Title
Aspatial models of zone pricing and parking

Author
Lehe, Lewis

Publication Date
2016

Peer reviewed|Thesis/dissertation

Aspatial models of zone pricing and parking

by

Lewis Lehe

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering — Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Carlos Daganzo, Chair
Professor Michael Cassidy
Assistant Professor Victor Couture

Fall 2016

Aspatial models of zone pricing and parking

Abstract

Aspatial models of zone pricing and parking

by

Lewis Lehe

Doctor of Philosophy in Engineering — Civil and Environmental Engineering

University of California, Berkeley

Professor Carlos Daganzo, Chair

It is often practical to collapse spatial information about a transportation system into aggregate variables related by an *aspatial* model—that is, a model in which individual interactions and directions of travel are not explicitly accounted for. This dissertation brings aspatial modelling to bear on two topics: downtown congestion pricing, referred to as "zone pricing," and parking policy.

Regarding zone pricing, a survey of the history of zone pricing shows that all existing systems fail to toll vehicles according to their use of the downtown network. To explore whether it would be advantageous to charge higher tolls to travelers who travel farther within the network, a static traffic model with probabalistic choice and variable trip lengths is proposed. Distance-based tolling turns out to be more socially efficient than charging all travelers the same price, but it leaves drivers themselves worse off since most of the welfare gains are converted to toll revenues for the government.

Regarding parking, the thesis considers the idea of a "feedback loop" among landowners' uncoordinated decisions about how much off-street parking to provide on their parcels, when parking competes with floorspace as a use of land. One landowners' decision affects others' via two opposing channels. First, crowding: when there is too little off-street parking, on-street parking becomes crowded, making off-street parking relatively more valuable. Second, accessibility: when floorspace takes the place of off-street parking, the neighborhood becomes walkable through its higher density, making floorspace more valuable. A stylized model of development decision-making shows that each force leads to positive or negative feedback, resulting in multiple equilibria and counterintuitive results from policy.

Contents

To my parents, Jim and Linda.

Acknowledgments

I arrived in Berkeley four years ago. My experience in graduate school could not have gone any better. The department, Carlos Daganzo and UC CONNECT have given me incredible freedom to pursue the research contained in this thesis as well as work on three papers with no imprint here. I am also grateful to University of Leeds for giving me the scholarship that got me into the transportation field before coming to Berkeley. My friends also deserve a mention for supporting me outside of school: Victor Powell, Jacob Brunner, Nicole Johnson, Jacqueline Moreno and others. I have been further blessed to work alongside some other great students that made the journey more interesting and fun: Nathalie Saade, Juan Argote, Jean Doig, Tim Braithwaite being among the most important.

Chapter 1

Introduction

When ecologists analyze an ecosystem, they do not try to predict which wolves will eat which caribou or how many offspring each particular doe will bear. Rather, they build models of the system in which a few, aggregate variables such as population size and food stocks can be related in logical ways from measurements. Likewise, as the name *macro*economics suggests, economists who study large labor markets do not make lists of all the workers and firms in a region or solicit information on who would be willing to work and hire and under what circumstances. Rather, they build aggregate-level search and supply/demand models of the market as a whole with abstract quantities such as the unemployment rate.

By contrast, much of transportation engineering consists of looking at particular roads and bridges and transit lines, and of estimating travel patterns with individual surveys. There is a good reason for this: unlike the ecologist or the the labor market economist, traffic engineers actually participate in designing and building the systems they analyze; they contribute to them piece-by-piece with particular expansions and upgrades.

But for the academic researcher of transportation systems, there is value in aggregation. In the first place, to obtain rules-of-thumb that can guide major decisions, an analyst needs to take a "big picture" perspective. In the second place, transportation systems—like markets and ecosystems—often exhibit gestalt properties at the level of the system that are hard to see in fine-grained analysis of particular pieces. For example, relieving a bottleneck at a freeway exit pours cars into city streets, causing new congestion miles away if the streets cannot handle such a rapid rate of arrivals.

The difference between the two perspectives in transportation can be boiled down to the treatment of *space*. Miroscopic analysis of transportation systems, such as microsimulators of traffic or crowds, explicitly model interactions among agents and vehicles in space—on particular links or among particular vehicles. Macroscopic analysis trades in explicit spatial representation for statistics that approximate the system-scale results of individuals' spatial interactions. That is one reason this dissertation prefers the term *aspatial* to the more common term *macroscopic*. In traffic theory, the term-of-art "macroscopic" sometimes refers to relations among vehicles on a single link pointing in a single direction.

This dissertation is concerned with two aspatial applications of transportation theory.

The first application is downtown congestion pricing, which is referred to as "zone pricing." These are tolling systems that charge drivers for access to an entire area of the city (its downtown), rather than particular links like road and bridge tolls do. The second is parking policy. The applications each serve their special purposes of course; the chapters on each application have something important to say about policy and outcomes. But together they serve a broader purpose, which is to illustrate the usefulness of aspatial modelling to the clear understanding of transportation in cities.

1.1 Zone pricing

Chapters 2-4, the majority of the work, are devoted to zone pricing. Chapter 2, "Zone pricing in practice," consists of a large-scale review of how zone pricing actually works in the real world. Written as a hisory, the chapter describes the politics, designs and results of systems in Singapore, Hong Kong, London, Stockholm, Milan and Gothenburg (Sweden). Chapter 3, "Review of the literature," provides context for Chapter 4. It consists of a literature review of the theories of relevant traffic flow and economic traffic models. Chapter 4, "Self-selection and downtown congestion pricing," proposes an original economic model of commuting into a downton zone, and is based on a paper with the same title that has been submitted to *Transportation Research Part B*. The goal of the chapter is to analyze the difference between a "trip toll," levied on the mere act of making a trip of any length into the downtown, and a "distance toll," levied on distance traveled within the downtown.

1.2 Parking

Chapter 5, "Feedback and the use of land for parking," is based on a forthcoming paper, Lehe (2017), of the same name. The central idea is to show, by way of aspatial models of real estate development, how land-use equilibria in an urban neighborhood may be subject to "feedback" with regard to the amount of parking landowners provide. Feedback here means the way in which neighbors' decisions of how much parking to provide influence future decisions about how much parking to provide, and hence generate complex dynamics for the path of the neighborhood's built environment. The model is aspatial because what matters to decision-making is not the proximity of a certain parcel of land to others' off-street parking, but rather the total amount of off-street parking in the neighborhood as a whole.

Chapter 2

Zone pricing in practice

There exists a considerable theoretical literature on downtown congestion pricing, but, in practice, special considerations wind up altering the design of schemes in ways that are hard to model in theory. One example is that the boundary of the Area C congestion charge in Milan follows the path of old fortifications. Another is that the price of the current London Congestion Charge (£11.50) was chosen, partly, "to ensure the charge level is clear and memoriable to customers by rounding to the nearest 50p."[1] These are not considerations that a theory paper would usually take into account.

To keep theory grounded in practice, this chapter reviews the history of the practice of zone pricing. Its organization is historical. Section 2.1 covers developments that led up to the enactment of zone pricing. Afterward, each section looks in detail at a real scheme that was either enacted or, in the case of Hong Kong, trialed and cancelled. These schemes are: Singapore's Area License Scheme (1975-1997), Hong Kong's Electronic Road Pricing (trialed 1983-1985), Singapore's Electronic Road Pricing (1997-present), the London Congestion Charge (2003-present), Milan's Area C and Ecopass systems (2008-present), the Stockholm Congestion Tax (2006-present) and the Gothenburg Congestion Tax (2012-present). The chapter is most similar in scope to Gómez-Ibáñez and Small (1994) and Anas and Lindsey (2011), which both review a number of downtown pricing schemes, but it incorporates information on many more schemes than the former reference and provides more detail than the latter.

2.1 Early steps toward zone pricing

The first concrete proposal for zone pricing is probably a 1959 testimony to the US Congress on the state of traffic in Washington, D.C., by the economist William Vickrey (Vickrey, 1959). Vickrey advocated that radio transmitters in cars broadcast to receivers embedded in roadbeds, and computers linked to those receivers would then tally charges that could be varied, in proportion to demand, by time and place. Vickrey (who studied electrical

[1] https://consultations.tfl.gov.uk/roads/cc-changes-march-2014/user_uploads/cc-impact-assessment.pdf

engineering as an undergraduate) even built a working prototype of the system. He installed with a transmitter in his car and a receiver/computer combination in his driveway, and would print out a record of his own comings-and-goings to anyone skeptical of the proposal (Harstad, 2005).

Two decades would pass before a city (Hong Kong) considered anything so targeted as Vickrey's scheme. But the idea garnered attention almost right away in the United Kingdom, where the Ministry of Transport convened a committee to study road pricing in cities. Its report, Ministry of Transport (1964), is known as the "Smeed Report" for the committee's chairman, Reuben Smeed of the Road Research Laboratory; and most of its pages are devoted to ingenius techniques for zone pricing using the technologies of the era.

An example of one especially remarkable concept will convey the type of thinking in the Report. According to this proposal, special timers would be mounted outside of vehicles to count how long the vehicle spends in color-coded zones of a city. The counting mechanism in the timers come purchased with a fixed budget of money, which can be replenished for a fee at service stations or post offices. Because installing roadside communication infrastructure would not have been practical at the time, it is incumbent on the driver to set the timer to different as she drives. For instance, when she passes into London's "pink zone," she sets it to the pink zone setting—whereupon it counts down the timer's budget at the pink zone's predetermined rate. The zones in cities across Britain are to be colored and divided up into zones such that the most in-demand areas deplete the timer's budget most quickly. To ensure the driver has set the timer honestly, the timer shines a light (Recall it is mounted outside.) of the color it is set to, and wardens stationed by the side of the road record the number plates of vehicles with timers set incorrectly.

But in spite of its detail (there are even expense estimates for the mechanisms involved), none of the zone pricing proposals from the Smeed Report were adopted in Britain. A later Ministry of Transport report with input from traffic wardens (Ministry of Transport, 1967) judged the report's schemes to be as yet impractical, with the exception of higher parking charges in downtown areas. Nontheless, the Reports principles and practical bent would prove to be influential.

2.2 Singapore — Area License Scheme

2.2.1 History

Singapore implemented the first zone pricing system, the Area License Scheme (ALS), in 1975. The impetus was rising congestion in Singapore's central business district, which followed from a profound and growing mismatch between demand and infrastructure: between 1960 and 1970, the private vehicle population of Singapore doubled, but the total length of public roads rose only 35 percent (Santos et al., 2004, p.211-212). Consequently, just prior to ALS, downtown traffic during the morning rush moved at an estimated 27 kph (Watson and Holland, 1978, p. 66). While 27 kph is respectable for the CBD of an international

city like Singapore at rush hour, forecasts predicted the situation would worsen. The Singapore Concept Plan of 1971—a land use and transportation plan produced with help from the World Bank—involved extensive traffic demand modelling, and it called for a massive program of road building and the establishment of the nation's first rail transit network. Work on new arterials and a network of expressways began immediately, thought the Mass Rapid Transit rail system would not open until 1985. But since the Concept Plan's models suggested that infrastructure alone would prove insufficient, the government also organized a Road Traffic Action Committee (RTAC) to study demand-side solutions (Gómez-Ibáñez and Small, 1994). The RTAC produced a menu of policies. These included consolidating the island's many independent bus services into a network, introducing commuter buses and encouraging staggered work hours (Chor, 1998). But the most most powerful of the RTAC's proposals were ways to make driving expensive: (i) taxes on car purchases, (ii) increased parking rates in the CBD and (iii) a "supplementary license" scheme along the lines of a design recommended for its practicality in the Smeed Report. The license plan became reality as the Area License Scheme on June 3, 1975.

2.2.2 Design

ALS was administratively simple. At post offices, gas stations, convenience stores and roadside booths located along roads leading downtown, a driver paid S$3 to buy a daily "license", or S$60 for a monthly one, that permitted entry to a 6.2 km^2 Restricted Zone (RZ), Monday through Saturday morning (Gómez-Ibáñez and Small, 1994, p. 14).[2] The "license" was a paper decal that a driver would placed in the windshield—like those used for event parking in the US today. For enforcement, wardens standing at 22 access points to the RZ inspected passing vehicles and wrote down the plate numbers of those lacking licenses.

The Singapore experience with zone pricing is distinguished by its administrators' constant willingness to adapt and tinker with every feature of the scheme's design in response to data. Refinements continued from 1975 until the system's replacement in 1997 with an electronic system, described later. At the launch, the license was only required for private vehicles with fewer than three passengers; taxis, delivery vehicles, public vehicles, motorcycles, carpools and buses were exempt. The taxi exemption, however, was ended within three weeks due to a surge in taxi traffic. Also, while the charging period was 7:30-9:30 AM at the launch on June 3, authorities noticed a surge of traffic just after 9:30 AM, and so charging was extended to 10:15 AM on August 1, 1975. ALS was limited to an inbound morning charging period, because it was hoped that curbuing in-bound traffic in the morning would dampen the flow of out-bound trips in the evening (Gómez-Ibáñez and Small, 1994). Moreover, an outbound evening toll would require the operation of license sales booths inside the downtown, adding to the system's costs and complexity. In 1976, the price of a license was increased to S$4 and a double rate charged to registered company cars, because firms

[2]This charge is the only one levied on Saturdays, because the Singaporean work-week extends to mid-day Saturday.

were able to deduct the ALS as a business expense. In 1977, charges for taxis were cut to S$2, because it was difficult to find a taxi downtown. In 1980, the standard license rose again to S$5 (S$10 for company cars). Throughout the 1980s, the boundaries of the RZ were expanded to enclose new real estate developments. The year 1989 saw a package of reforms: first, carpools, goods vehicles and motorcycles lost their exempt status (motorcycles would pay only S$2); second, ALS added an evening charging period. Note that the evening period was not an outbound charge, and a single license served for travel in both periods of the same day. The last major change occured in January 1994, when a S$2 charge was added for entry from 10:15 AM to 4:30 PM (Phang and Toh, 2004). A vehicle entering the RZ in either or both peaks needed a a Whole Day license (also good between the peaks), while a driver who entered only between the peaks only needed the Part Day license.

Three complementary policies supported the launch of ALS. These were (i) the doubling of downtown parking rates (by fiat for public parking and by taxation for private ones), (ii) a massive Park-and-Ride scheme and (iii) a pair of high-quality commuter bus services (Watson and Holland, 1978, pp. 24-25). However, the Park-and-Ride scheme was so little used as to be shut down within the first few months of operation.

2.2.3 Results

The immediate result of implementing ALS was a sharp fall in entries to the RZ. Between March and October of 1975, entries during the 7:30-10:15AM charging interval fell by 44%—well beyond RTAC's desired 25-30% reduction (Watson and Holland, 1978). Speed results appear in Table 2.1, although a lack of reliable measurements means that pre-charging speeds are estimates. Commute trips into the zone primarily switched to bus and carpool, but substantial rescheduling to earlier times of day was also observed. Travelers who had traversed the RZ in the morning en route to destinations outside tended to switch to a ring road.

In spite of the substantial changes in flows and behavior, a household survey of travel times conducted just after the launch yielded disappointing results. There was almost no effect on traffic in the charge-free evening peak; and, due to the fall in speeds on the ring-road and mode shifting to slower modes such as bus and carpool, average journey times actually worsened in the short term.

An important point about the launch of ALS is that the initial price of a license was not decided by detailed demand or traffic studies (Watson and Holland, 1978). Rather, authorities simply noted that the rate of entries to the RZ was about 25-30% lower between the peaks, when traffic was considered acceptable, and estimated that a S$3 charge would discourage about this proportion of trips. Since the actual result exceeded expectations so drastically, some observers—including Wilson (1988); McCarthy and Tay (1993) and Watson and Holland (1978)—have concluded that, at least initially, charges were set too high.

The reforms of 1989—the end of myriad exemptions, a price cut for cars and taxis and introduction of an evening charging period—yielded large results. Between May 1989 and May 1990, traffic composition in the morning charging period rebalanced as a rise and entries

	before ALS (kph estimated)	after ALS (kph observed)
Restricted Zone	27	33
inbound radials	29	32
outbound radials	35	35
ring road	25	20

Table 2.1: Singapore speeds before and after Area License Scheme implementation (Watson and Holland, 1978, p.10)

by car and taxi partly offset roughly 50% falls in entries by truck and motorcycle. In the new evening charging period, entries fell by 54% and exits by 34% (Gómez-Ibáñez and Small, 1994, p. 19). Also, although less data are available for the 1994 introduction of the Part-Day pricing, traffic and congestion fell in the periods just after the morning and just before the evening period (Phang et al., 1997).

2.2.4 Finances

Because of its simple administration, ALS could be said to have the highest financial rate-of-return among all schemes considered. While the capital cost of implementation was S$6.6 million, in fact 95 percent of that cost was sunk into the park-and-ride system; only 5 percent went toward building out the system's infrastructure and publicity Watson and Holland (1978, p. 38). Initial revenues from license sales were S$225,000 per month and operating costs were S$50,000 per month. By 1993, annual revenues were S$47 million, of which operating costs consumed 9 percent (Phang and Toh, 2004). In September 1998, when the scheme ended, annual revenues were about S$100 million (Chin, 2010).

2.3 Hong Kong — Electronic Road Pricing

2.3.1 History

Like Singapore in the early 1970's, Hong Kong in the late 1970's struggled with the traffic consequences of a booming economy and scarcity of space: during the 70's, real per capita income and the number of registered vehicles both doubled, but the aggregate length of roads rose only 17 percent (Hau, 1992). Therefore, in 1979, the government announced its plans for transportation in a White Paper (Branch, 1979), whose recommendations stemmed from the city's First Comprehensive Transport Study (Wilber Smith and Associates, 1976). Like the Singapore Concept Plan of 1971, the White Paper called for new expressways and rail transit, complemented by demand-side measures to make better use of road space. In 1982, while infrastructure work progressed quickly, the government adopted its demand-side package: it

(i) tripled the annual license fees for private cars, (ii) doubled the initial registration fees such as to reach 70-90 percent of the purchase price and (iii) doubled the gas (petrol) tax.[3]

These so-called "fiscal" measures succeeded in reducing car ownership, but their severity and bluntness renewed interest in road pricing among leaders (Gómez-Ibáñez and Small, 1994, pp. 22-25). Therefore, in March 1983, the government contracted with consultants from Britain's Road Research Laboratory—origin of the Smeed Report—to design and test a zone pricing system capable of precisely targeted congestion in space and time. More complex than ALS, the system—called Electronic Road Pricing (ERP)—would rely on wireless communication to levy time-varying charges on traversals of several downtown cordons. The ERP study lasted from July 1983 to March 1985 and involved two main avenues of investigation. On the strategy side, consultants carried out modelling studies to compare various cordon locations and charging structures. On the technological side, field trials involving 2,500 equipped vehicles and 18 tolling sites showed the technology to be highly reliable (Dawson and Catling, 1986).

Despite succesful tests and modelling, Hong Kong did not adopt ERP when study concluded in 1985. Many factors have been put forth to explain the scheme's failure (Hau, 1990; Borins, 1988), but two seem to have been most important. First, the trials ended exactly as outside developments had begun to mitigate congestion on their own: in 1984 a large expressway (the Island Eastern Cooridor) opened; in 1985 a large rail line line (the Island Route) opened; market troubles had weakened the economy; and the fiscal measures of 1982 had begun to curb car ownership. The second major factor was political: in December 1984, Britain had agreed to deliver Hong Kong to China in 1998, prompting worries about spying. ERP's form of administration—a monthly bill listing all of one's vehicle movements—turned out to be highly vulnerable to such concerns.

