scatter plots of passenger movements vs train flow



(a) Wong Tai Sin Station

(b) Lok Fu Station

(c) Kowloon Tong Station

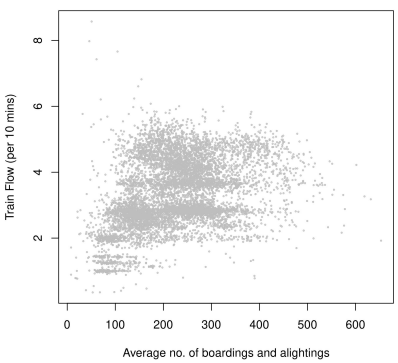
(d) Shek Kip Mei Station

(e) Prince Edward Station

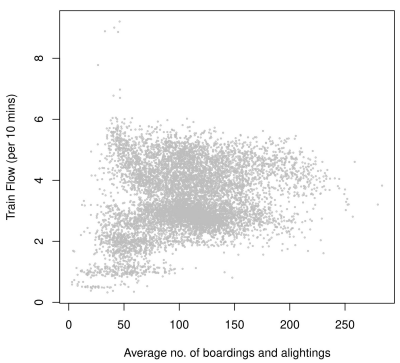
(f) Mong Kok Station

(g) Yau Ma Tei station

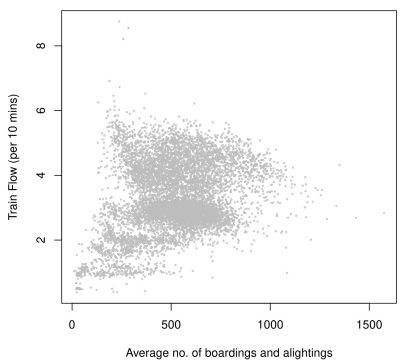
Figure D.2: Variation of observed train flow in the downward direction over passenger movements for the stations highlighted in Figure 6.5.



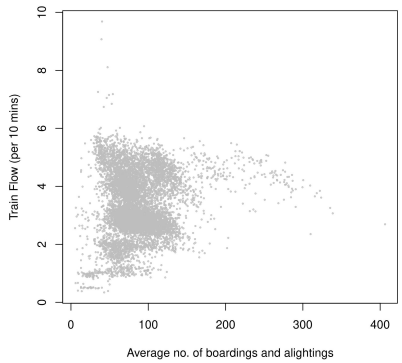
(a) Wong Tai Sin Station



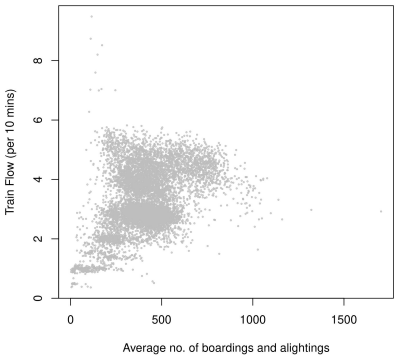
(b) Lok Fu Station



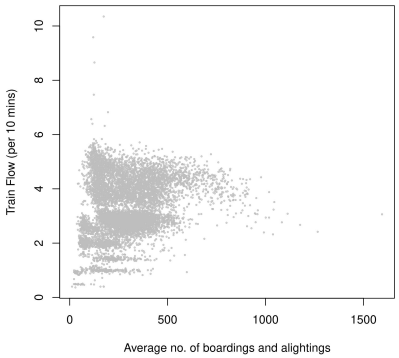
(c) Kowloon Tong Station



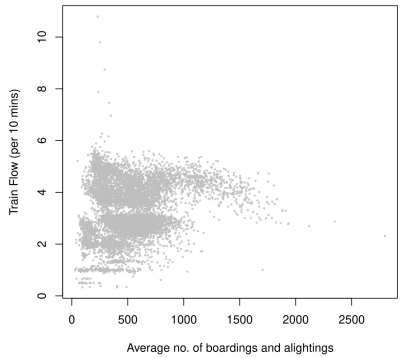
(d) Shek Kip Mei Station



(e) Prince Edward Station



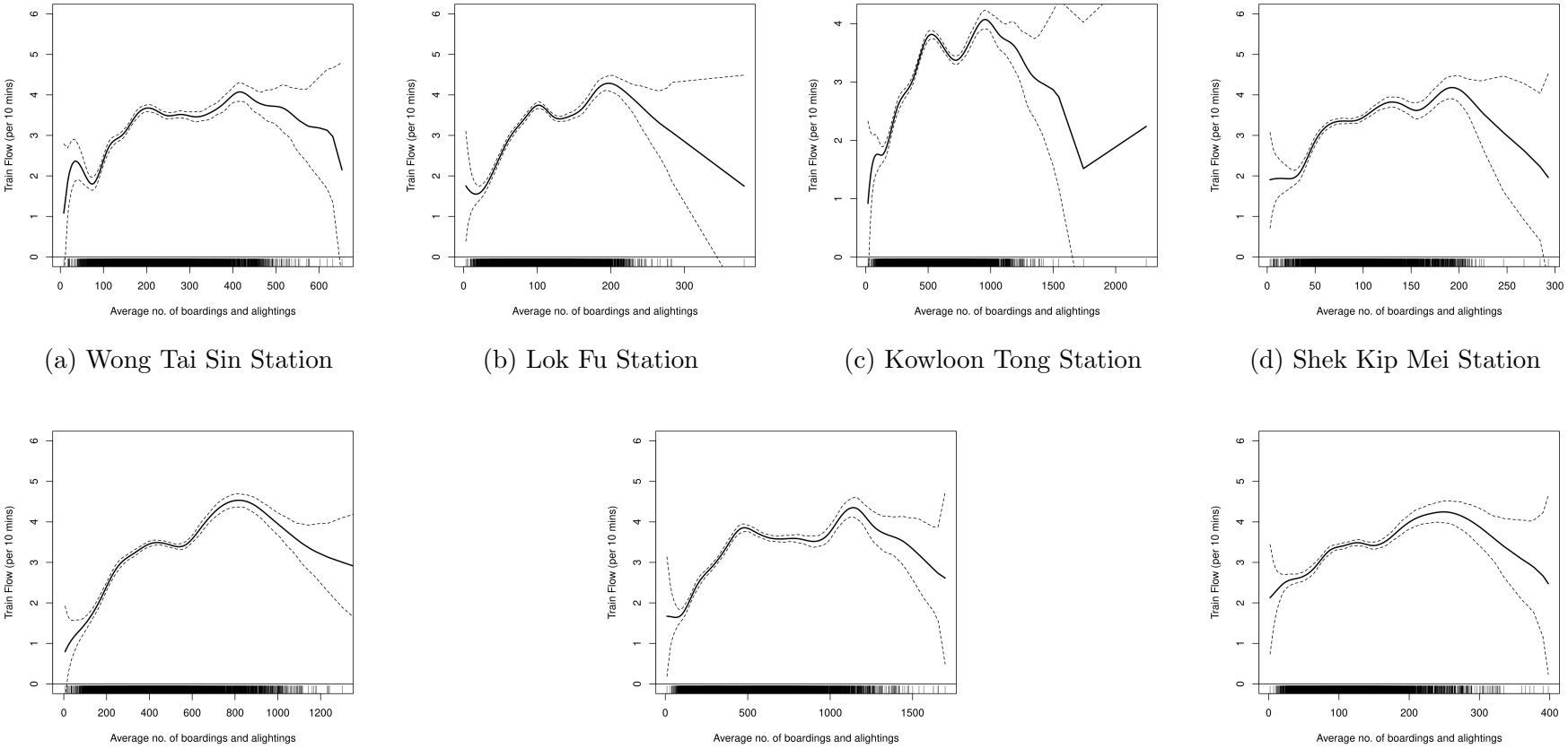
(f) Mong Kok Station



(g) Yau Ma Tei station

Figure D.3: Variation of observed train flow in the upward direction over passenger movements for the stations highlighted in Figure 6.5.