2.3.2 Design

The Hong Kong ERP was to be a scheme in which vehicles identify themselves to a central system (Dawson and Catling, 1986, pp. 130-131). A solid-state device called an Electronic Number Plate (ENP) (described in the literature as being about the size of a VHS tape) would be attached beneath the vehicle. The ENP had no power supply of its own, but an inductive power loop embedded in the pavement would become energise it as the car passed, whereupon the ENP would transmit data—including an identification number—back to the loops. Attached to the loops was a roadside "cabinet" with computer, which decoded the relevant information and tranmsited it, via modem, to a central control center. And the control center, at the end of every month, mailed users a bill (like a long-distance telephone bill) listing the time, place and times of all cordon traversals. Meanwhile, CCTV cameras at every loop site would record plate numbers to catch violaters.

[3]In US dollars of 2016, the annual registration fee would be $1,100 for a car with a standard engine, and the gas tax would be $1.26 per gallon.

The modelling studies considered three main scenarios for the charging structure and cordon topology (see Gómez-Ibáñez and Small (1994, Table 11, p. 23) or Small and Gomez-Ibanez (1998, Table 10.3, p. 218) for a summary). In all three, charges varied by time-of-day. Scheme A would have five zones with 130 toll sites and would charge the same price in both directions. Scheme B's layout was a geographically simplified version of A's, having only 115 sites, but for certain boundaries it would charge different prices for different directions of travel, depending on the time of day. Finally, Scheme C was the most complex, involving 13 zones, 185 tolling sites and directionally-dependent charging.

There are no real results or finances to speak of for Hong Kong's ERP, since it was not implemented on a large scale.

2.4 Singapore — Electronic Road Pricing (ERP)

2.4.1 History

While cheap to administer in a simple form, a manual system such as ALS becomes intractable if prices vary over space, time and user types. For example, Chin (2010, p. 4) describes the functioning of ALS once multiple charging periods and vehicle classes were added:

> About 60 enforcement personnel and another 60 officers at dedicated licence sales booths were required each day. The enforcement duties were demanding, given the long hours spent under the sun and rain, not to mention the dust and the noise. In addition, there were 16 different types of licences in use at its peak, and much concentration by the enforcement officers was required to ensure that they identified them correctly.

To gain precision without such difficulties, Singapore began making preparations to switch to some form of electronic enforcement as early as the 1980s. In the early 1990s, the government solicited bids for an electronic system using smart cards and In-Vehicle Units [pp. 19-20](Gómez-Ibáñez and Small, 1994). The government chose a smart card/IU system, rather than an AVI design, partly to alleviate the privacy concerns that had bedeviled Hong Kong's ERP (Phang and Toh, 1997). After extensive field trials and publicity, the electronic system—called, like Hong Kong's Electronic Road Pricing (ERP)—went into full operation in September 1998.

2.4.2 Design

Menon and Chin (2004) describes the technology of ERP as being comprised of three components, outlined below:

- The first component is the In-vehicle Unit (IU), a transponder device that sits on vehicles' dashboards or handlebars. The IU has a slot for a "CashCard" that the user

PCU's	Vehicles
0.5	motorcycles
1	cars, taxis, light goods vehicles
1.5	heavy goods vehicles/small buses
2	very heavy goods vehicles/large buses

Table 2.2: Passenger Car Units (PCU's) for ERP. Vehicle charges are weighted by the PCU number—e.g., a very heavy goods vehicle pays four times what a motorcycle does. (Land Transport Authority, 2016)

can load up with money at various places.[4] There are IU's for each of six vehicle classes (see Figure 2.1), because the toll charged to a vehicle is a multiple of its "Passenger Car Unit" (PCU), an index of how much roadspace the vehicle takes up (see Table 2.2).

- The second component is the gantry (see Fig. 2.2). Gantries are positioned in pairs that work together. When a vehicle passes below, the first gantry fires a signal (on a 2.54 GHz radio-frequency band) to the vehicle's IU, telling it to charge the CashCard appropriately. An optical sensor on the second gantry confirms the vehicle is the type attested by the IU. If there is a discrepancy or error, a camera on the first gantry captures the vehicle's rear number plate.

- The third component is the central control system, which verifies charges and issues notices when there is a violations or error.

In contrast to ALS, ERP varies tolls over the course of the day. Figure 2.3 illustrates a weekday toll schedule for the CBD cordon. For the first five years of ERP, tolls changed only at half-hour intervals. But since February 2003, whenever tolls change by more than S$1, there is a five-minute interval in which tolls rise or fall by half the amount of the change, in order to discourage cars from slowing down or speeding up when tolls are about to change (Menon and Chin, 2004). The Land Transport Authority (LTA) of Singapore updates the schedule every three months to maintain speeds of 45-65kph on expressways and 20-30 kph on roads in the RZ, because engineering studies have shown these were the speeds at which flows are maximized (Li, 1999).

The substantial changes to the design of ERP have been spatial. ERP started in 1998 with 33 gantries that approximately reproduced the ALS cordon. Beginning in 1999, the LTA added gantries gradually to enclose the first cordon in a second "Outer Cordon" (Chin, 2010). In 2005, a shopping area on the border of the CBD Cordon became a sub-cordon called the Orchard Road Cordon, because shops open later than offices. The same year,

[4]The CashCard can also be used to pay for parking. In March 2015, a sevice called vCashCard was announced that allows drivers to pay with debit or credit without having a physical CashCard in the IU. (http://www.straitstimes.com/singapore/transport/virtual-cashcard-aims-to-solve-erp-woes)

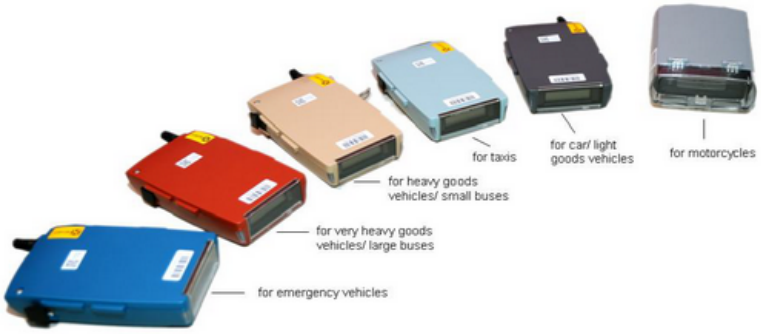

Figure 2.1: ERP In-vehicle units (Land Transport Authority, 2016)

Figure 2.2: ERP gantry (Land Transport Authority, 2016)

gantries were added to charge outbound trips in the evening on one expressway. Finally, in 2008 a line of gantries was planted down the middle of the CBD Cordon. These operate only between 6 and 8PM, when intra-CBD congestion would otherwise be severe. By December 2014, ERP made use of 80 gantries (Chu, 2015, p. 406).

2.4.3 Results

The transition to ERP was not as thoroughly documented as the launch of ALS. One significant effect is that entries to the CBD fell by about 15%, largely due to a decline in repeat trips by the same vehicle (Menon, 2000). Since ALS permitted unlimited same-day entries, under ALS about 23% of trips had been repeat trips—e.g., office workers using cars

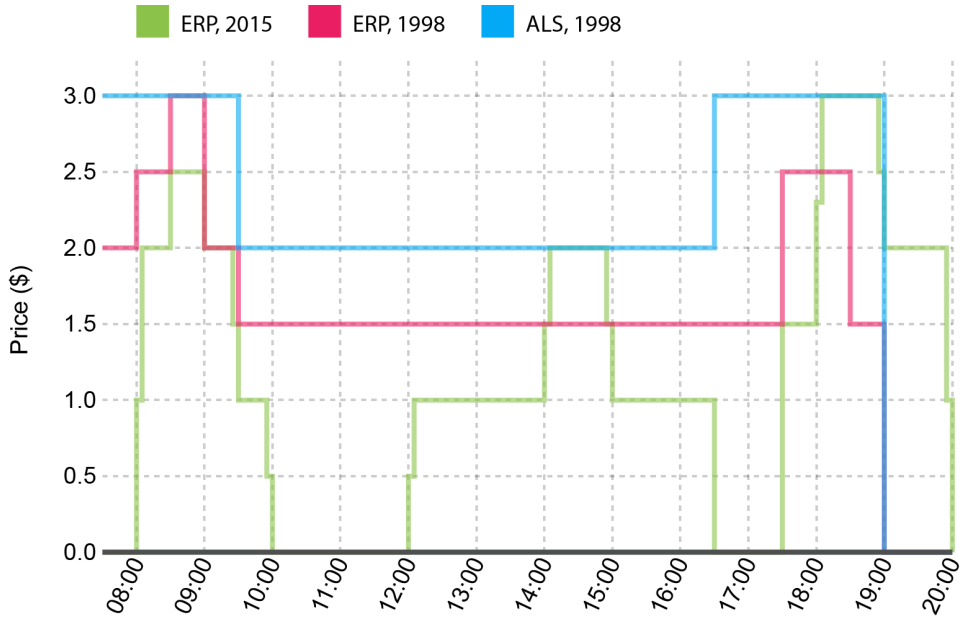

Figure 2.3: Singapore prices for different years. The schedule becomes more variable as years pass. (Land Transport Authority, 2016)

for lunches and meetings in the middle of the day (Chin, 2010, p. 23). Olszewski and Xie (2005) conclude, using data from before and after ERP, that the LTA's charging structure has done a good job controlling congestion and spreading traffic flow over the peaks.

2.4.4 Finances

It would be inaccurate to consider ERP a revenue device. In the first place, ERP's revenues are not hypothecated to particular projects or even transport expenditures generally; they flow into the government's overall budget, of which they form an insignificant fraction. More directly, the switch to ERP actually lost significant revenue. Revenue immediately fell 30%-40%, relative to the S$100 million earned by ALS the year prior.[5] The reason for the revenue loss—as Figure 2.3 shows—is that ERP tolls were cheaper than an ALS license (Chin, 2010, p. 8). Also, the launch of ERP coincided with a one-off rebate on the annual "road tax" that Singaporeans vehicle-owners pay. Singapore does not regularly report revenues, but in 2012 an LTA officer's answer to a question in the national parliament revealed revenues were S$159 million in 2010.[6] This is on par with the revenues from the Gothenburg Congestion Tax, which covers a significantly smaller area. Finally, since the Singaporean government charges very high taxes on gasoline and ERP reduces travel and idling, the overall impact of ERP on revenues is ambiguous.

[5]Note that Chin (2010, p. 8) reports revenue fell 30% while Menon (2000, p. 43) reports 40%.
[6]https://sg.news.yahoo.com/erp-system-collects-about–150-million-each-year.html

Implementation cost S$197 million in 1998, of which S$100 million paid for IU's (at a cost of S$150 each) and S$97 million to build out the infrastructure (e.g., to buy the gantries) (Santos et al., 2004). Annual operating costs have measured about 20-30 percent of revenues.

2.5 London — London Congestion Charge

2.5.1 History

By the time the London Congestion Charge launched in 2003, the study of zone pricing in London had become traditional. The 1960's saw the aforementioend Smeed Report (Ministry of Transport, 1964). In the 1970s, the Greater London Council funded simulation studies into supplementary license schemes, but decided to proceed with parking taxes instead, due to equity concerns (May, 1975). In the late 1980s, the London Planning Advisory Committee commissioned several modelling studies to examine the potential for charging schemes involving several concentric cordons (May and Gardner, 1989). Afterward, from 1991 to 1994, the UK Department of Transport conducted a very elaborate study—or sequence of studies—called the London Congestion Charging Research Programme (MVA Consultancy, 1995; Richards et al., 1996), but results suggested that administration and enforcement would use up too large a share of revenues at reasonable prices, and so plans were shelved for the time being (Small and Gomez-Ibanez, 1998).

Concrete steps toward implementation finally begain in the late 1990's, when Britain's new Labour government undertook to establish a central executive office for Greater London called the Mayor of London (TfL, 2007b, pp.115-120). Among the Mayor's powers would be the ability to implement congestion charging without a referendum, so the government arranged one more research group: the Review of Charging Options for London (ROCOL). In March 2000, ROCOL published its final report, ROCOL (2000), recommending a daily license, which would resemble Singapore's ALS but with two main changes: (i) it would be enforced with cameras and ANPR; (ii) it would apply to all travel within the zone, not merely traversals of a cordon. In May 2000, Ken Livingstone was elected Mayor of London with a campaign promise to consult the public on zone pricing. Following an 18-month consultation in which he softened some aspects of the original ROCOL (2000) plan—offering discounts to residents, eliminating a higher rate for heavy goods vehicles—Livingstone released the Mayor's Transport Strategy in July 2001 with a final description of the charging scheme. After a contracting process and some final revisions, the scheme opened to the public on February 17, 2003.

2.5.2 Design

The LCC launched on February 17, 2003 as a £5 license for travel within the 22 km^2 Charging Zone (CZ) between 7 AM and 6:30 PM on weekdays. The solid region in Figure 2.4 depicts the original and present-day CZ. Since 2003 the LCC has undergone many changes. The

toll was raised from £5 to £8 in July 2005, £10 in January 2011 and £11.50 in June 2014—making it the most expensive scheme. In 2007, the end of charging was moved up from 6:30 to 6:00 PM. In February 2007, TfL added a 19 km^2 area called the "Western Extension" (the striped area in Figure 2.4). But after a consultation showed the public heavily opposed (TfL, 2008), it was abolished in January 2011.

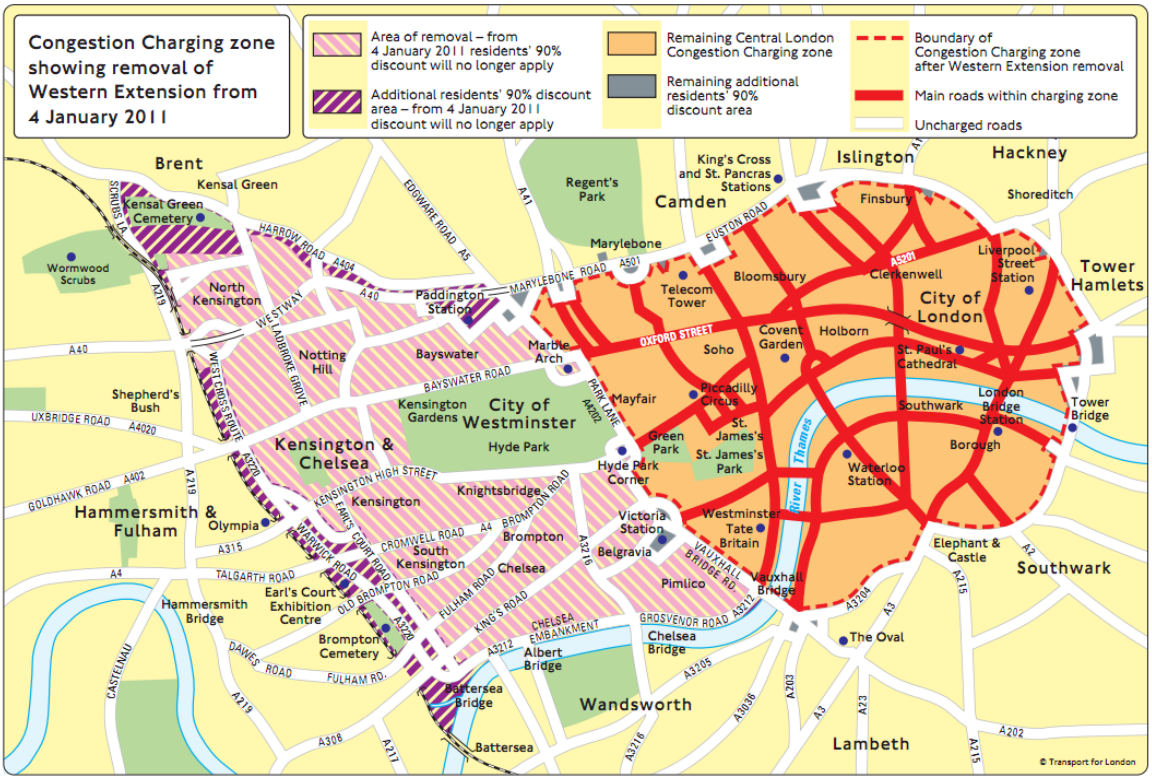

Figure 2.4: London Congestion Charging Zone (Transport for London, 2015)

The LCC's design is distinctive in at least three ways. First, it is the only scheme to charge for all travel, not merely traversals of a cordon. Thus, at the launch enforcement involved 500 cameras at 250 sites throughout the CZ (TfL, 2003, p. 240). Second, transactions are the user's responsbility: a driver without an Auto Pay account must pay online, by phone or in participating stores before midnight on the day of entry to avoid a fine. Third, it has always offered an extraordinary number of discounts and exemptions. Today, these are:

- *Exempt:* motorcycles/mopeds, registered taxis and private hire vehicles, certain government vehicles
- *£1 discount*: Auto Pay users
- *90% discount*: residents of the zone and designated areas contiguous to the zone

- *100% discount:* Blue Badge (a Europe-wide handicap identifier), motor tricycles, vehicles with 9+ seats, breakdown, cars/vans weighing less than 3.5 tonnes that emit 75 or fewer grams of CO2 per kilometer and meet the Euro 5 air quality standard.

2.5.3 Results

The introduction of the LCC led to a sharp fall in entries by private car and a rise in bus ridership. In the first year of operation, entries by private car during charging hours fell 33% (65,000 per day) and entries by all chargeable vehicles—cars, trucks (lorries) and vans—fell 27% (73,000 per day). By contrast, entries by private taxi, which were exempt from charging, jumped 18% (10,000). Data on traffic flos shows a similar patern except that taxi circulation by a proportionately greater amount than did taxi entries, suggeting longer taxi trips. (See Table 2.3 TfL 2007, p. 26).

There was little sign of a switch to subway (metro), but bus ridership grew rapidly. The Central Area Peak Count — whose provence includes the charging zone — rose 18 percent in 2003 and another 12 percent in 2004 (TfL, 2007b, p. 58). Bus ridership into the Charging Zone itself seems to have only been measured in the first year: passenger-entries by bus rose 37 percent, of which increase TfL attributed half to charging and the other have to service improvements.

2.5.4 Finances

Financially, the LCC failed to live up to expectations; it cost more money and raised less than expected. ROCOL (2000) predicted setup costs would be about £30-50 million, annual operating costs in the same range and £230-270 million in annual revenue. More conservatively, at the launch, TfL expected £130 million in *net* revenue per year (TfL, 2003, p. 34). In reality, implementation costs turned out to be £162 million (TfL, 2007b, p. 135), while operating costs were higher and revenues lower than forecast (see Table 2.3). A chief reason for the shortfall was that car entries fell more (30%) than expected (20%) (Leape, 2006, p.169). TfL (2007a) reports that only 40 percent of daily entries pay the full charge. Note that penalties (for the years when such data are available) accounted for substantial revenue. Years 2008-2011 in Table 2.3 are starred to highlight the years of the Western Extension—a time of higher costs and revenues. Net revenues have mainly funded bus service, in accordance with Section 295 of the London Authority Act hypothecating the first ten years of revenues to transport.

year	tolls	penalties	revenue	cost	net revenue
2003	18	1	19	17	2
2004	116	55	171	93	78
2005	117	75	192	90	102
2006	144	66	210	88	122
2007	158	55	213	90	123
2008*	-	-	328	191	137
2009*	-	-	326	177	149
2010*	-	-	313	155	158
2011*	-	-	287	113	174
2012	-	-	227	90	137
2013	-	-	220	88	132
2014	-	-	235	85	149
2015	-	-	257	85	172

Table 2.3: London Congestion Charge costs and revenues. Years 2008-2011 are starred to show when the Wesern Extension was in place. 2003-2008 data from TfL's *Congestion Charging Monitoring* reports. Later data from TfL's *Travel in London* series.