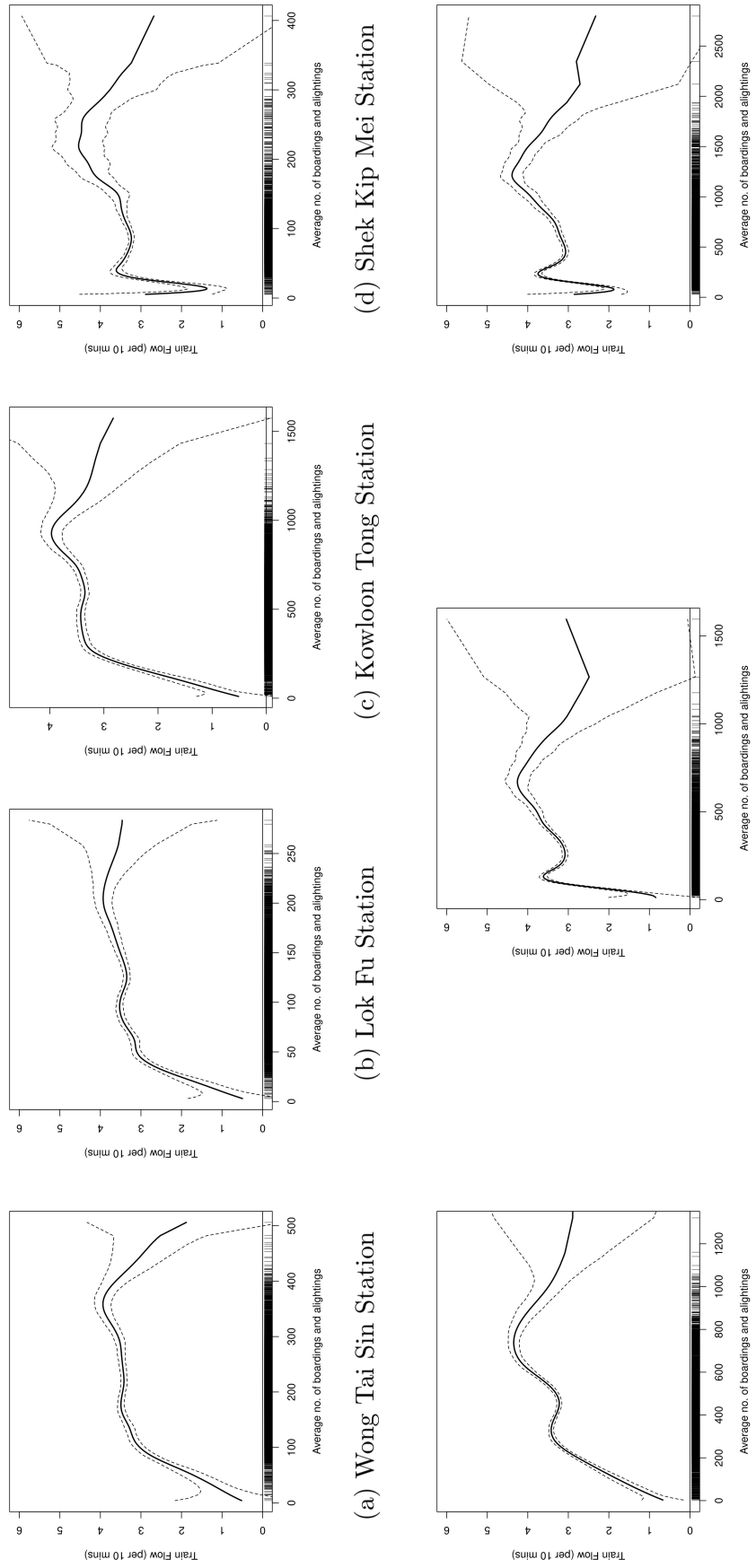
D.3 Results from Bayesian NP (non-IV) estimation



(a) Wong Tai Sin Station (b) Lok Fu Station (c) Kowloon Tong Station (d) Shek Kip Mei Station

(e) Prince Edward Station (f) Mong Kok Station (g) Yau Ma Tei station

Figure D.4: Non-parametric (non-IV-based) based estimation results for train movements in the downward direction along the Kwun Tong Line.