2.6 Stockholm — Stockholm Congestion Tax

2.6.1 History

Like the London Congestion Charge, the Stockholm Congestion Tax, enacted on a permanent basis in 2007, is the last iteration of a long chain of proposals. Its early ancestor was a rule—argued in local politics throughout the 1980's but never enacted—that would require all cars entering downtown Stockholm to display a transit pass (Ramjerdi et al., 2004).[7] In the early 1990's, an electronic cordon toll was made a primary revenue source for the "Dennis Package"—a bundle of infrastructure projects, including tunnels, the completion of a ring-road and a major bypass, which were designed to keep traffic out of central Stockholm. In 1997 the Dennis Package fell apart for various political reasons, but zone pricing's inclusion in such a heavily-debated package had elevated its political visibility, particularly among Sweden's strong environmentalist community.

After the 2002 round of Swedish elections, the Social Democratic Party reached out to the Green Party to join a coallition government, and the Greens made their support contingent on carrying out a trial of zone pricing in Stockholm. The Social Democrats consented, prompting loud complaints from opposition leaders: prior to the election, the Social Democrat candiate for Mayor of Stockholm had stated on television, "My message to the voters of Stocholm is that there will be no road charging during our next term of office, and that is a manifesto pledge on our part." Consequently, the Social Democrats decided it

[7] See Arnott et al. (2005, Ch. 5) for a theoretical analysis of this class of policy.

would be politically expedient to conclude the trial with a popular referendum on whether to make zone pricing permanent, which greatly complicated implementation but laid the groundwork for a very interesting case study in public opinion.

The trial lasted from January 3, 2006 to July 31, 2006 (Eliasson, 2008). To accompany the trial, the City also offered a temporary public transport extension involving new bus lines, extra rail capacity and park-and-ride that lasted August 22, 2005 to December 31, 2006. The referendum on making the Congestion Tax permanent occured in September 2006 as part of the Swedish national election cycle, and 53% of eligible votes were in favor. This was a surprise, because public opinion and media coverage prior to the trial had been strongly negative. During the same time, however, the Social Democrats lost control of the government to a new center-right government composed of the parties who had opposed charging. In the end, the new government agreed to accept the results of the referendum after negotiating a €10 billion infrastructure package, in which toll revenues would be matched by national funds to build new roads and transit.

2.6.2 Design

The Stockholm Congestion Tax is a time-variable toll on weekday trips in either direction across a cordon around central Stockholm. Crossings in both directions pay the same toll. See Figure 2.5 for a map of the approximately 35 km^2 charging zone. Note from the map that water boundaries require only 18 access points to establish the cordon. Vehicles are charged each time they pass under gantries, which function in sets of three: when a laser on the middle gantry detects a vehicle passing below, cameras mounted on the first and third photograph the front and rear plates (Transportstyrelsen, 2015b). Foreign, emergency, diplomatic, alternative-fuel and foreign-registered vehicles were exempt at first, but the alternative-fuel exemption was ended in August 2012.

2.6.3 Results

The Tax has cut both traffic flow and travel times. Traversals of the cordon have fallen by 20-21%. This figure was in excess of the forecast 16% fall, due to the fact off-peak travel fell by almost the same amount (21%) as peak travel (20%) (Eliasson et al., 2013). Models had predicted off-peak travel would fall only 14% due to lower mid-day charges. The fall in person-trips was similar for commuters (24%) and discretionary travelers (22%), but the two groups adapted differently: commuters who stopped driving switched to public transit, while discretionary travelers cancelled or rerouted their trips (Franklin et al., 2009). Eliasson et al. (2013) estimate that the Tax raised public transit ridership 4-5%, or 8-10 thousand riders per day, and that vehicle-kilometers-traveled within the zone fell 10-15 percent. Karlström and Franklin (2009) find only weak evidence of departure time rescheduling.

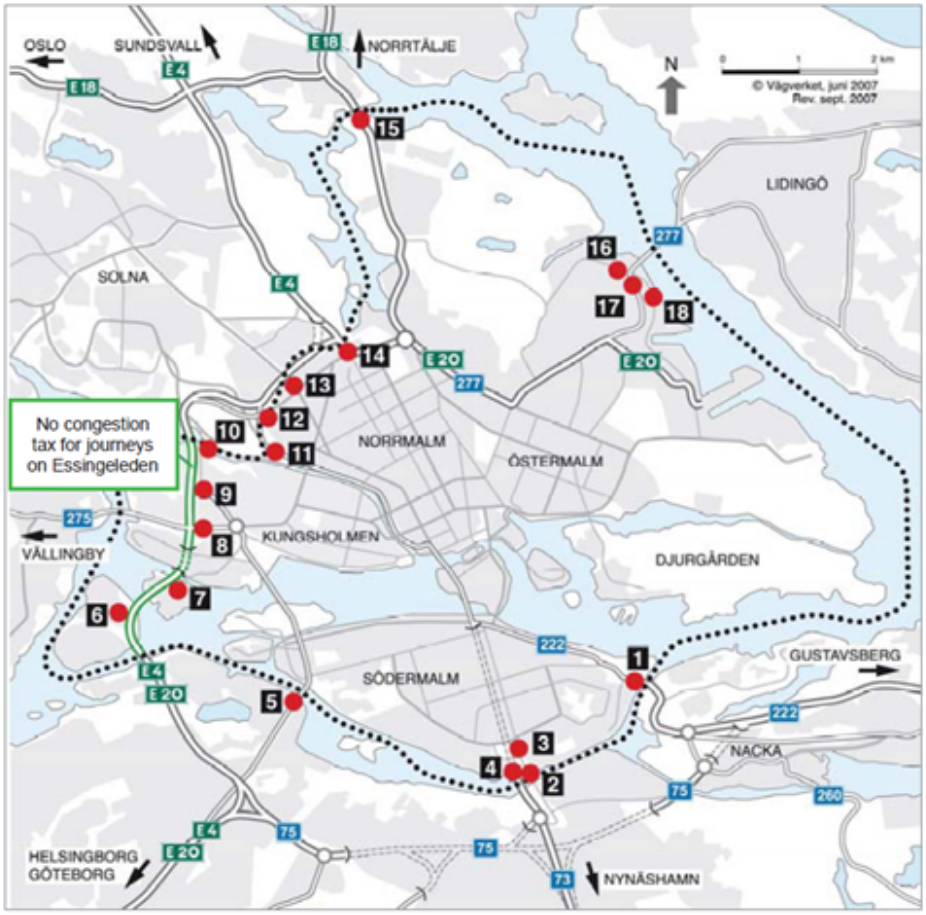

Figure 2.5: Stockholm Congestion Tax, access points in red. (Transportstyrelsen, 2015a)

2.7 Milan — Area C

2.7.1 History

The next city to adopt zone pricing was Milan. The Milan experience depends critically on two pieces of context. First, Milan — like several other Italin cities — operates a number of "limited traffic zones," or "zonas trafico limitado" (ZTL's). These are areas, often historic, where vehicle access is restricted by time and purpose. A ZTL, for instance, might ban private vehicles during the working day except for the zone's residents. Some are enforced by a gate, but others use cameras to identify entrants' plates, issuing fines to catch violators. Thus, Milan required little new infrastructure. Moreoever, the driving public and the city officials were acquainted with the idea of controlled access. The second piece of context is the region's problem with air pollution. Milan has among the highest rates of car ownership in Europe, and sits in the relatively windless Po Valley. Consequently, between 2002 and 2011,

the city exceeded European Union standards for PM10—a fine form of particulate matter pollution—an average of 133 days per year (Danielis et al., 2011).

The Milan zone pricing schemes have applied to the Cherchia dei Bastioni, a ring of 16th century fortifications around the city center. During his 2002-2007 term, Mayor Gabriele Albertini discussed some form of charging for access to this area (Mattioli et al., 2012). But it was Albertini's predecessor, Letizia Moratti, who took up the cause in earnest following her election in 2006. Consequently, an ANPR-enforced daily license called "Ecopass" was implemented January 1, 2008 with the specific goal of curbing access by high-emission vehicles.

Ecopass had a complex charging structure (see Table 2.4) and liberal exemptions that made entry free for many vehicles. The share of chargeable vehicles entering the zone fell from 42 percent before Ecopass to just 16 percent by 2009 (Danielis et al., 2011, p. 5, Table 3). Observers believed Ecopass was not having its advertised effect, particularly since air pollution worsed in early 2010 (Mattioli et al., 2012). In 2010, activists helped to organize a petition drive for a number of environmental and transportation referenda, including a strengthening of Ecopass. Voting took place in July 2011, and all referenda passed by large margins: support was 79 percent for the Ecopass one. Around the same time, Moratti was replaced as mayor by Giuliano Pisapia, who set about remaking Ecopass. The reorganized system, called Area C, launched in January 1, 2012 with a simpler toll structure (See Table 2.5) and a mission more weighted toward congestion reduction. Area C continues in use today, although a court order temporarily suspended the program from July 25 to September 17, 2011—creating a natural experiment that has since been used for studies such as Gibson and Carnovale (2015) and Percoco (2014).

2.7.2 Design

The charging zone of both Area C is an 8 km^2 area of central Milan called the Cerchia dei Bastioni (see Figure 2.6). The cordon consists of 43 access points where cameras read the number plates of entering vehicles. Charging operates from 7:30 AM and 7:30 PM on weekdays (Muncipality of Milan, 2015). The standard way to pay is to buy a digital "ticket" at banks, parking meters, online, ATM's or in stores and then "activate it"—that is, associate it with a plate number on a particular day—by phone, SMS, online or at municipal offices. Since the switch to Area C, users can also sign up for a Telepass radio-frequency transponder to pay by debit automatically. What most distinguishes Ecopass from Area C is the charging structure: Ecopass involved a schedule of emission classes , while Area C involves a schedule of user classes. See Tables 2.4 and 2.5 below. Both systems have also come with an enormous number of exemptions, including motorcycles and scooters, a certain number of free days for residents and vehicles delivering perishable and refrigerated food Danielis et al. (2011). Ecopass also offered residents of the zone a 50% discount on their first 50 entries per year, and a 40% discount on the next fifty.

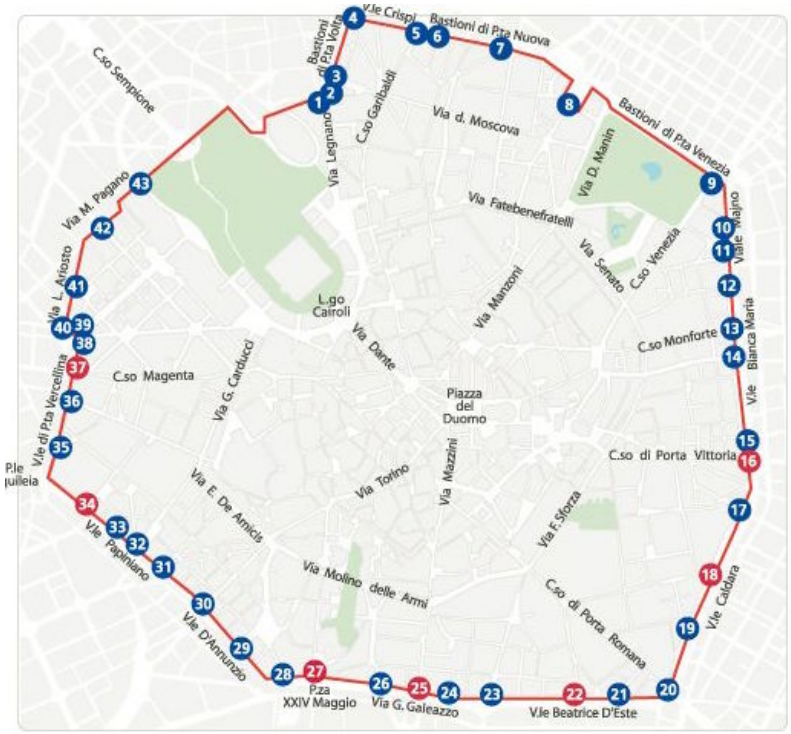

Figure 2.6: Milan Ecopass/Area C (Rotaris et al., 2010)

Emissions Class	Charge (€)
1	0
2	0
3	2
4	5
5	10

Table 2.4: Ecopass prices. Lower classes are less polluting. Class I includes hybrid and electric cars. Class V low-efficiency diesel and buses. (Rotaris et al., 2010)

2.7.3 Results

During the Ecopass trial, congestion inside the cordon fell by 12.3%, vehicle-kilometers traveled by 14.2% and accidents by 20.6% while bus speeds and private vehicle speeds rose, respectively, 7.8% and 4% (Rotaris et al., 2010). Using time-series data during the suspension in summer 2012, Gibson and Carnovale (2015) estimate that the suspension raised entries to the charging zone during charging hours by 27,500 per day (14.5 percent), CO concentrations

User Class	Charge (€)
standard	5
residents	2
commercial	3

Table 2.5: Area C prices (Muncipality of Milan, 2015)

by 6 percent and PM10 concentrations by 17 percent.

2.8 Gothenburg — Gothenburg Congestion Tax

2.8.1 History

The most recent place to adopt zone pricing has been Gothenburg—Sweden's second largest city. The infrastructure deal that Stockholm obtained from its Congestion Tax attracted the attention of Gothenburg's leadership, who in 2009 negotiated a 3.4 billion EUR package funded partly by a zone pricing scheme (Börjesson and Kristoffersson, 2015). The outcome of that deal, the Gothenburg Congestion Tax, launched in January 2013. In September 2013, a non-binding public consultation showed 57% of Gothenburg voters opposed the Tax, but the scheme has been kept in order to fulfill the terms of the infrastructure agreement. Börjesson and Kristoffersson (2015) attribute the unpopularity of the Congestion Tax to revelations that the infrastructure it purchases—especially a massive underground rail tunnel—will be significantly less beneficial and more expensive than advertised.

2.8.2 Design

Gothenburg's scheme was based on Stockholm's. Cameras identify drivers crossing in either direction a cordon around the city center (see Figure 2.7). The toll schedule has the same structure as in Stockholm, though shifts occur 30 minutes earlier, and rates differ slightly at certain times.

As a city, Gothenburg differs critically from Stockholm. Gothenburg is less than half the size of Stockholm and has less transit usage: in 2012 public transit accounted for 26% of trips among OD pairs involving the congestion tax in Gothenburg, while the statistic for Stockholm is 77% (Börjesson and Kristoffersson, 2015). Also, since the CBD is not on a penninsula, Gothenburg has 38 gantries to Stockholm's 18, and several must be situated in residential neighborhoods in order to prevent traffic from diverting onto quiet streets. The "single charge rule" states that, no matter how many times a vehicle traverses the cordon within 60 minutes, it is charged only once (Transportstyrelsen, 2015a). In this case, the amount of the toll is the highest among the applicable traversals.

Figure 2.7: Gothenburg Congestion Tax zone (Transportstyrelsen, 2015a)

2.8.3 Results

The effects in Gothenburg were qualitatively similar to but milder than those observed in Stockholm. Traffic over the cordon during the charged hours fell 12%, rather than the 15% forecast by models (Börjesson and Kristoffersson, 2015). The discrepancy was largest during the peaks: whereas models had predicted peak travel would fall 18%, it fell only 13%—approximately the same reduction observed for off-peak traffic. Travel surveys show that commuters switched to public transport, while discretionary travelers traveled less frequently or switched destinations. Accounting for external factors, the charge is estimated to have raised public transport ridership by about 4.5-6.5%. As for congestion, Figure 2.8 shows travel times in Gothenburg for different classes of road from before and after the charge. Likewise, in Stockholm commuters switched to public transport, public transport ridership rose by about 5%, discretionary travelers canceled trips, peak and off-peak traffic fell by similar amounts and travel time reductions were strongest on inner arterial roads.

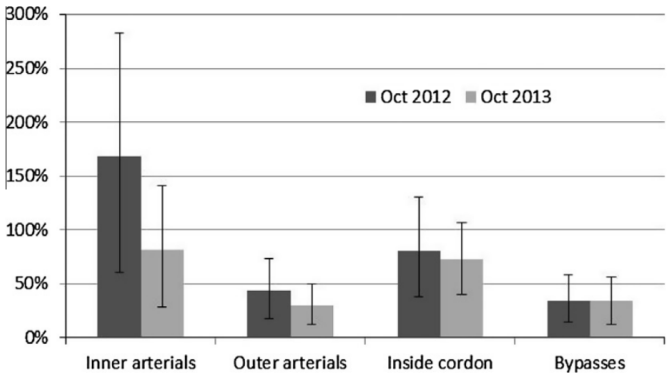

Figure 2.8: Gothenburg 7-8AM. Increase in travel times on selected categories of road relative to free-flow speeds, before (Oct 2012) and after (Oct 2013) the Congestion Tax. (Börjesson and Kristoffersson, 2015)

2.8.4 Finances

In 2013 the Gothenburg Congestion Tax raised 846 million kr, of which 76 million kr came from tolls and 83.8 million kr from penalties (Transportstyrelsen, 2013). This was short of the 930 million kr ($103 million) forecast, largely because 45% of traffic turned out to be exempt rather than the 30% forecast (Börjesson and Kristoffersson, 2015).

2.9 Discussion

This chapter has reviewed the history of zone pricing from its conception as an idea in the 1950's and 1960's to its proliferation in the new millenium as a result of number-plate recognition technology. Surveying the zone pricing experience, two themes emerge that are worth highlighting.

First, note that all the systems surveyed produced similar traffic reductions, between 10 and 20 percent of all entries. This is surprising given that the amount of money charged varies considerably. In Stockholm, the maximum charge is only about $3. In London it was about $18 before the recent devaluation of the pound. Moreoever, the models in Stockholm and Gothenburg were mistaken in that they predicted different traffic reductions in the peak and off-peak, due to the time-varying toll. Instead what was observed is that a similar share of traffic stopped driving at both times of day. As a rule-of-thumb, it might be worthwhile to assume that, in any pre-charging population of drivers, about 10-15% of drivers are simply unwilling to pay anything to drive, almost regardless of the size of the toll.

Second, while most of the theory of congestion pricing focuses on prices, in practice exemptions are critical. Some of these exemptions—such as those for the handicapped or medical vehicles—are easy to justify by appeals to welfare or social justice. But many

exemptions—such as those for taxis or even, in the case of Milan, refrigerated delivery trucks—lack much of a rationale outside politics.

Chapter 3

Review of the literature

Historically, the facts of transportation have been a reliable stimulus to original economic thought—often with implications far beyond transportation. John Stuart Mill offered lighthouses as an instance of a public good that markets cannot provide. To explain how differences in the costs of shipping various crops leads to bands of land rents around towns, Johann von Thunen created what has been called the first economic model. The civil engineer Jules Dupuit is credited with inventing consumer and social surplus in his analysis of what toll to charge on a new bridge.

This chapter traces traffic physics have underlain certain economic models leading up to the model of Chapter 4 in four sections. Section 3.1 describes the development of the orthodox theory of traffic on links. Section 3.2 then reviews "static" economic models meant to analyze pdecisions about whether to drive on links. Section 3.3 returns to traffic theory to describe the evolution of network-level traffic relations. Section 3.4 covers the model in Gonzales (2015). Before beginning, a note on notation: The following conventions and notation will be used throughout this chapter and the next.

- *Units*: Time quantities are denominated in minutes (min), and utility in minutes of travel time savings, which implies everyone values time in the same way. The nouns "traveler" and "vehicle" are synonymous.

- *Trip properties*: There is a confusing difference between how economists and traffic flow theorists describe the start and end of a trip. Economists typically say that vehicles reach a facility at about the time they "depart" from home and leave the facility upon "arriving" at work (or vice versa for the evening commute). On the other hand, specialists in traffic and queues embrace the opposite terminology: vehicles "arrive" at a facility and later "depart." The rule used here is to say vehicles "arrive" at a facility and later "exit" it; this rule is clear, because no one speaks of "exiting" their home in the morning. "Trip length" is the physical length of a trip, while "trip duration" is the amount of time it takes from beginning to end.

- *Flow*: Here, the "exit rate" or "out-flow" of vehicles will be e (veh/min). This is the rate at which vehicles cross the boundary out of the zone in the evening commute or park at their destinations for the morning commute. The "arrival rate" or "in-flow" is a (veh/min); this is the rate at which cars leave their parking spaces and enter city streets during the evening commute or arrive cross the boundary into the zone during the morning commute. Finally, the rate at which cars will pass a observer stationed anywhere along a section of stationary traffic will be called the "circulation" or, generically, "flow" and denoted by q (veh/min).

3.1 Traffic flow on a link

Much of traffic flow theory can be built up from a single observation: after a point, vehicles slow down as they approach one another. Causation could happen in either direction: on the one hand, drivers may want more space when traveling at higher speeds; on the other hand, when cars are close together, drivers might feel safer if they slowed down. In either case, there is a weakly-declining relationship, $v(k)$, between speed v (km/min) and traffic density k (veh/km)—which is the inverse of the average distance between vehicles. Greenshields (1935) was among the first studies to measure $v(k)$, finding a linear speed-flow relation.

The speed-density relation sketches a bigger picture when joined to the Fundamental Identity of Traffic Flow, $q = v \cdot k$ (veh/min). The passing flow, q (veh/min), is the rate at which vehicles pass a point anywhere along a segment of stationary density k and speed v. Greenshields (1935) applies the Fundamental Identity to plot the locus of points (q, v) defined, parametrically, by $[k \cdot v(k), v(k)]$ (See Fig. 3.1). But it is more common to depict the relationship as a single-peaked function, $q(k) \equiv v(k) \cdot k$, called the *Fundamental Diagram* $q(k)$ (see Fig. 3.2). The speed corresponding to any point $(k, q(k))$ on the Fundamental Diagram can be visualized as the slope of a ray from the origin to that point. The maximum q_0 of $q(k)$ is called the *capacity* of the link, and the density k_j for which $q(k_j) = 0$ is the *jam density*. States on the left branch are normally called *uncongested* or *free-flow* (when traffic moves at the permissible limit), while those on the right branch are called *congested* by engineers and *hypercongested* by economists. Congested/hypercongested states can also be identified as points on the upper, faster branch of Greenshields' speed-flow curve in Figure 3.1. When traffic is hypercongested, speed will be observed to rise with flow—although the relationship is not causative.

Different studies have given different shapes to the Fundamental Diagram, but the most popular is triangular. A triangular FD implies traffic moves at a free-flow speed, v_f until a certain critical density, k_0; thereafter, it falls with a value proportional to $1/k$. Cassidy (1998) suggests this shape fits highway traffic well, and Newell (2002) offers a simple car-following model that gives rise to a triangular FD.

In the 1950's, traffic flow theory advanced tremendously through the emergence of the so-called "kinematic wave" theory—also called the Lighthill-Witham-Richards (LWR) model. Lighthill and Whitham (1955) and Richards (1956) treat traffic on a highway like a contin-

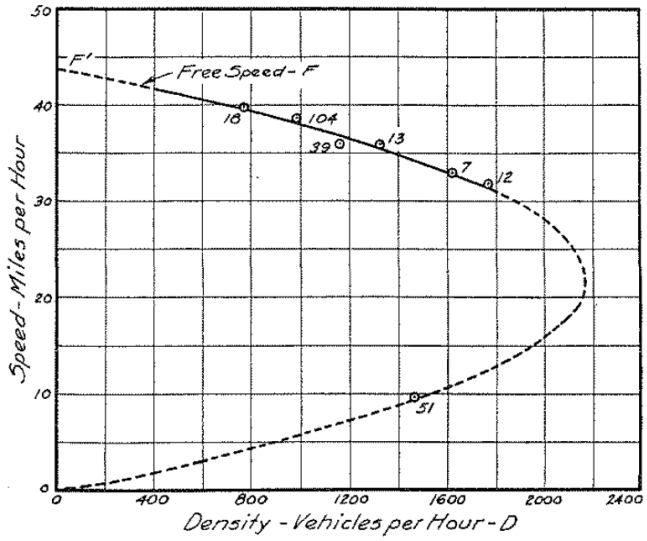

Figure 3.1: Greenshields' proto-Fundamental Diagram comparing speed (vertical axis) against flow across a point (horizontal axis)

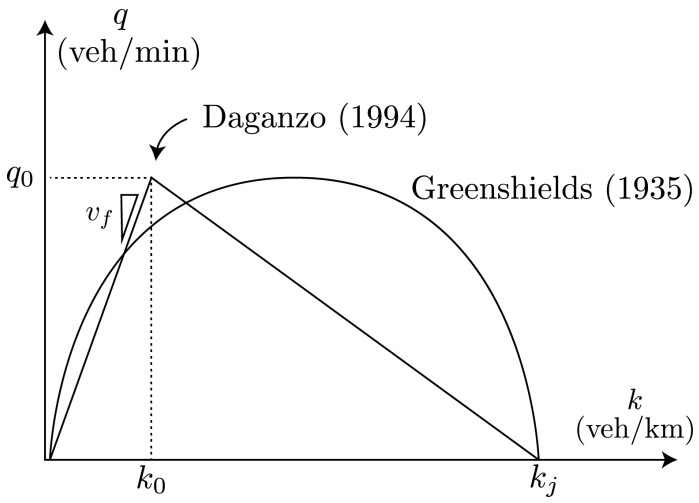

Figure 3.2: Fundamental Diagram, $q(k)$, for traffic on a link, expressed as flow (vertical axis) against density (horizontal axis)

uous fluid with density, flow and speed defined at every point along a road. The number of vehicles in a stretch of highway is analagous to the mass of fluid contained in a segment of pipe. The essential insight is that, given initial conditions (the density at each location along a link) and boundary conditions (e.g., the path of in-flow to the link being modeled), the Fundamental Diagram can be used to find the density, flow and speed at each location at every moment in the future. This is accomplished by recognizing that vehicles are conserved: the instantaneous outflow, $e(t)$, from one segment is instantaneous in-flow, $a(t)$, to the next; and any gap between a segment's a and e is reflected in a rise or fall in its density. These facts lead directly to a conservation equation—the first-order partial differential equation

$$k_t + q_x = 0, . \tag{3.1}$$

(subscripts denote derivatives).

Using the conservation equation, characteristics of a certain density can be traced through space and time. These characteristics represent waves of traffic that roll up and down a road viewed from above. When characteristics intersect, they form shockwaves. Because the process of tracing the characteristics can be tedious, streamlined solution methods have been developed; see e.g., Newell (1993) and the Cell Transmission Model (CTM) of Daganzo (1994), which discretizes the road into a finite number of small segments.

3.2 Static economic models of traffic

In addition to the physics of link, there also exist economic models of traffic on links. These are usually classified as either "static" or "dynamic." Dynamic models are those descended from Vickrey (1969) and involve the decision of *when* to travel. By contrast, in static models the question is whether someone travels and, if so, what route or mode they take. Since Chapter 4 does not involve scheduling, this summary only touches on static modelling.

3.2.1 Pigou (1920)

Many of the enduring insights of static modelling can be found in a single paragraph of Pigou (1920). These insights are: (i) user equilibrium (UE); (ii) social optimum (SO); and (iii) corrective tolling.

> Suppose there are two roads, ABD and ACD both leading from A to D. If left to itself, traffic would be so distributed that the trouble involved in driving a representative cart along each of the two roads would be equal. But, in some circumstances, it would be possible, by shifting a few carts from route B to route C, greatly to lessen the trouble of driving by those still left on B, while only slightly increasing the trouble of driving along C. In these circumstances a rightly chosen measure of differential taxation against road B would create an artificial situation superior to the natural one.

To couch Pigou's example in the language of traffic engineering, we let "carts" be vehicles, "trouble" stand for travel time and allow that a route's travel time is given by a rising function of flow—as in Greenshield's diagram. The first meaningful claim is that "If left to itself, traffic would be so distributed that the trouble involved in driving a representative cart along each of the two roads would be equal," which suggests that self-interested drivers will choose between the two routes until the travel time on both is equal. This is the situation Wardrop (1952) calls a "user equilibrium" (UE), because no individual traveler can do any better by switching routes.

The next meaningful claim in the passage is that, "it would be possible, by shifting a few carts from route B to route C, greatly to lessen the trouble of driving by those still left on B, while only slightly increasing the trouble of driving along C." The point here is that since travel time on each route may respond differently to flow, then there is no guarantee that the flows in the user equilibrium minimize the total travel time of all users. For example, suppose that B is long but completely uncongestible (that is, its travel time is the same even if everyone travels along it), whereas C is short but highly congestible. In this case, it must be possible to improve on the user equilibrium: if we were to move a single "cart" from C to B, then neither the switched cart nor the other users of B would be any worse off, but all the travelers on C would be better off. The split between the routes that minimizes total travel time is another vector of flows that Wardrop (1952) calls the "social optimum" (SO). Later, Beckmann et al. (1956) brought new optimization techniques to bear on finding the UE and SO on a complex network of routes, giving birth to the "traffic assignment" literature.

The last point in the passage concerns "differential taxation," better framed as tolls or a congestion price. When the SO and the UE differ, there are vectors of tolls across the routes that push drivers to choose the SO—thereby minimizing aggregate travel time. In our example with the uncongestible route B and congestible C, the appropriate vector of tolls would privilege B, and encourage travelers to switch to it.

3.2.2 Walters (1961)

Although Pigou's example could be called the first economic traffic model, it is not detailed or abstract enough to be taught or much extended. Rather, the textbook treatment of road pricing comes from Walters (1961). The idea of Walters (1961) is to map the physical quantities of congestion into the vocabulary of conventional economics: traffic flow mirrors the quantity of goods sold, travel time is akin to price, the relation between flow and travel time is a supply curve, and the schedule of traffic flow that travelers desire at different travel times is a demand curve. To formalize this, let $w_s(q)$ give travel time w (min) as a function of flow q (veh/min), and let $w_d(q)$ be an inverse-demand curve. Fig. 3.3 depicts Walter's functions diagramatically like the supply-and-demand diagrams of economics. Note that the function $w_s(q)$ is not directly found in traffic flow theory; it is a side-effect of the relationship between density and flow and the Fundamental Identity of Traffic Flow.

This framework is especially helpful for illustrating the role of tolls. The total travel time on a route is $q \cdot w_s(q)$ (veh/min), so the marginal social "cost" of adding a traveler to some

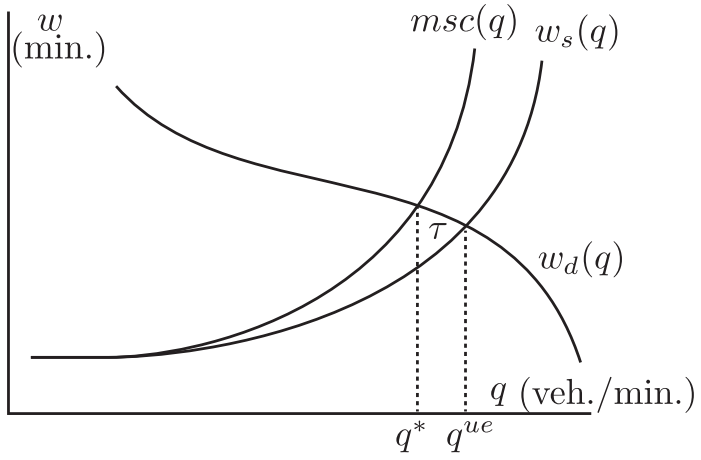

Figure 3.3: Walters (1961) supply/demand framework, finding optimal flow q^* where the Pigovian tax bridges private cost, w_s, and marginal social cost, msc.

route is

$$msc(q) := q \cdot w_s'(q) + w_s(q) \tag{3.2}$$

The individual traveler, though, endures only the $w_s(q)$ component—the average cost. Therefore, vehicles will enter the route whose drivers value the trip less than the social cost of the journey. The ineffieciency is eliminated if every traveler has to pay a toll equal to the marginal cost of the journey. The optimal size of this tax is equal to the gap, τ, between $w_s(q)$ and $w_d(q)$ at the point where $w_d(q)$ intersects $msc(q)$. With such a tax, not all congestion is eliminated, but every traveler who is willing to pay the toll deems the congestion externality of her trip to be worthwhile.

3.3 Traffic flow in a network

The traffic flow theory discussed above pertains to links. There has also been vigorous research in macroscopic traffic relationships among average descriptors at the level of entire networks, where the world "network" is used loosely; it could mean a long freeway, ring road, downtown grid or some other type of facility. What most distinguishes network-level relations from link ones is probably the absence of an explicit spatial representation. Such theories do not trace characteristics or the locations of particular vehicles. Instead, they reproduce for a facility of positive dimension the same type of relationship among average flow, density and speed that conventional theory gives for a dimensionless point on a link. This section reviews the development and key findings of this agenda.

3.3.1 Early efforts

The idea of network-level traffic relations has invited research since at least the late 1960's. Most early work (Thomson, 1967; Smeed, 1966; Wardrop, 1968; Zahavi, 1972) gave monotonically-rising speed-flow relationships for downtowns. Zahavi (1972), for instance, states that $v \cdot q = \alpha^2$ for some network parameter α. This implies

$$v(k) = \alpha^2/\sqrt{k}. \tag{3.3}$$

As is realistic, this relationship declines. But, unrealistically, it declines so slowly that flow rises monotonically with density by $q = \alpha \cdot \sqrt{k}$. This monotonocity, observed for the network as a whole, invites the question: "If the individual links of the network can become hypercongested, why can't the network?"

Godfrey (1969), on the other hand, originated a macroscopic theory capable of handling flow that falls with density. Using ground-level travel time measurements and density measurements from aerial photogrpahy, the study finds a $\cap$-shaped $q(k)$ relationship between the average flow and density for the town of Ipswitch. This function has since been named the *macroscopic fundamental diagram* (MFD)—the network-level counterpart of the Fundamental Diagram on a link (Geroliminis and Daganzo, 2008).

3.3.2 Two-fluid model

In the 1980's, Robert Herman and colleagues investigated network-level traffic relations through a line of inquiry revolving on a particular model of traffic physics—the "two-fluid theory." By this theory, laid out in Herman and Prigogine (1979), there is supposed to be steady relationship between the instantaneous fraction of stopped cars in a network and (i) the average speed, and (ii) the average density. These relationships are power laws resembling the statistical physics of condensing molecules. (Herman and Prigogine were physicists.) The two-fluid theory drove a series of studies using both empirical measurements and, especially, simulation to test its validity (Mahmassani et al., 1984; Herman and Ardekani, 1984; Ardekani and Herman, 1987; Williams et al., 1987). In the end, the studies find the two-fluid model hard to support, but they establish the promise of aggregate relations surprisingly similar those on a link and give clear definitions of network-level speed, passing rate and density. Williams et al. (1987, p. 87) notes "the most important conclusion...is that it is possible to characterize traffic flow in urban street networks using relatively simple macroscopic models relating the principal networkwide traffic variables."

Though prescient, neither Godfrey (1969) nor the Herman series seems to have influenced economic models of traffic. (Small and Chu (2003) is the exception.) Various reasons could be given: empirical foundations were tenuous (the aerial photography and microsimulations involved relatively few data points); the theoretical component emphasized a particular model of traffic physics; the results did not circulate widely among economists. But the critical reason is probably that these studies do not treat vehicles' arrivals and exits, only their motion.

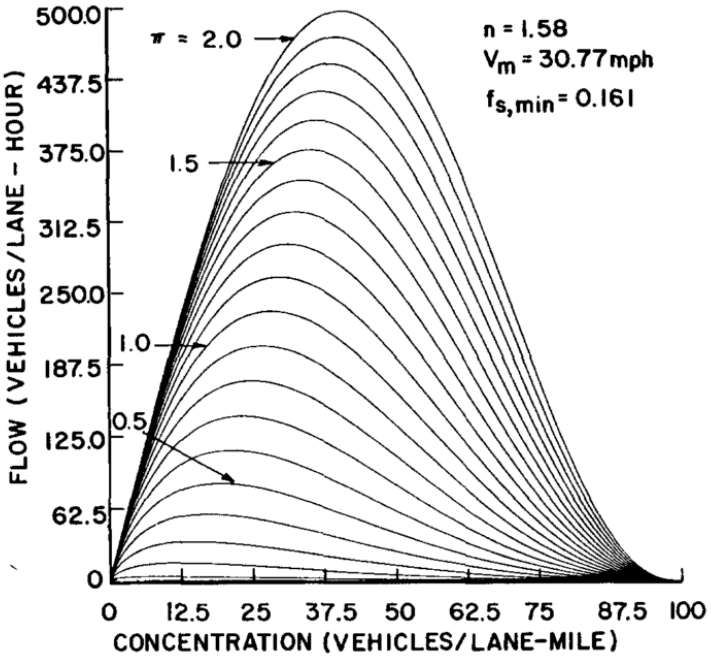

Figure 3.4: contours of early Macroscopic Fundamental Diagram for different physical parameter values from Ardekani and Herman (1987)

3.3.3 The Network Exit Function (NEF)

Since 2007, a new vein of network-level traffic flow theory has emerged, distinguished from previous work in four ways: (i) an orientation toward adaptive control; (ii) strong support from measured data and microsimulation; (iii) microfoundations in the "variational theory" of traffic flow; and (iv) explicit consideration of in- and out-flow. The last point is the one that has influenced traffic modelling.

The first such study, Daganzo (2007), argues that out-flow, e, from certain networks can be written as a function of the network accumulation, n—i.e., the number of vehicles circulating in the network at once. This function will be written $e(n)$ and has been called the Network Exit Function (NEF). The first half of the argument is that, if vehicles spread themselves evenly over a network, there will be a stable $q(k)$ giving the passing flow averaged over all points as a function of density averaged likewise. Since vehicles are distributed evenly, if M is the network's total lane-length, its average density is $k = n/M$. Vehicles are supposed to spread themselves evenly as they route around congestion.