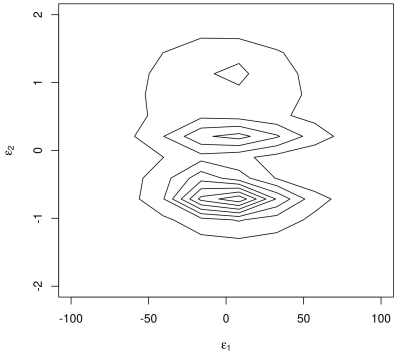


(a) Wong Tai Sin Station (b) Lok Fu Station (c) Kowloon Tong Station (d) Shek Kip Mei Station

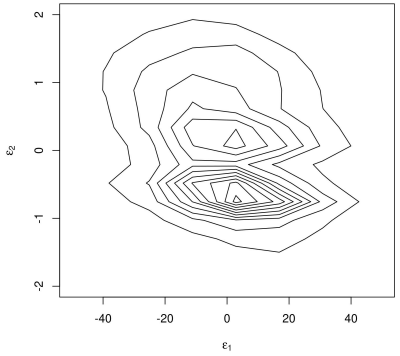
(e) Prince Edward Station (f) Mong Kok Station (g) Yau Ma Tei station

Figure D.5: Non-parametric (non-IV-based) estimation results for train movements in the upward direction along the Kwun Tong Line..