The argument's second half is that, if average trip length, L, is stable over time and traffic conditions change slowly (not substantially within the life of a trip), then out-flow is a fixed multiple of passing flow. To see why, first note that if density is roughly constant over time and space, then a vehicle moves at about the same speed, v, during the entirety of its trip; the vehicle does not come upon an oasis of free traffic no a ticket of congestion. It follows

that the average trip duration is L/v. And so if there are n vehicles in the network, then, during an interval lasting L/v min., n vehicles will exit. The number of exits per minute is thus $e = n/(L/v) = kMv/L = Mq/L$. Since $q = q(k)$, we have

$$e(n) = \frac{M}{L} \cdot q(n/M). \tag{3.4}$$

It is easy to imagine a situation where the NEF might not hold, even when a network is homogeneous. Consider a typical downtown at 5:05 PM. Since 5 PM, cars have poured out of parking garages, swelling n and q. But the exit rate has hardly budged since 5PM, because few drivers have trips so short as to finish in five minutes. The lesson is that stationarity is critical to the NEF.

The NEF and MFD are not merely hypothesized. Geroliminis and Daganzo (2008) used speed and density data from loop detectors and taxis on city streets of downtown Yokohama, Japan to show the existence of both relations. Regarding the MFD, while data from individual detectors were widely scattered, when the data were averaged by time-of-day clearly-defined $q(k)$ appeared, (see Fig. 3.5). With caveats, MFD's have been shown to exist using real-world data: Toulouse (Buisson and Ladier, 2008), Minneapolis/St. Paul (Geroliminis and Sun, 2011). They are also confirmed in large simulations: Nairobi (Gonzales et al., 2011), San Francisco (Geroliminis and Daganzo, 2008). Regarding the NEF, Geroliminis and Daganzo (2008) also shows, using taxi data, that the average trip length in Yokohama remained constant over the day and that the exit rate remained proportional to average flow.

In addition, it turns out that the MFD for certain types of facilities can even be derived from knowledge of the traffic signal parameters (Daganzo and Geroliminis, 2008; Daganzo and Lehe, 2016). Generally, the MFD has a the shape of a smoothed-out trapezoid whose height is the highest possible average flow through the traffic signals.

3.4 Gonzales (2015)

The last study we will examine is Gonzales (2015), which presents a model of rush-hour traffic flowing at a steady rate into a downtown governed by an NEF over a fixed period of time. Travelers in the model choose between taking car and transit, and the study solves for the socially optimal difference in price between the two modes. The study is not exactly a dynamic model in the tradition of Vickrey (1969), because travelers do not choose among times-of-day; but it is not quite a static model either, because the state of traffic varies at the beginning and end of the rush.

3.4.1 User equilibrium

All the car trips into the downtown have the same length. Therefore, it is possible for the downtown to exhibit an NEF; and, in fact, the downtown is assumed to do so at all times,

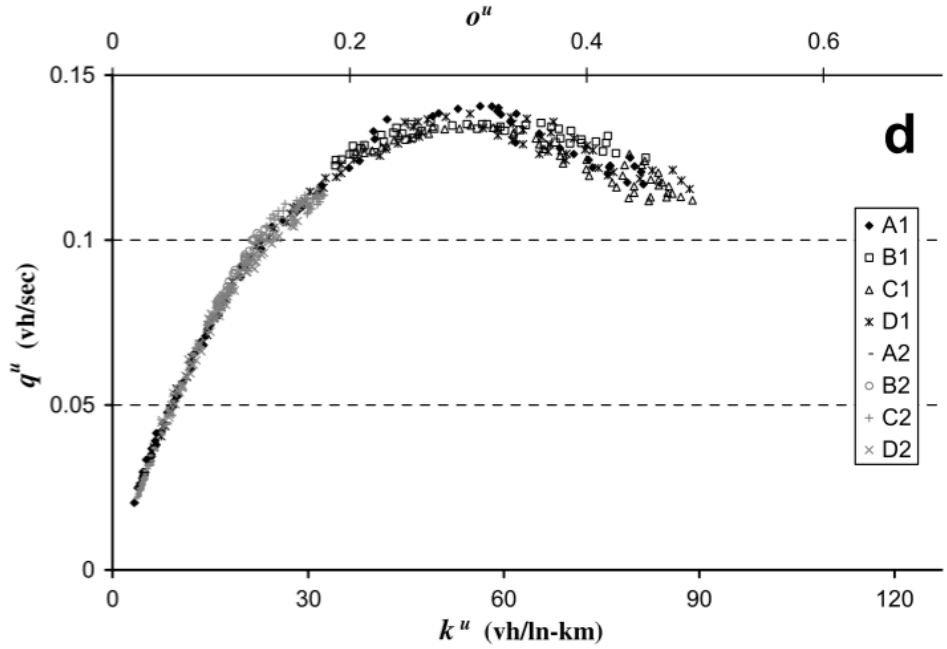

Figure 3.5: Macroscopic Fundamental Diagram from data in Yokohama, Japan. The low scatter results from tendency of vehicles to spread themselves around the network. (Geroliminis and Daganzo, 2008)

so that the rate of exits is a pure function $e(n)$ (veh/min) of the accumulation, n (veh). Moreover, it is assumed that the duration of a trip depends only on the value of the exit rate at the time the trip ends, so that $w(n) = n/e(n)$ (min) gives travel time. Note this is not very realistic if conditions change over the course of the trip, which they actually do even in the context of the model's equilibrium. In the next chapter, we will look at a simulation model that takes this into account. But for now, the assumption that trip duration depends only on the exit rate, which Arnott (2013) calls the "bathtub" assumption, is taken for granted.

There is a fixed cost, w_{c0}, to car travel, and so the "cost" of a trip by car that exits when accumulation is n is

$$z_c(n) = w_{c0} + n/e(n). \tag{3.5}$$

Let $e(0) = 0$ and assume $e(n)$ is concave. It follows that $z_c(n)$ is monotonically increasing to infinity as $n \to \infty$.

The alternative to driving is to take transit, which has its own cost function. In the study's initial model, which we build on in Chapter 3, transit is unaffected by car traffic. The cost of a trip by transit is $z_t(\lambda_t)$, where λ_t is the instantaneous flow of arrivals into the network by transit. Transit exhibits economies of scale due to the fact that, when more people take transit, the agency can break even running buses more frequently. Therefore,

$z_t(\lambda_t)$ is declining. Note that, whereas car cost depends on a stock variable n, transit cost depends on a flow variable λ_t.

During the rush, the flow into the downtown by both modes is constantly λ between times-of-day 0 and T. Travelers who arrive at each moment decide between the two modes so as to minimize their costs. It happens to be the case that $z_c(0) < z_t(\lambda)$. Otherwise, at the beginning of the rush when the network is empty, it would be an equilibrium for everyone to just take transit for the whole rush, which is not very interesting.

Now, let n_{crit} be the value of accumulation such that

$$z_c(n_{crit}) = z_t(\lambda - e(n_{crit})). \tag{3.6}$$

Although the assumption is unstated, the study assumes that travelers decide between the two modes in a short-sighted way. Even though a traveler arriving in the network at time t will be in the network for a while, she bases her mode choice on the value of n at time t. Therefore, at any moment t for which $n(t) < n_{crit}$, then everyone arriving at t will drive; and at any moment where accumulation is changing everyone will be using one mode or the other (travel time on both modes cannot be the same if $n(t)$ is changing). Accumulation is changing whenever the rate of car arrivals is different from the exit rate, so it must be that people only use both modes if cars arrive at exactly the rate they exit. This fact allows us to sketch from formal relations a picture of what the rush looks like. Let $A_c(t)$ represent the cumulative number of car arrivals in the network by time-of-day t, so that $\dot{A}_c(t)$ is the rate of car arrivals at t. It follows from our deductions that

$$\dot{A}_c(t) = \begin{cases} \lambda & n(t) < n_{crit} \\ e(n) & \text{otherwise.} \end{cases} \tag{3.7}$$

As stated, car is cheaper than transit when the rush begins. Thus, the rush must unfold in the following manner: at first everyone takes car until the accumulation reaches n_{crit}, after which the rate of arrivals by car is $e(n_{crit})$ and that by transit is $\lambda - e(n_{crit})$, until the rush is over and accumulation fizzles out. Let $E_c(t)$ be the cumulative number of exits by car at time-of-day t (the integral of $e[n(t)]$ from t to t). The rush must appear as in Fig. 3.6.

3.4.2 Tolls

Although tolls cannot push travelers between routes in this model or cause them to stop traveling, they can get travelers out of their cars and onto transit. Doing so has two benefits: auto congestion is lessened and greater economies-of-scale in transit can be realized. Tolls are given in terms of travel time.

Tolls are pure transfers from drivers to the government, and so the optimal toll level is one that minimizes aggregate travel cost. Aggregate travel cost over the whole rush can only be calculated by integrating moment-to-moment until all cars have departed, but if the steady state lasts a long time then we can approximately minimize aggregate travel cost by

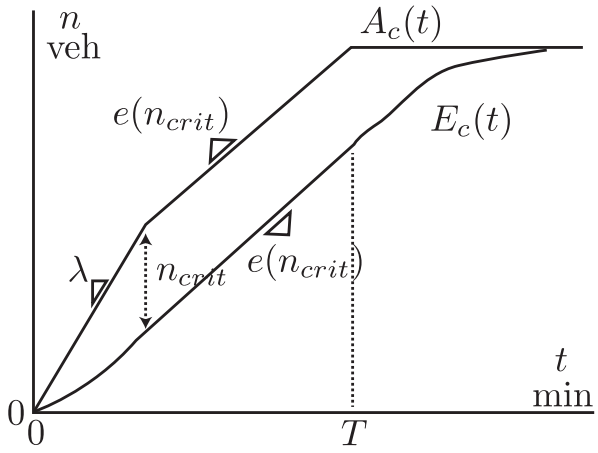

Figure 3.6: In equilibrium of Gonzales (2015), all travelers drive until accumulation (vertical gap between A_c and E_c) reaches n_{crit}. Then they reach a steady state where they arrive and exit at rate $e(n_{crit})$ until $t = T$.

minimizing the cost per minute during the steady state. The aggregate cost (aggregate time spent traveling) during each minute of the steady state with accumulation n is

$$Z(n) = e(n) \cdot z_c(n) + [\lambda - e(n)] \cdot z_t(\lambda - e(n)), \tag{3.8}$$

where the first term on the RHS is the total cost of car travel and the second term the total cost of transit travel. Therefore, the cost-minimizing accumulation is

$$n^* = \arg\min_n \{Z(n); n \in [0, n_j], e(n) < \lambda\}. \tag{3.9}$$

Once this value is determined numerically, it is straightforward to find the toll that achives this accumulation as a decentralized equilibrium. The toll τ^* that does so is given by the level

$$\tau^* + z_c(n^*) = z_t[\lambda - e(n^*)] \tag{3.10}$$

that equalizes the costs of car and transit when accumulation is n^*.

Of course, there is much unrealistic about the model and this method of solution: the travelers are shortsighted, travel times depend only on accumulation at the time of exit and we choose the optimal accumulation based only on conditions during the steady state, not over the whole rush. Nonetheless, the analysis contains an important lesson: even though the model is somewhat dynamic, the fact it has a steady-state enables us to look at the situation in a more-or-less static framework, minimizing the cost per unit of time to obtain a single, invariant level of the toll.

Chapter 4

Self-selection in downtown congestion pricing

All of the tolls analyzed in Chapters 2 and 3 charged for the act of making a trip. Thus, a cab or a vehicle employed for ridesharing may circulate many kilometers over the course of hours but pay the same toll as a commuter who parks immediately upon entry. Therefore, these systems will be classified as "trip tolls" because a "trip" (the mere act of traveling inside the zone somehow) is the unit-of-account in the transaction between the driver and the toll authority.

A shortcoming of a trip toll—relative to, say, a toll that varies by distance-traveled—is that it fails to distinguish among drivers for the delay they impose on others. There are two ways of thinking about the consequences. First, at the extensive margin of choice (how much people drive, given that they do drive), there is the problem of "moral hazard": unless they are charged for their externalities, travelers may drive farther than they would otherwise. Second, at the intensive margin of choice (whether people drive), there is a problem we will call "self-selection": even if trip lengths are fixed, a trip toll may discourage short trips with little cost to society and allow long trips that impose externalities larger than their private value. For example, if a traveler's destination is one block inside the tolled zone, then a trip toll may encourage cause her to park down the block and then walk to the destination, even though the additional block of car travel would add little to congestion.

The focus of this chapter is on self-selection, and in particular the case where a driver's willingness-to-pay for a car trip rises with its length. To explore self-selection issues and to compare a trip and distance toll, we present an original, static traffic model. In this model, travelers enter a downtown zone, and each traveler independently chooses between doing so by car or by an alternative mode that is called "transit." The choice is probabalistic, and travelers have significantly different trip lengths. Speed is the same everywhere in the zone, and it depends at each moment only on the zone's density—which is to say that the zone exhibits a well-defined Macroscopic Fundamental Diagram. This setting, of course, is similar to that of Gonzales (2015), but simpler in some ways and more complex in others. The probabalistic nature of the choice and the variance of trip lengths are enrichments of

that model, and it uses the more basic MFD relation rather than the NEF relation, which cannot be said to hold when trip lengths vary. On the other hand, for parsimony, we simplify the model by omitting the true fact of scale economies in transit.

The chapter is organized as follows. Section 4.1 sets up the principal model, for which Sec. 4.2 derives the user equilibrium (UE). Sec. 4.3 shows that equilibrium to be the steady state that prevails in the middle of the rush for a dynamic model. Afterward, Sec. 4.4 considers the model's social optimum (SO), and Sec. 4.5 compares the impact of a distance and a trip toll. Next, to show the same effects can be derived in a model with only car traffic, Sec. 4.6 briefly exposits a model of cars choosing between traversing the charging zone or avoiding it on a free, capacious ring-road. By way of a conclusion, Sec. 4.7 summarizes the chapter, discusses policy insights and lists several directions for future research.

4.1 Model setup

This section sets up the main model of the chapter. The model has two components: a "physics" side describing the congestion technology, and a "demand" side describing mode choice.

4.1.1 Physics

The physical setting of the model is a downtown zone governed by an MFD. As such, a declining function $v(k)$ (km/min) gives the traffic speed v (km/min) given a density of k vehicles per km. But for our purposes, it will be more convenient to work, not with speed v, but with its inverse $p := 1/v$ (min/km). Traffic engineers call the inverse of speed *pace*; it is the time required to traverse a fixed distance—which is, in our case, arbitrarily one km. Consequently, the instantaneous "circulation," q (veh/min), is

$$q(k) = \frac{k}{p(k)}.$$

The function $p(k)$ is weakly rising to infinity, in such a way that $q(k)$ takes a unimodal "hump" shape, tending to 0 as $k \to \infty$. The Macroscopic Fundamental Diagram, equivalently expressed as $p(k)$ and $q(k)$, appears in Fig. 4.1. The density $k = k_0$ is where the zone achieves its the maximum circulation, q_0. The pace $p = p_f$ is the "free flow" pace $p_f = p(0)$ at which cars drive unobstructed. Blue highlighting denotes uncongested traffic states—those where pace and circulation rise together. For $k > k_0$, traffic is hypercongested, and pace falls as circulation rises.

4.1.2 Demand

Each traveler who enters the network—whether by car or by transit—is defined by a tuple (ε_i, l_i), where l_i is her trip length and ε_i her draw of a random cost component. The random cost ε_i is distributed with zero mean and has a cumulative distribution F.

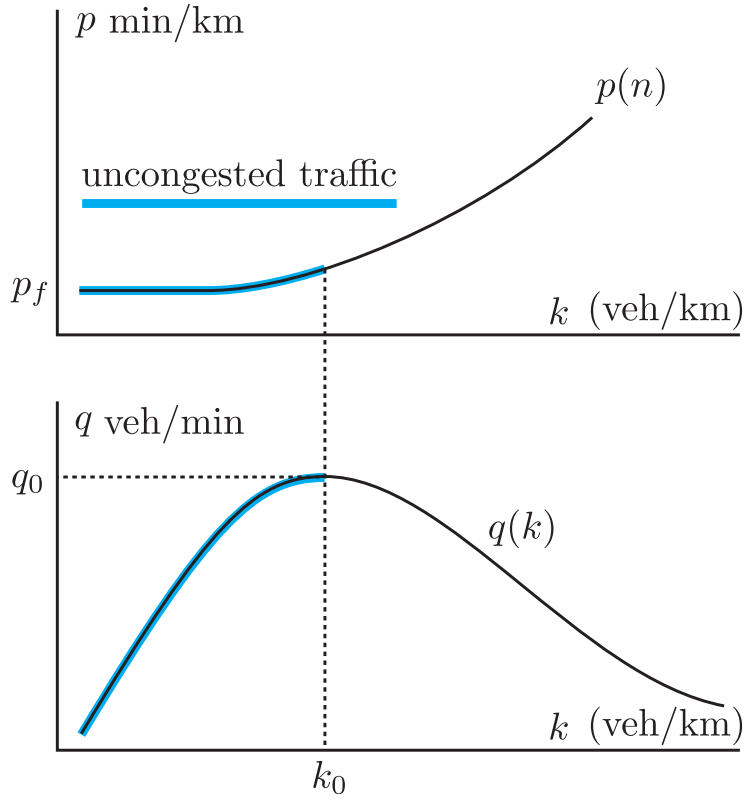

Figure 4.1: Macroscopic Fundamental Diagram expressed equivalently in terms of pace p or circulation q. Highlighted segments are uncongested regime, where circulation rises with density and falls with pace.

The network is homogeneous, and so, to avoid creating an extraneous variable for the size of the zone, we treat a particular area with a lane-length of one kilometer and define other variables per lane-km. This makes it so that the density and the number of vehicles in the zone are the same. Every minute, $g(l)$ (veh/km/lane-km) trips of length l arrive in the zone. These may enter by crossing the boundaries of the zone in the in-bound direction using either car or transit, by leaving a parking space located inside the zone or by boarding transit inside the zone.

Arriving travelers decide between driving a car at pace p and taking an alternative mode—called "transit"—that travels at pace β regardless of what car traffic does. "Transit" thus could mean subway, buses in segregated lanes, bicycling or simply walking from a garage near the edge of the zone. The difference in fixed cost, excluding tolls, between the two modes is γ (measured in units of travel time). Therefore, for a traveler with trip length l_i who drives when the pace of traffic is p, the deterministic part of the *net* advantage of driving is

$$V(p, l_i) := (\beta - p)l_i - \gamma. \tag{4.1}$$

By assumption, $\gamma > 0$ and $\beta > p$ near the model's equilibrium; that is, driving has a higher fixed cost than transit and is faster. Thus, the first term of V represents the travel time savings from driving.

The traveler's decision and utility depend not only on V but on her ε_i. This attribute accounts for unobserved differences in the preferences for car and transit. There are many conceivable origins for ε; it may take into account, for instance, the distance between the nearest parking garage or transit stop and the traveler's home or destination. Without loss of generality, the utility of every traveler i is indexed to zero under the transit alternative. A traveler chooses to drive if and only if $V(p, l) - \varepsilon \geq 0$, so i's utility when pace is p is $\max\{V(p, l_i) - \varepsilon_i, 0\}$.