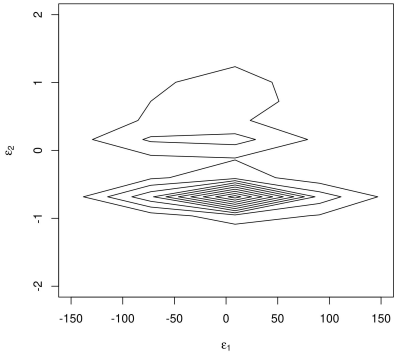
D.4 Distribution of Errors



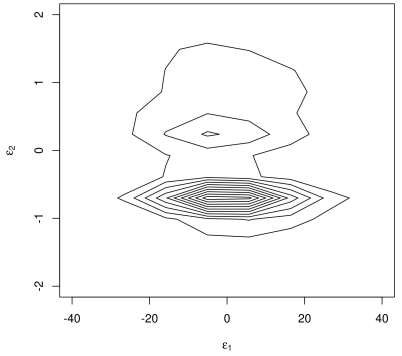
(a) Wong Tai Sin Station



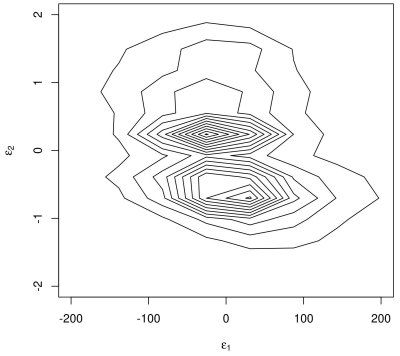
(b) Lok Fu Station



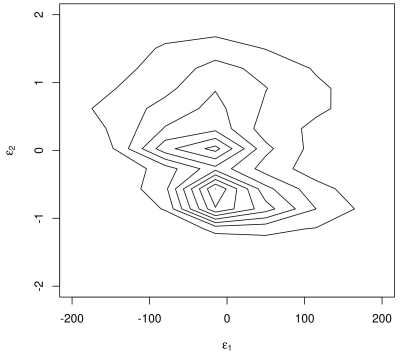
(c) Kowloon Tong Station



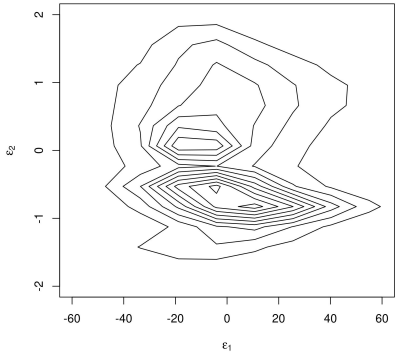
(d) Shek Kip Mei Station



(e) Prince Edward Station

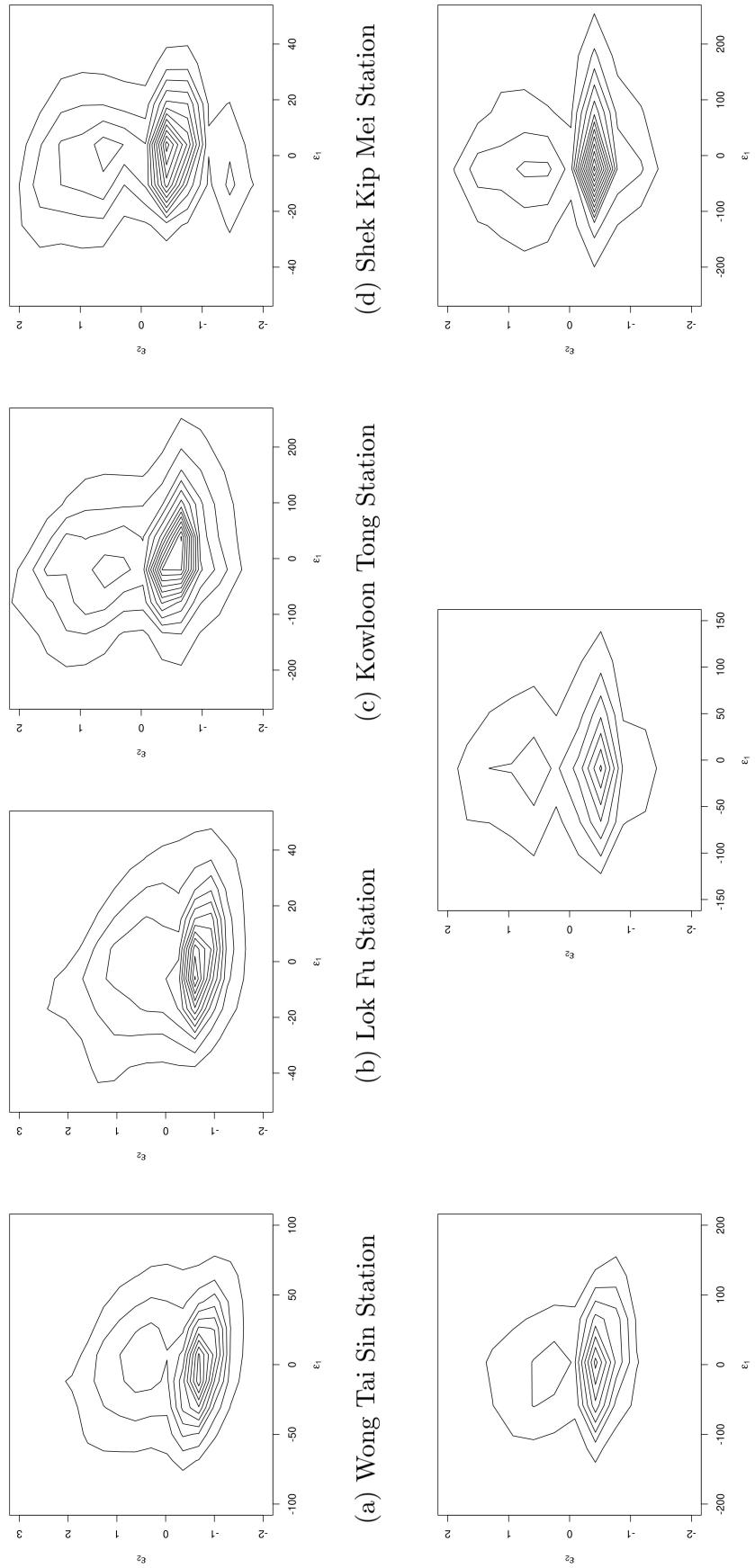


(f) Mong Kok Station



(g) Yau Ma Tei station

Figure D.6: Distribution of errors from analyses of train movements in the downward direction along the Kwun Tong Line.