4.2 User equilibrium (UE)

4.2.1 Derivation

Having specified the model, we now turn to an examination of the UE that prevails without tolls. For traveler i to drive, it must be the case that

$$\varepsilon_i \leq V(p, l_i), \tag{4.2}$$

and so the probability i drives, given p, is

$$P\{i \text{ drives}\} = F[V(p, l_i)]. \tag{4.3}$$

When every i satisfying (4.2) drives, by Little's Law, the circulation, q, must be

$$q_d(p) := \int_0^\infty g(l) \cdot l \cdot F[V(p, l)]dl. \tag{4.4}$$

The function $q_d(p)$ will be called the "demand curve." It weakly declines with p, because fewer travelers drive as traffic slows.

Note that the demand circulation is to be distinguished from the demanded *arrival* flow given by

$$a_d(p) := \int_0^\infty g(l) \cdot F[V(p, l)]dl. \tag{4.5}$$

The circulation flow, q (veh/min), measures what flow an observer would measure if she stood on a street of the zone and counted how many cars pass every minute. The arrival flow, a (veh/min), measures how many cars enter the network every minute.

The UE is found by pairing the demand curve $q_d(p)$ with the physics of the zone. It is a situation in which the individual choices expressed in (4.4), conditiononal on p, are consistent with the traffic physics that give rise to that very p. Consistency between demands and physics holds when the density k is at a level k^e such that

$$q_d[p(k^e)] = q(k^e). \tag{4.6}$$

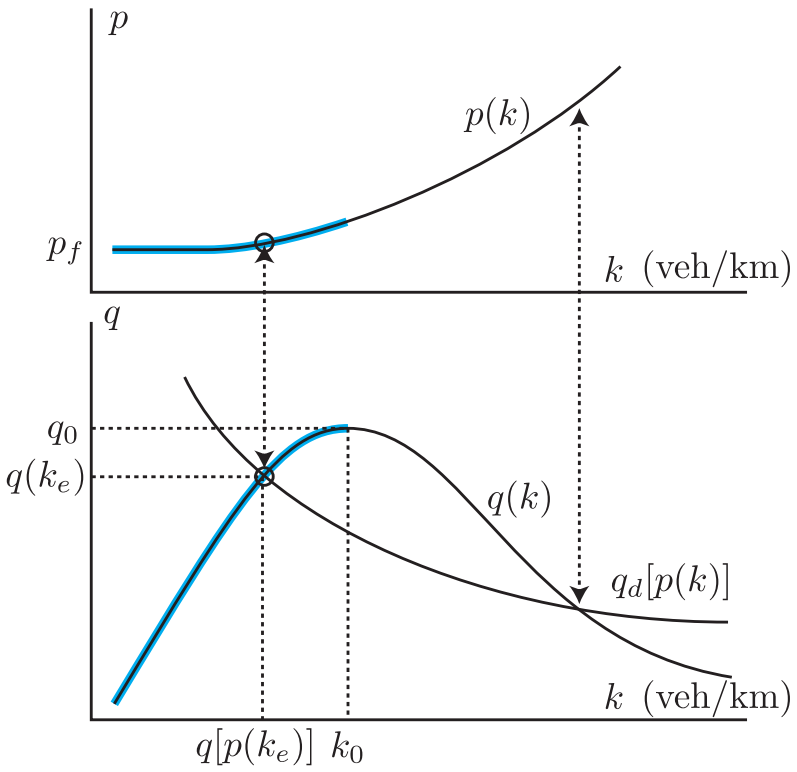

Figure 4.2: User equilibria along the MFD. Leftward equilibria is uncongested, rightward is hypercongested. Equilibria are points of intersection between the demand curve q_d and the MFD.

The LHS of this expression is the circulation demanded when density is k^e. The RHS is the circulation that the network's physics imply when $k = k^e$. The two sides are plotted together in Fig 4.2 below. Their intersections are the model's UE: at any other k, pace would have to be either rising or falling.

4.2.2 Simplifications

Now we will make two simplifications that make the situation more amenable to economic analysis: a change-of-variables and a plausible assumption.

- *Flow congestion.* First, we excise k and put things in terms of p and q, as in the canonical congestion pricing model of Walters (1961). Such "flow congestion" invites convenient comparisons to the supply-and-demand curves of microeconomics and to concepts such as marginal cost. To make the switch, first let $p_s(q)$ map $q(k) \to p(k)$ on the interval $[0, k_0]$; and let $\tilde{p}_s(q)$ do so on the interval $(k_0, \infty]$. The first function,

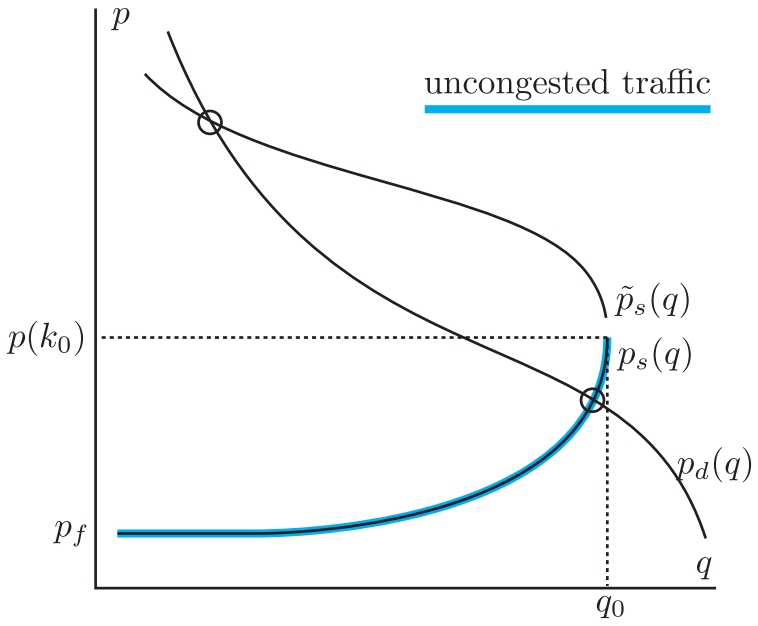

Figure 4.3: User equilibria expressed in terms of pace and circulation only—without density—as in Walters (1961).

$p_s(q)$, links pace and circulation for uncongested traffic; the latter, $\tilde{p}_s(q)$ describes hypercongested traffic. Next, let $p_d(q) := q_d^{-1}(q)$ give the pace that induces a demanded circulation of q; this is a simple inverse operation, since $q_d(p)$ monotonically declines. Figure 4.3 shows $p_s(q)$, $\tilde{p}_s(q)$ and $p_d(q)$, which meet at the equilibria of Fig. 4.2.

- *Ignore hypercongested equilibria.* There is no monotonic physical relation between the circulation and pace. Hence multiple equilibria—expressed graphically as multiple intersections between the "supply" and "demand" curves—can arise. Since $p_s(q)$ rises weakly, and $\tilde{p}_s(q)$ and $p_d(q)$ both decline monotonically, there can no more than one uncongested equilibrium but many hypercongested equilibria. Issues of hypercongested equilibria—principally, their stability, relevance and interpretation—have long been debated for other static models (see Verhoef (1999), Verhoef (2001) and Small and Chu (2003)). This analysis sidesteps that debate. By assumption, $p_d(q)$ intersects $p_s(q)$, and this is the only equilibria of interest. The justification for ignoring hypercongested equilibria is that we are interested in tolls, and authorities can always choose one sufficient to preclude hypercongestion.

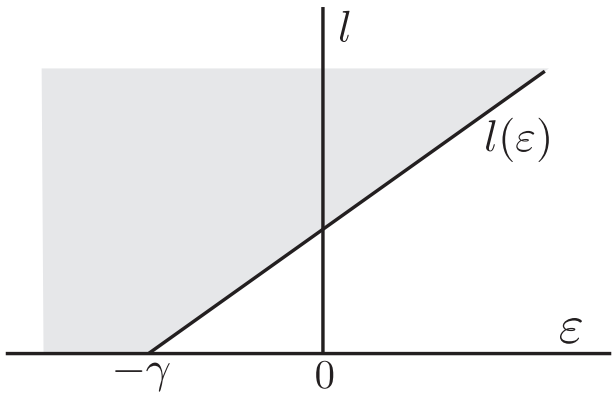

Figure 4.4: User equilibrium mode split. Only drivers with trip-length/random cost tuples in the shaded region drive.

4.2.3 Mode split in the user equilibrium

We will now explore the mapping of travelers to modes—hereafter, called the "mode split"—in the UE. To start, note that the drive condition

$$V(p, l_i) \leq \varepsilon_i,$$

implies

$$l_i \geq \frac{\gamma + \varepsilon_i}{\beta - p}. \tag{4.7}$$

Thus, along a given slice of ε, only travelers with trips longer than $(\gamma + \varepsilon_i)/(\beta - p)$ drive. The bar to drive also rises with ε_i: travelers with a higher random cost must go farther for driving to be worthwhile. The resulting mode split is demonstrated in Fig. 4.4, where travelers whose (ε_i, l_i) tuple lie above in the grey space above

$$l^{ue} = \frac{\gamma + \varepsilon}{\beta - p}$$

drive, while those below ride transit. The root of l^{ue} is $-\gamma$; the intercept is the level of ε_i below which everyone drives. The intercept of l^{ue} is $\gamma/(\beta - p)$; half of travelers with a trip length equal to the intercept drive, while half take transit.

4.3 Dynamic model

Obviously, a static model does not accurately describe real traffic, which ebbs and flows. However, it turns out that our static model well describes the steady state of a dynamic model in which travelers enter the network at a constant rate over a finite rush, as in Gonzales (2015).

Suppose that the network begins empty at time $t = 0$. Thereafter, travelers of different lengths arrive every minute as given by $g(l)$ until time $t = T$, so that T is the rush duration. The travelers who arrive in each moment independently choose between car and transit, depending on the modes' relative costs. Unlike in the static model, here pace can vary during the life of a trip: if $k(t)$ is the time-path of density, then $p[k(t)]$ is the time-path of pace. Let $w_i(t_a)$ give i's trip duration for a trip arriving in the network at time t_a. This duration is determined implicitly by

$$\int_{t_a}^{t_a+w_i(t_a)} 1/p_s[k(t)]dt = l_i, \tag{4.8}$$

where the upper bound of the integral, $t_a + w_i(t)$, is the time i exits. Note that these physics are more complete than the physics of Gonzales (2015), where travel time depended only on the accumulation/density of the network at the time of exit and where travelers were shortsighted. It follows that the traveler i will drive if and only if

$$\beta \cdot l_i - w_i(t_a) - \gamma \geq \varepsilon_i \tag{4.9}$$

For this model, it is hard to prove that there is necessarily a UE or a what it is, but we can hypothesize one and then test it with an agent-based simulation. Our hypothesis is the following: At first, pace will be low because the zone is empty, and therefore driving will be attractive to many travelers to whom it is not in the static UE. As the zone fills up, pace will rise and so fewer travelers will find driving attractive. Eventually, pace and circulation will rise (speed will fall) until stabilizing at the levels predicted by the static UE. This situation will persist as a sort of steady state until near the end of the rush, when the zone will gradually clear out.

The details of how the simulation works are in the Appendix. The first step to building a simulation is to choose functional forms and parameters. The pace-density relation will be $p(k) = p^f \cdot \exp(k/k_0)$, where p^f (km/min) is the free-flow pace. The trip length distribution, $g(l)$, is uniform between an $l_{\min}$ and an $l_{\max}$ and has a scale of λ. The uniform distribution was chosen because it is easy to simulate and integrate, although a lognormal might be more realistic. The random cost ε is distributed logistic with scale parameter α and mean 0, giving the logit model popular in discrete-choice analysis. Consequently, $F(v) = [1+\exp(-v/\alpha)]^{-1}$. Table 4.1 contains the parameters chosen. The free-flow pace is $p_f = 2.0$ (30 km/hr) and β, the pace of the transit alternative, is 5.0 (12 km/hr). The value k_0 is 70 veh/lane-km, implying circulation is maximized at 14 m per vehicle Other values were chosen somewhat haphazardly in order to guarantee stability and an illustrative plot in the static numerical simulations conducted later.

The results of the simulation, for the parameters used earlier and $T = 100$, are plotted in Figures 4.5 and 4.6. Fig. 4.5 shows the cumulative number of vehicles that have arrived and exited the network by car at each time, where the density/accumulation equals the difference between the number of arrivals and exits. At first the slope of $A(t)$, which is the gross arrival rate, is large because many travelers wish to drive when the network is empty, and the slope

Table 4.1: Parameter values and specifications

parameter	value/specification
$g(l)$	$U(l_{\min}, l_{max})$
$p(k)$	$p_f \cdot \exp(k/k_0)$
$F(v)$	$[1 + \exp(-v/\alpha)]^{-1}$
p_f	2.0
k_0	70
λ	4
β	5.0
γ	6.0
α	3.0
$l_{\min}$	2.0
$l_{\max}$	8.0
T	100

of $E(t)$ is low because no drivers have traveled for long enough to complete their trips. But the slope of $A(t)$ falls as pace rises (as traffic slows down), and that of $E(t)$ rises as travelers complete their trips. By about time $t = 30$, the slopes of $A(t)$ and $E(t)$ level out and become equal, so the density/accumulation in the zone remains constant. This is the steady state, and it persists until $t = 100$.

As predicted, the steady state is the same as the one in the static model. This fact is illustrated in Fig. 4.6, where $q(t)$ shows the circulation in the experiment and q^{ue} the circulation of the static UE with the same parameters (except for T). While $q(t)$ is not completely stable, due to the probabalistic nature of the choice and the discreteness of the simulation, it is *reasonably* stable and remains in the vicinity of q^{ue}.

4.4 Social optimum (SO)

This section derives the SO and compares it to the UE. To begin, note that the expected utility of a traveler with trip length l_i is the expectation

$$S[V(p, l_i)] := E\left\{\max\left[V(p, l_i) - \varepsilon, 0\right]\right\}. \tag{4.10}$$

Therefore, aggregate utility, CS, when pace is p, must be

$$CS = \int_0^\infty g(l) \cdot S[V(p, l)]dl. \tag{4.11}$$

Now, let us consider the impact of a particular traveler i's decision to drive on aggregate welfare. By driving, i earns utility $V(p, l_i) - \varepsilon_i$ and slows down traffic by a small amount.

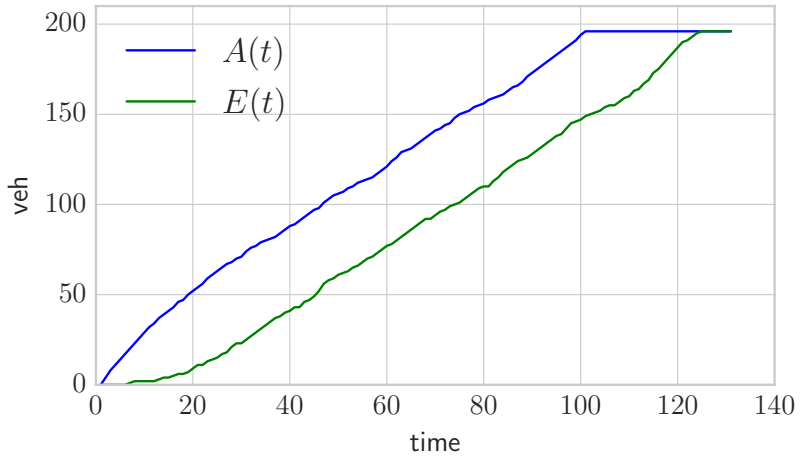

Figure 4.5: Cumulative arrival ($A(t)$) and exit ($E(t)$) curves from the agent-based simulation. As in Gonzales (2015), the slopes are constant during the steady state in the middle of the rush.

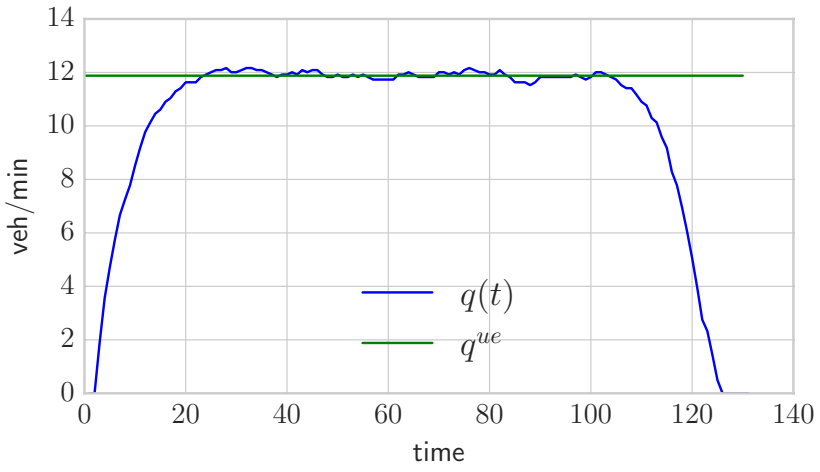

Figure 4.6: Circulation over time from the agent-based simulation. During the steady state, circulation converges to the value q^{ue} predicted by the static model.

Since i's car trip raises circulation by l_i, the trip's externality is

$$l_i \cdot p_s'(q) \cdot \frac{\partial CS}{\partial p}. \tag{4.12}$$

Two convenient facts will give $\partial CS/\partial p$. First, $\partial V(p,l)/\partial p = l$. Second, per the result of Williams (1977), $S'(V) = F[V]$. These substitutions yield

$$\frac{\partial CS}{\partial p} = \int_0^\infty g(l) \cdot l \cdot F[V(p,l)]dl = q_d(p). \tag{4.13}$$

That is, the derivative of consumer surpus with respect to pace is simply the circulation. It follows that the externality of i's car trip is

$$l_i \cdot c(q),$$

where

$$c(q) := p_s'(q) \cdot q \tag{4.14}$$

is the externality per km-driven. Therefore, i's car trip only raises aggregate welfare if

$$V(p, l_i) - \varepsilon_i > l_i \cdot c(q), \tag{4.15}$$

which also implies

$$V[p_s(q) + c(q), l_i] < \varepsilon_i. \tag{4.16}$$

At the SO circulation, q^*, only travelers satisfying this condition drive, and so

$$q^* = q_d[p_s(q^*) + c(q^*)] \tag{4.17}$$

Using this expression, q^* can be found graphically—as in Fig. Walters (1961)—after inserting both sides into $p_d(q)$ to obtain

$$msc(q^*) = p_d(q^*), \tag{4.18}$$

where

$$msc(q) := p_s(q) + c(q) \tag{4.19}$$

is the marginal social cost of a kilometer traveled. Since $msc(q)$ is always at least as high as $p_s(q)$, the UE circulation must lie weakly to the right of the SO circulation, and there can be no more than one SO and one UE.

Now we will compare the mode split of traffic in the SO and UE. See Figure 4.8. The SO mode split is given by the line

$$l^{ue} = \frac{\gamma - \varepsilon}{\beta - msc(q^*)} \tag{4.20}$$

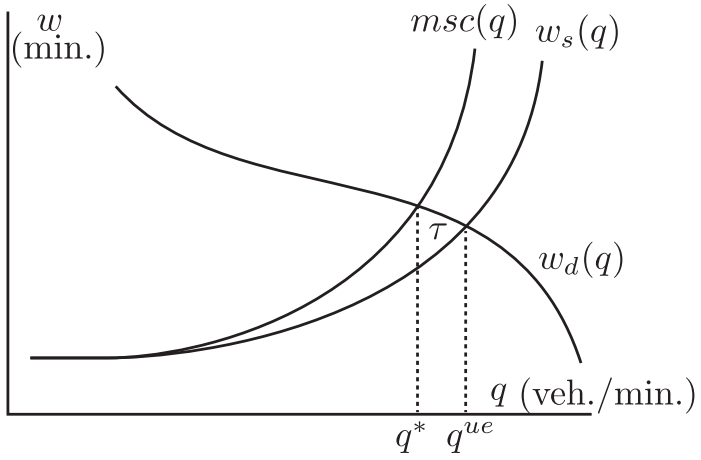

Figure 4.7: Supply, demand and marginal social cost with optimal tax τ, as in Walters (1961).

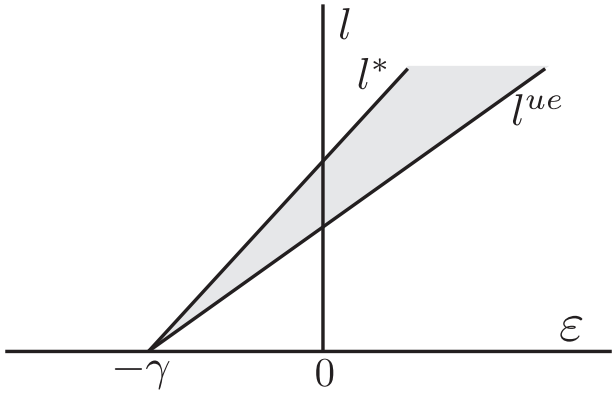

Figure 4.8: Social optimum vs. user equilibrium mode split. The user equilibrium population is a superset of the social optimum, and both mode split lines have the same root.

while the UE mode split is given by the line

$$l^* = \frac{\gamma - \varepsilon}{\beta - p_s(q^{ue})}. \tag{4.21}$$

Both lines have the same root at $\varepsilon = -\gamma$. For values greater than $-\gamma$, l^* is at least as high as l^{ue}, since it must be that $msc(q^*) > p_s(q^{ue})$. Hence, the set of drivers in the UE contains all the drivers in the SO. The grey area of Figure 4.8 contains those travelers who drive in the UE but not in the SO.

4.5 Tolls

So far we have established that the static UE can be suboptimal. We now ask what effects tolls may have on total social surplus and on consumer surplus. To this end, this section contrasts a distance toll that charges in proportion to a traveler's trip length and a trip toll that charges a traveler for the act of driving. Let τ_t give the size of a trip toll and τ_d the size of a distance toll. Like other cost measures, tolls are given in units of travel time. To analyze the tolls, enrich $V(p, l)$ with arguments for τ_t and τ_d. The deterministic gain from driving becomes

$$V(p, l, \tau_d, \tau_t) := l \cdot (\beta - p - \tau_d) - \gamma - \tau_t. \tag{4.22}$$

4.5.1 Total social surplus (TSS)

A common assumption about social surplus, which we make here, is that money is weighted the same way in driver's hands and in the toll authority's coffers. Because the distribution of revenue is irrelevant to TSS, a toll that assigns travelers to modes according to (4.16) maximizes TSS. An appropriately-chosen distance toll does so: if $\tau_d = c(q^*)$, i drives if and only if (4.16) holds. This is the classic Pigovian result: a toll equal to marginal external cost ensures every driver's gain outweighs her externality.

By contrast, only under special circumstances does some τ_t obtain the SO mode split. Figure 4.9 makes the point. The lines $l^{\tau_t^1}$,$l^{\tau_t^2}$,$l^{\tau_t^3}$ and $l^{\tau_t^4}$ give the mode split induced for different levels of τ_t, where $\tau_t^1 < \tau_t^2 < \tau_t^3 < \tau_t^4$. To see why these lines fan out as pictured, note the mode split for a given τ_t is given by

$$l_i^{\tau_t} = \frac{\gamma + \tau_t + \varepsilon_i}{\beta - p^{\tau_t}}, \tag{4.23}$$

where p^{τ_t} is the pace induced by τ_t. Thus, $l_i^{\tau_t}$ has a root at $-(\tau_t + \gamma)$, and so the lines slide leftward along the ε axis. Moreover, since the trip toll reduces p (by discouraging trips), the lines' slopes rise. Because they influence both the slope and the root of the dividing line, a trip toll permits both what we will call "Type I" and "Type II" errors. The region R_1 contains Type I errors; they are "false positives," travelers who ought not to drive but who do drive. R_2 contains Type II errors or "false negatives," travelers who ought to drive but who take transit. For a high-enough trip toll, such as $l^{\tau_t^4}$, the slope is higher than that of l^*, permitting only Type I errors. Such a level cannot possibly be optimal, because, at this level, slightly lowering τ_t would only add car trips whose private net benefit exceeds their delay externality. Therefore, the best possible trip toll must permit both type I and type II errors, athough whether these errors actually manifest depends on the distributions of ε and l.

The trip toll's efficiency problems are moot if either l_i or ε_i does not vary. When there is a universal trip-length $\bar{l}$, all the population lies along a single horizontal line $l = \bar{l}$, and if τ_t is chosen such that l^{τ_t} intersects $l = \bar{l}$ where l^* does, traffic will be divided optimally. The trip toll that does so is $\tau_t = c(q^*) \cdot \bar{l}$. Similarly, in the deterministic demand case where $\varepsilon_i = 0 \, \forall i$,

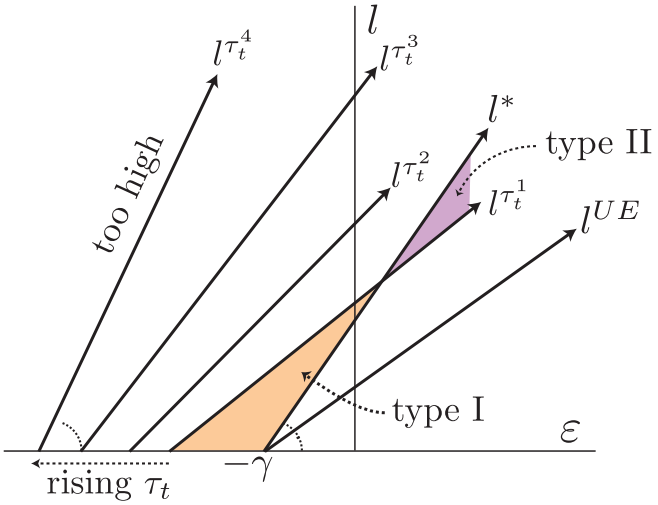

Figure 4.9: Inefficiency with trip toll. Type I errors are drivers needlessly priced off the road by the trip toll; Type II errors are drivers who drive but should not with the trip toll. The optimal trip toll must have both, or else, like $l^{\tau_t^4}$, it will exclude too many drivers.

the population is dispersed only along the vertical axis. Thus, if τ_t is chosen such that l^{τ_t} has the same intercept as l^*, then traffic will be divided efficiently. In either case—constant trip length or deterministic demand—R_1 and R_2 still exist, but they are void of travelers; their existence has no practical consequence.

4.5.2 Consumer surplus

We have just seen that, generally speaking, the distance toll outperforms the trip toll at raising total social surplus. The same cannot be said of *consumer* surplus, meaning the aggregate welfare of travelers only and not of the government.

A distance toll necessarily lowers consumer surplus, because it reduces V for all trip lengths. To see why, suppose a rise $\Delta\tau_d$ in the distance toll changes pace by $\Delta p^d < 0$. The change in V, ΔV_i^d, for traveler i is

$$\Delta V_i^d = -l_i \cdot (\Delta p^d + \Delta\tau_d). \tag{4.24}$$

It must be true that $-\Delta p^d < \Delta\tau_d$ (i.e., that the fall in pace is smaller than the rise in the distance toll), or else the toll increase would not eliminate any trips. Hence, ΔV_i^d is negative.

A rise in the trip toll, on the other hand, will lower V for short trips and raise it for long-enough trips. Let $\Delta p^t < 0$ be the change in pace resulting from a rise of $\Delta\tau_t$ in the trip toll. The change in V_i is now

$$\Delta V_i^t = -l_i \cdot (\Delta p^t) - \Delta\tau_t. \tag{4.25}$$

Hence, the longer her trip, the more the traveler stands to gain (or less the traveler stands to lose) from a rise in the trip toll. There is a certain trip length at which ΔV_i^t changes sign from negative to positive: ΔV_i^t is positive when $l_i < -\Delta\tau_t/\Delta p^t$ and negative when $l_i > -\Delta\tau_t/\Delta p^t$.

Note that ΔV_i does not necessarily give the change in utility that i will experience form a toll increase. Travelers who take transit before and after the rise are unaffected. Travelers who switch from car to transit are worse off, but not by as much as ΔV, because their ex post utility is bounded below by 0. Travelers who switch from transit to car (This will happen to some long trips after the trip toll increase.) are better off, but not by as much as ΔV, because these travelers' ex ante utility is 0. Still, the qualitative point holds up: the trip toll increase will leave certain long-distance travelers better off, while the distance toll increase will hurt certain traveler's utility but not improve any traveler's utility.

4.5.3 Numerical example

For the numerical example, we use $p(k)$, α, λ, β and γ from Sec. 4.3 and again a uniform distribution of trips between $l_{\min}$ and $l_{\max}$. As explained in Small and Rosen (1981), when ε is distributed logistic with scale α,

$$S(v) = E\left[\max(v-\varepsilon, 0)\right] = \alpha \cdot \ln[1 + e^{v/\alpha}] + C, \tag{4.26}$$

where C is a constant. We ignore C when we compare the two tolls, though, because it is the same for the same traveler in both regimes. Consequently,

$$TSS = \alpha \cdot \int_0^\infty g(l) \cdot \ln[1 + e^{V(p,l,\tau_d,\tau_t)/\alpha}] \cdot dl \tag{4.27}$$

$$+ \int_0^\infty g(l) \cdot F[V(p,l,\tau_d,\tau_t)] * (\tau_t + l\tau_d) \cdot dl, \tag{4.28}$$

where the first term on the RHS represents consumer surplus and the second term toll revenue.

Table 4.5.3 presents summary statistics for the static model under different scenarios with TSS-maximizing trip and distance tolls, including consumer surplus (CS), total revenue (TR) and average trip length $\bar{l}$. Two scenarios are presented, one with short trips and another with long trips. Meanwhile, Figures 4.10 and 4.11 show the mode split diagram introduced earlier under each trip-length distribution, with colors for the population density. As Table 4.5.3 shows, either toll significantly improves total social surplus, but only in the second scenario with long trips ($l_{\max} = 8.0$) does the distance toll improve things much beyond the trip toll. In both cases, the trip toll permits a larger $\bar{l}$, because more long trips enter the network.

The size of the advantage of the distance toll is underscored by the mode split diagrams, too. In Fig. 4.10 (long trips), the R_1 and R_2 regions are fairly small, meaning not many

Table 4.2: Results of numerical simulation for static model

$l_{\min}$	$l_{\max}$	toll type	toll	CS	TR	TSS	$\bar{l}$	q	p
2.0	5.0	none	0	10.72	0	10.72	3.69	7.34	2.64
		trip	2.31	7.6	3.73	11.32	3.77	6.08	2.48
		distance	0.67	7.36	4.00	11.36	3.71	6.00	2.47
2.0	8.0	none	0	8.65	0	8.65	5.52	12.3	4.13
		trip	7.99	5.99	12.25	18.24	6.23	9.56	3.02
		distance	1.57	5.68	13.65	19.33	5.42	8.71	2.85

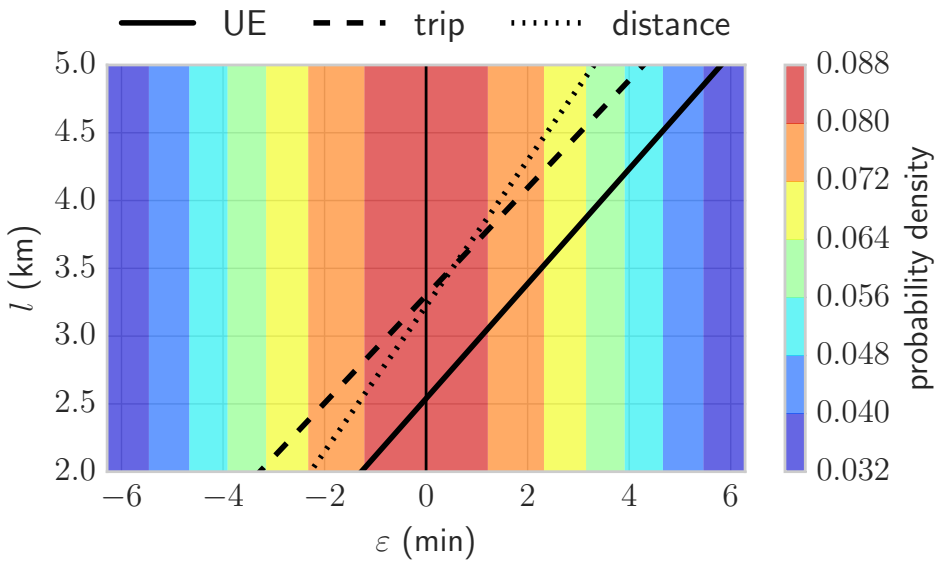

Figure 4.10: Mode split for $l_{\min} = 2.0$, $l_{\max} = 5.0$ under different toll regimes. Colors show population density.

travelers change modes between the distance and trip tolls. By contrast, in Fig. 4.11 (short trips), the lines demarking the mode split between the two tolls are quite different, and so R_1 and R_2 are fairly large.

In both cases, moreover, the trip toll improves consumer surplus more, because the distance toll raises more money. Altogether, the magnitudes of these results should likely be taken with a grain of salt, but they show that the model is computable and that the qualitative results derived above hold up.

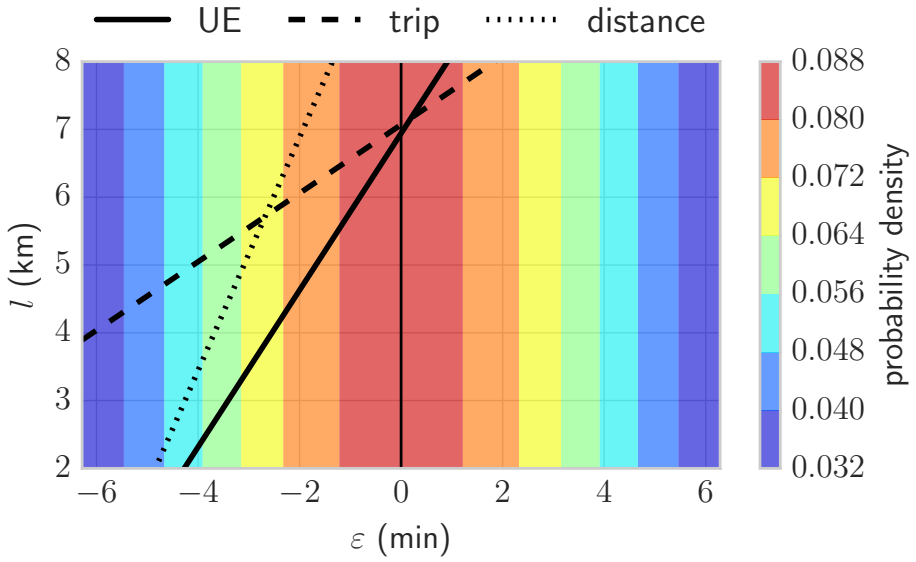

Figure 4.11: Mode split for $l_{\min} = 2.0$, $l_{\max} = 8.0$ under different toll regimes.

4.6 Ring-road model

In this section we will briefly look at another plausible model—an alternative to the mode choice model given above—that generates the same effects, albeit in a less tractable way. It is included to show that the qualitative results derived can also be obtained in a place without much transit and to spark further research.

Consider a perfectly circular zone of diameter D surrounded by a ring-road. The ring-road's capacity is large enough that its pace is not affected by the mechanizations described in the model. Suppose that this zone lies between a great many origin-destination pairs that lie outside the zone. In this case, drivers between these pairs have the choice of either cutting through the zone or taking the ring-road.

For some such trip between outside points, let l_z be the distance traveled within the zone if the driver decides to cut through, and let l_r be the distance she would travel along the ring-road if she avoids the zone—whether due to tolls or the zone having a high pace (low speed). Geometrically, because the zone is assumed to be a circle, l_z is the length of a chord, and l_r is the length of the short arc contained by the chord. The situation is depicted in Fig. 4.12. The relationship between l_z and l_r is by

$$l_r = D \arcsin\left(\frac{l_z}{D}\right) \tag{4.29}$$

Let the pace of traffic internal to the zone be p_z and ring-road traffic p_r. And let $V(p_z, p_r, l_z)$ now give, rather than the net advantage of driving, the net advantage of driving

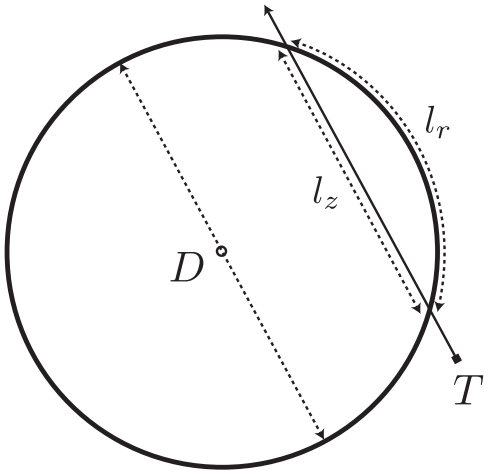

Figure 4.12: Setting for the ring-road model: a circular zone with external trips. Each trip involves a cut-through distance l_z and a circumferential distance l_r.

through the zone instead of along the ring-road for a driver with internal trip length l_z, given p_z and p_r. It follows that

$$V(p_z, p_r, l_z) = D \arcsin\left(\frac{l_z}{D}\right) p_r - l_z p_z. \tag{4.30}$$

Since $D \arcsin(l_z/D) > l_z$ (the arc is always longer than the chord), the ring-road alternative will only be attractive to someone if $p_r < p_z$—that is, if the ring-road is faster than the zone. Now suppose that inequality holds. In this case, $V < 0$ for all $l_z < l_z^0$, where l_z^0 is defined implicitly by $0 = V(p_z, p_r, l_z^0)$, and $V > 0$ for all $l_z > l_z^0$. Figure 4.13 illustrates the contour $V(l; p_z, p_r)$ for $p_z = 3.0$, $p_r = 2.7$ and $D = 2.0$. For these values, all trips with $l_z < 1.41$ will tend to be made by the ring-road, and all longer trips will cut through the zone. Moreover, since $\partial V/\partial p_z = l_z$, most of the equations of Sections 4.2 and 4.4 hold up, and much of the analysis can be repeated. The main challenge to any researcher pursuing this model is that the mode split lines are non-linear and must be computed numerically.