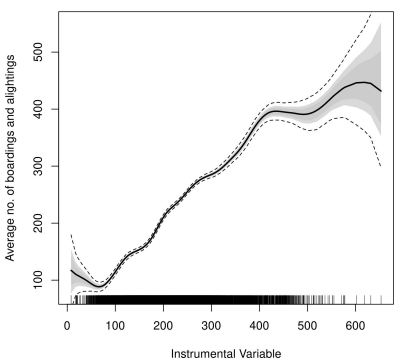


(a) Wong Tai Sin Station (b) Lok Fu Station (c) Kowloon Tong Station (d) Shek Kip Mei Station

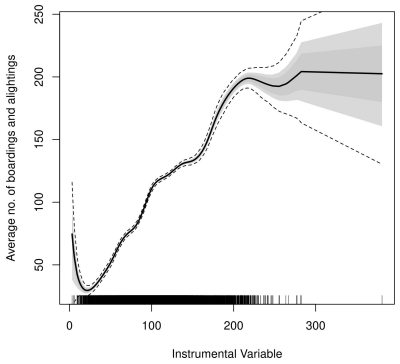
(e) Prince Edward Station (f) Mong Kok Station (g) Yau Ma Tei station

Figure D.7: Distribution of errors from analyses of train movements in the upward direction along the Kwun Tong Line.

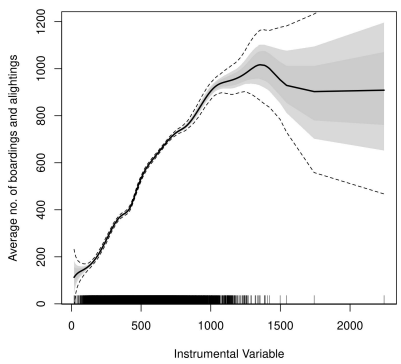
D.5 Strength of Instruments



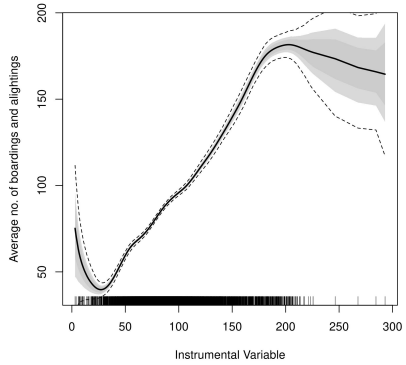
(a) Wong Tai Sin Station



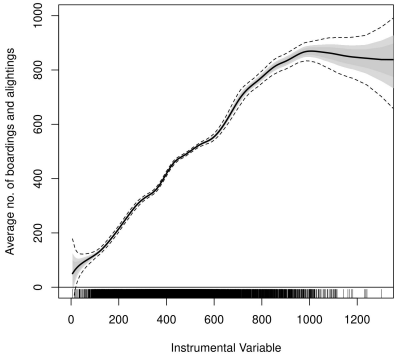
(b) Lok Fu Station



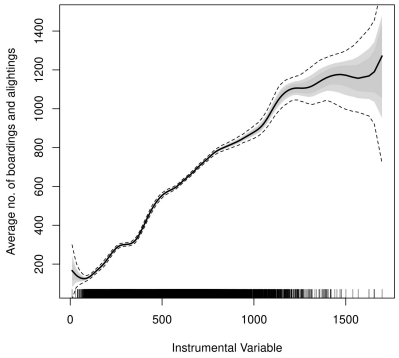
(c) Kowloon Tong Station



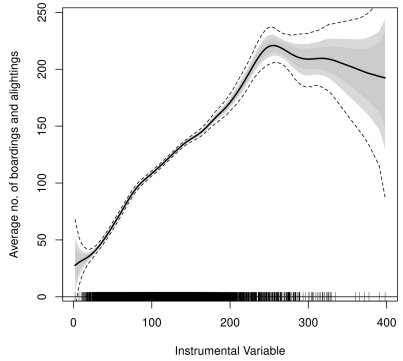
(d) Shek Kip Mei Station



(e) Prince Edward Station

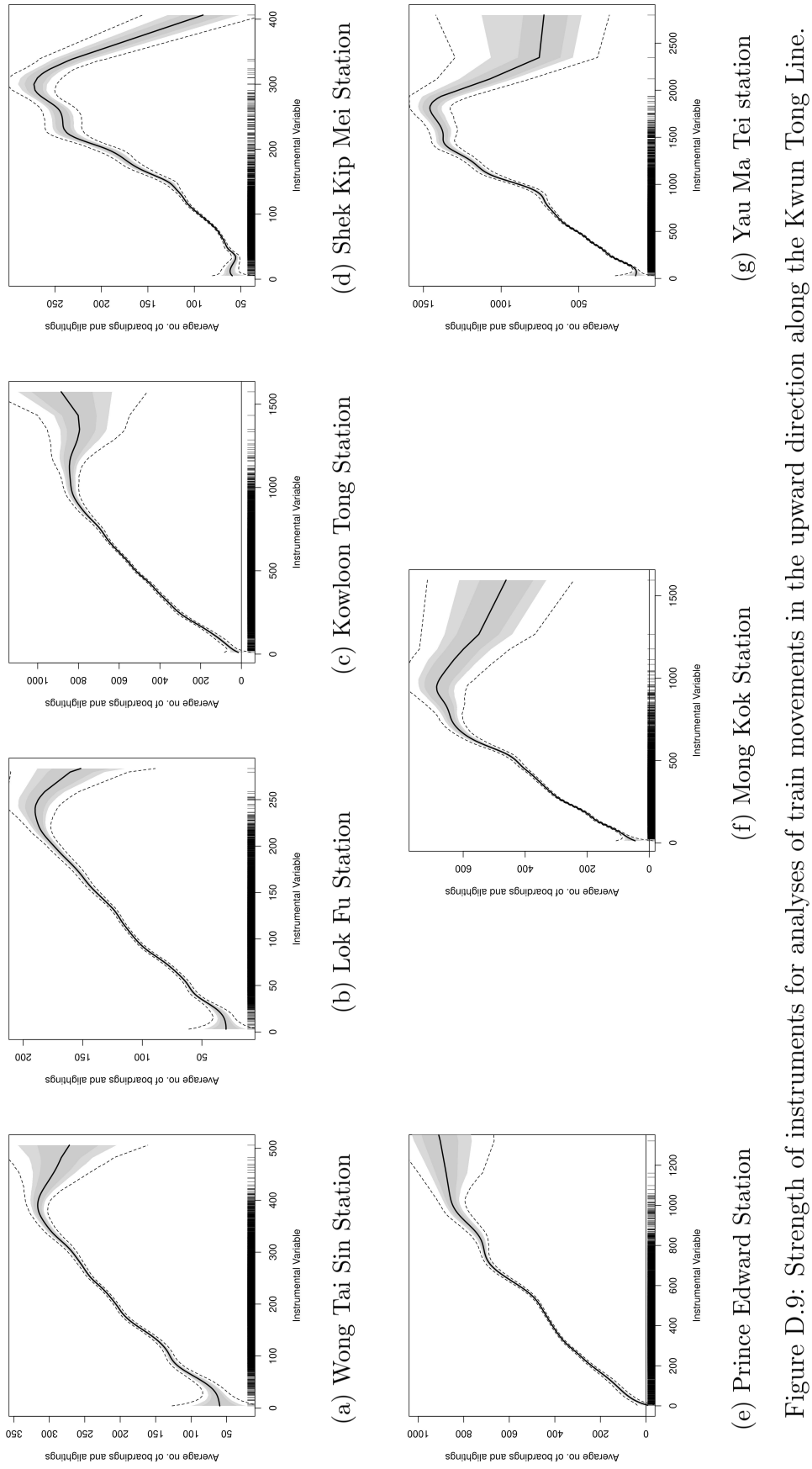


(f) Mong Kok Station



(g) Yau Ma Tei station

Figure D.8: Strength of instruments for analyses of train movements in the downward direction along the Kwun Tong Line.



(a) Wong Tai Sin Station (b) Lok Fu Station (c) Kowloon Tong Station (d) Shek Kip Mei Station

(e) Prince Edward Station (f) Mong Kok Station (g) Yau Ma Tei station

Figure D.9: Strength of instruments for analyses of train movements in the upward direction along the Kwun Tong Line.