4.7 Conclusion

4.7.1 Summary

This chapter proposed an original, static model of traffic into a downtown with an MFD, in which commuters with varying trip lengths make a probabalistic choice between a fast car and a slow transit option. While this situation seems complex, an application of Little's Law allows the user equilibrium to be derived from curves in pace/circulation space. Another graphical framework for dividing the population between drivers and transit-riders allows for

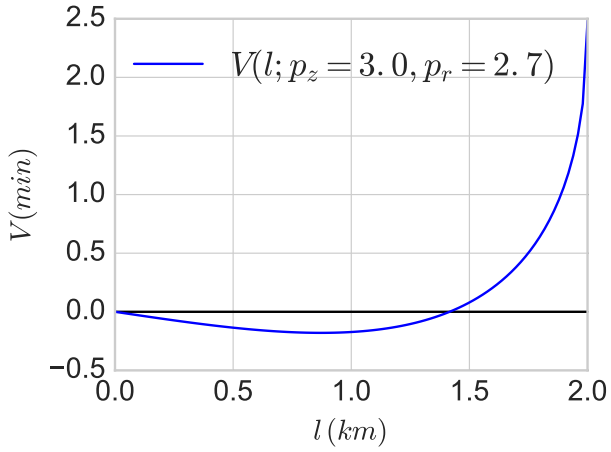

Figure 4.13: The net advantage of cutting through in the ring-road model. Zero-length trips are indifferent. Short trips prefer to take the ring road. Long trips prefer to cut through.

convenient illustration of the different impacts that a distance and a trip toll have. Theory shows that the trip toll cannot raise total social surplus as much as the distance toll (except in the special cases where demand is deterministic or trip lengths invariant), because it allows certain long, low-value trips to be made by car and forces certain short, high-value trips to be made by transit. However, numerical simulations showed the trip toll does nearly as well as the distance toll with regard to total social surplus, and actually improves travelers' surplus more since it raises less revenue. Moreover, the static model turns out to be the steady state of a dynamic model of the rush in which travelers arrive at a constant rate in the downtown.

Before ending, we will look at several policy implications that may be inferred from the model as well as several extensions for future research.

4.7.2 Policy results

One policy result has already been mentioned: in spite of being more efficient from the standpoint of society as a whole, the distance toll is easily worse for travelers than the trip toll. Consequently, even if the considerable technological barriers to distance tolling can be overcome, political resistance will remain high—at least until there is more public appetite for government revenue. For example, in the author's private correspondance with Singapore's Land Transport Authority, officials were quick to note that there are no plans to introduce distance tolling as part of the satellite-based ERP 2.0 system. On a hopeful note, as already mentioned, the trip toll did nearly as well in simulation as the distance toll at raising total social surplus. However, results in real life may vary considerably, especially considering the diversity of trip types (e.g., delivery trips) exhibited by real cities. Only a

detailed survey and demand model will reveal whether moving from a trip to a distance toll is worthwhile for a given city.

Another result has to do with the difference between two uses of transit subsidy. In the chapter's model, since τ_d is added to β in every equation, an investment in transit that raises the convenience or speed of transit (thereby lowering β) is equivalent to a spike in a distance toll. Likewise, since τ_t is always added to γ, a subsidy that lowers the transit fare boosts γ (the difference in the fixed cost of the two modes), and thus functions like a spike in the trip toll. This point is worth making, because there are many investments that cities can make to speed up their transit systems or make the time spent on transit more pleasant: a city can run buses at higher frequency (which speeds up the journey by reducing the number of boardings and alightings per stop), install WiFi or make seats more comfortable. These measures will specifically tend to divert drivers with longer trips, just as the distance toll does, because such travelers spend more time on the transit mode. Notwithstanding the equity consequences, it may be worthwhile for some agencies to raise fares in order to run more frequent service.

A final policy impliciation of the model is that the practice of using arrival flows—that is, flows across the boundary of the zone, as opposed to circulation within the zone—as a metric of toll performance and a target for toll setting is defective. For example, the Stockholm Congestion Tax was designed with the goal of reducing traffic across the cordon by 10-15 percent (Eliasson, 2008, p.396), and the designers of Singapore's Area License Scheme set the prices to reduce inbound traffic by 25-30 percent (Watson and Holland, 1978, p. 12). The problem is that the circulation, density and speed of traffic depend not only on the *magnitude of* the arrival flow but also its *composition*—how many long and short trips it contains—and tolls may effect recomposition. This fact is illustrated by the longer average trip lengths permitted by the trip toll in our numerical simulation. Of course, arrival flows are popular, in part, because they are more readily measured than circulation and density. But it should be practical to measure and target some metric of average speed inside the zone, rather than arrival flows. More philosophically, targets for arrival flow may also be popular because the canonical model of congestion pricing taught in planning programs is the Walters (1961) diagrammatic approach, which treats arrival flow. Some confusion arises with the application of this model to downtown zones, built to describe traffic on a highway, because on a highway all trips entering a segment travel the whole length of the segment, while in a downtown zone trips travel different distances.

4.7.3 Extensions

The model is easy to extend, and several avenues seem promising, in addition to the ring-road model already presented. The most important is probably letting β vary among travelers. While the additive ε parameter used here goes part of the way towards incoporating realistic heterogeneity, letting β vary is essential for recognizing that different people have different values of time. Recognition of this fact would help with understanding the income equity implications of tolling, since the value of travel time savings is related to income.

Another interesting line of research would be to model the transit technology more explicitly. Two effects are important to consider. First, there are scale economies of transit that arise from higher frequency: as people are diverted to the transit mode, the transit authority can afford to run service more frequently, which raises the speed of the transit mode (via reducing boardings and alightings) and lowers the fixed cost of wait times. Small (2004) argues that the London Congestion Charge significantly improved welfare via this avenue. Second, while for the purpose of introducing the model it has been convenient to assume that the transit mode is not slowed down by cars, in practice it almost certainly is affected; even in cities with bus lanes, buses run in mixed traffic for much of their routes.

The last extension is the most complex: comparing the distance toll against an alternative "time toll," which would charge vehicles for how much time they spend circulating in the zone. For the static model, of course, this is a distinction without a difference, since the constant pace of traffic implies a fixed ratio of distance to travel time. But for dynamic models of the sort dealt with in Sec. 4.3 the time toll may have a real advantage: the distance toll charges travelers in the middle of the rush (when density is already high) the same as travelers at the bounds of the rush (when density is low). The time toll would treat them differently. Preliminary simulations indicate that the time toll raises total social surplus somewhat more than does the distance toll.

Chapter 5

Feedback and the use of land for parking

Off-street parking is among the most heavily regulated features of the built environment. In particular, parking minimums—rules requiring developers of new buildings to supply off-street parking in some proportion to the amount of built structure—have been nearly ubiquitous in the United States for decades (Ferguson, 2004; Jakle and Sculle, 2004). But the tide may be turning, and a drive to relax parking minimums is gaining traction. This trend invites questions about how cities would change if today's rules were liberalized or—as proposed in Barter (2010)—replaced by a new regime.[1] A robust and growing literature on parking seeks to provide answers and bring clarity to these questions.

Part of the parking literature is concerned with the actions of individual developers and landowners. Would they really supply much less parking if permitted? Li and Guo (2014) use as a natural experiment a campaign of parking reform—both the elimination of minimums and, in some places, the imposition of maximums—carried out in London during the early 2000s. Comparing nearby developments before-and-after, the authors estimate reforms reduced off-street parking by 0.76 spaces per unit, or 49 percent, and that eliminating minimums was more important than imposing maximums. Similarly, Manville (2013) looks at Los Angeles projects that take advantage of an ordinance change for the reuse of historic buildings, and find developers include less parking when allowed. McDonnell et al. (2011) find that 18 of 38 residential projects surveyed in Queens provided exactly the minimum number of parking spaces, which would be a coincidence were mandates not the limiting factor. Taking a more economic approach, Cutter and Franco (2012) ask whether the marginal parking space adds more value to the structure with which it comes bundled than it costs to provide. For six property types in suburban Los Angeles, the marginal space does not appear to pull its weight, suggesting it exists for compliance.

There is also strand of thought in the parking literature that takes what might be called the "neighborhood perspective." A theme is that, paved surfaces being neither origins (such

[1] For a more thorough discussion of the economics of parking, see Inci (2014).

as homes) nor destinations (such as shops, cafes, bars, groceries, offices, social facilities) unto themselves, off-street parking takes space from the enlivening land uses that make a place walkable. There is a competition between parking and floorspace. Weinberger et al. (2010) argues that, due to parking minimums, "The walking environment is undermined and the distance between destinations increases." Mukhija and Shoup (2006) agree: "Parking lots and garages tend to interrupt the streetscape, expand the distances between destinations, and undermine walkability." Taken far enough, parking minimums become what Willson (1995) calls "tacit policy for automobile use and urban sprawl." On the other hand, implicit in the rules themselves is another argument from the neighborhood perspective: without abundant off-street parking, vehicles will crowd the streets and overflow among neighbors' private lots.

This chapter is based on a forthcoming paper, Lehe (2017), that aims to theoretically link the two perspectives. Under consideration are city neighborhoods meeting two criteria: First, the amount of off-street parking can feasibly affect the neighborhood's walkability and crowding. Excluded, then, are places where land is so lightly developed that parking does not meaningfully limit density, and where on-street parking is plentiful for any level of off-street parking. Second, the neighborhood is subject to sequential land-use change in which lots are developed or redeveloped on occasion. When the time is right, a property-owner or developer may rebuild a dilapidated house, construct new apartments on an under-used lot or undertake a small finesee, such as turning a garage into an in-law unit. Both student neighborhoods and the older, residential areas of American cities are often good examples of the sort of area we have in mind.

For a neighborhood of this kind, the chapter considers the idea that there may plausibly arise a "feedback loop" between individual decisions and neighborhood outcomes. This feedback loop is a relationship of mutual influence: The individual developer's choice of how much off-street parking to supply influences neighborhood characteristics, which, in turn, influence that choice. Other features of urban life and land-use are rich in such chicken-and-egg links between gestalt properties and individual choices. An entrepeneur opens the tenth bar in a popular nightlife district because the crowds are there, and people go there for a night out precisely because there is such a variety of bars. It may be profitable for planners to consider feedback with respect to the parking supply, too.

One reason to look at feedback in parking is to enrich the conceptual vocabulary by which parking is understood. Urban neighborhoods can evidently function with widely different parking intensities—from relative parking deserts in Philadelphia and Boston to lands of milk and honey in Las Vegas and Phoenix. An appreciation of feedback invites interesting questions about this state of affairs. With parking minimums liberalized, is it possible that a parking-abundant, mildly-walkable neighborhood could, over the course of decades, become more like a parking-scarce, highly-walkable neighborhood—that a neighborhood in, say, Atlanta could become a little more like one in, say, Pittsburgh?

A second reason is to recognize that policy results can be hard to predict from even a thorough study of the status quo. Suppose, for instance, that a planner surveys recent projects in a neighborhood and learns that developers would have rather built only about

half the spaces required by law. It is tempting to deduce that, if the law were abolished, new projects would in fact turn out to have about half as many spaces. But while today's developers may build projects with only half as many spaces, once enough of them do so they will have collectively somewhat changed the neighborhood character, made it denser. Consequently, future developers may want to supply either more spaces or fewer. Perhaps developers will come to see parking as an underserved niche they can enter profitably. On the other hand, perhaps they will decide the neighborhood is becoming the sort of walkable place where people can make do with considerably less parking. These later decisions, in turn, will "feed back," shaping still later ones.

The chapter unpacks the above logic in the following way: First, Section 5.1 discusses qualitatively the idea of a feedback loop in parking provision—what mechanisms might drive it and to what effect. Next, Section 5.2 proposes an aspatial model in which feedback arises. Section 5.3 provides discussion.

5.1 The concept of feedback

The concept of a feedback loop has wide purchase across physical and social sciences. In engineering, it describes a system whose output is also one of its inputs—e.g., when the microphone picks up the sound of the speaker to which it is connected. Here, the feedback loop describes a relationship of causation between two quantities: (i) the amount of parking in a neighborhood; (ii) the amount of parking that individual property-owners or developers (hereafter referred to as "landlords") wish to provide upon redevelopment. One direction of causation is arithmetically true: how much parking individuals decide to provide affects the aggregate parking supply. The other direction of causation hinges on a question: how does the neighborhood influence the individual? Depending on the nature of that influence, feedback can be of either a "positive" or "negative" character.

5.1.1 Positive feedback

Positive feedback in general describes a process that is self-reinforcing—such as the example of a microphone cited above. It has long been of use to social scientists in explaining the multiple equilibria, path dependence, tipping points and other interesting phenomena observed in real societies—particularly the remarkable complexity and discontinuity of human geography. Widely-cited examples include racial segregation (Schelling, 1971), educational segregation (Benabou, 1993), differences in the frequency of riots among places (Granovetter, 1978), and manufacturing belts (Krugman, 1990).

Positive feedback here means a rise in the neighborhood's parking supply makes the individual landlord want to provide more parking. The converse holds, too: the scarcer parking becomes, the less the landlord wants to supply. Why should this happen? Setting aside for now facts about crowding and a market for parking, the force driving positive

feedback could be the same one that has invited concerns about a surfeit of parking: its effect on walkability, on the pedestrian vibrancy of the place.

It is hard to say with any precision how much land in cities is devoted to parking, and even statistics widely quoted may be unreliable (Manville and Shoup, 2005). Looking at several studies of different cities, McCahill and Garrick (2014) report estimates that between 10 and 40 percent of CBD land is parking. In any case, parking clearly consumes much land that could be otherwise devoted to origins or destinations. Thus, its prominence in the landscape constrains the count of destinations reachable from an origin by a reasonably short trip on foot or by bike. And by the same token, the aggregate amount of land devoted to parking also constrains the number of customers who can be pedestrian patrons of a destination.

There are several ways that walkability could reasonably cause a landlord to devote less space to parking and more to another use. Perhaps floorspace rents and parking rents rise as walkability increases, but floorspace does so more quickly. Or, per the Cutter and Franco (2012) model in which parking enhances the floorspace with which it comes bundled, parking might not "pull its weight" as much in a walkable area; it might not add enough value relative to its cost and the floorspace it displaces. Lastly, on land that is owner-occupied, a rise in walkability might lead the owner to do without the third car and build an extension over part of the driveway. And even when current owners do not plan these alterations, buyers who do wish to make them will raise their bids as walkability rises.

Note the difference between the sort of positive feedback advanced here and another, more "top-down" conception. Jacobs (1961, p. 350) explicitly cites positive feedback to explain the "erosion of cities" by automobile space—including parking. The principals in Jacobs' story are bureaucrats and civic leaders who endlessly convert land from other uses to parking and streets. Positive feedback describes the ironic way their interventions frequently wind up making traffic worse, leading the interventions to be iterated. Similar theories of positive feedback at the government level—though not necessarily stated in the language of feedback—appear in Shoup (2005, Ch. 5). The question here, by contrast, is what happens if the government strengthens or weakens a certain policy and lets things evolve, not what happens if the government keeps ramping up or down the policy. Change occurs as a side-effect of uncoordinated private action. The actors are landlords uninterested in, and likely ignorant of, what their own parking does to urban form.

5.1.2 Negative feedback

Negative feedback generally describes a process that is self-stabilizing, wherein a deviation in the output tends to undermine itself. An example is the body's temperature control system: when temperature deviates from the body's target level, the body takes note and either sweats or shivers to return to the target. Negative feedback here means a rise in the parking supply discourages the individual landlord from providing parking, whereas a fall inclines the landlord to provide more.

An obvious mechanism for negative feedback is supply-and-demand in a place where parking is traded as a priced commodity: a boost in supply sinks parking rents, muting the

incentive to add parking; and a scarcity of parking in a dense neighborhood will raise the price of parking, making it profitable to provide. Where parking comes bundled with floorspace, negative feedback can still plausibly arise via crowding: on-street parking becomes harder to find when the neighborhood lacks off-street parking, because people park on the street more often. Reasonably, this crowding means that the marginal parking spot adds more value to the bundled floorspace than when an on-street space is always ripe for the taking nearby. Thus, crowding makes profitable a higher ratio of parking to rentable floorspace.

5.2 Economic model

Above, Krugman (1990) was mentioned as a famous application of positive feedback concepts in geography. Decribing its goals, Krugman (2011) says he set out to create, "a model that was not intended to be realistic, indeed was aggressively unrealistic, but would serve as a demonstration." This section, similarly, works through a highly stylized model of parking provision. While this tack is unusual for the planning literature, it has advantages that compensate for strong assumptions. The model serves as an example—like the frictionless surface of introductory physics—that is easy to work with. It permits a graphical analysis in which certain qualitative ideas—such as equilibrium, change and whether requirements bind—can be spoken of concretely and memorably.

Section 5.2.1 establishes the basic assumptions of the model and shows how landlords decide how much on-street parking to provide. It is a Cutter and Franco (2012) situation in which parking adds value to bundled floorspace. The bundled case is chosen because it is more interesting than the simple mechanics of supply-and-demand, and because so much parking is bundled. Weinberger et al. (2010) reports that developers have even tried to evade a San Francisco initiative that requires unbundled parking, by offering special pricing to owners of associated units. Sections 5.2.2 and 5.2.3 model assumptions about demand that lead to positive and negative feedback, respectively. In Sec. 5.2.4, we discuss how equilibria and long-term change happen in the model's context. Lastly, Sec. 5.2.5 looks at several policies.

Note that a real neighborhood can and likely will have both forces of positive and negative feedback at the same time. We treat the all-or-nothing situations only for clarity of exposition.

5.2.1 Basic assumptions

Privately-owned land is divided into small parcels owned by landlords who are also, potentially, developers. It is assumed that all structures are β stories tall, that landlords provide only what yardspace is mandated by setback rules and that parking is at ground level beside the structure. See Figure 5.1 for a diagram of a parcel, in which the trees are setbacks. Consequently, the developable area of the neighborhood is split between off-street parking and structures, and for the neighborhood to gain x sq. ft. of parking it must lose βx sq.

ft. of floorspace. Let the *parking ratio*, p, be the ratio of parking space to floorspace on an individual parcel. It follows that $1/(1+p\beta)$ is the fraction of a parcel's developable land devoted to structure, while $p\beta/(1+p\beta)$ is the parking fraction. Let $\bar{p}$ stand for p's aggregate equivalent: the ratio of parking to floorspace for the whole neighborhood.

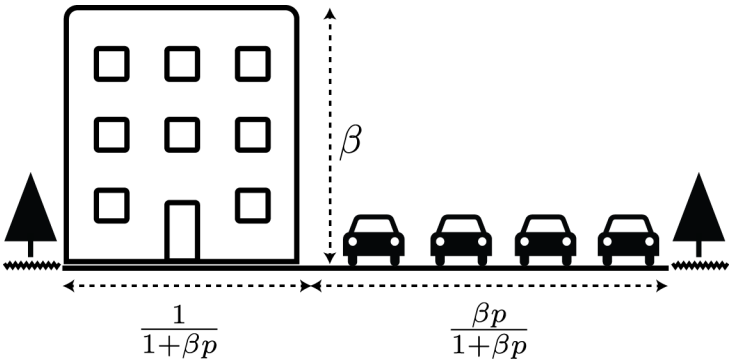

Figure 5.1: Division of a parcel between structure and parking. Greenspace, represented by trees, is fixed and thus omitted from analysis.

The "rent" per unit area of floorspace is R. This is not a literal rent; it is the amortized return after costs.[2] The rent depends partly on p, the parking ratio of a single parcel. Parking cannot be practically rented or sold as a commodity; rather, it comes bundled with floorspace for use by the parcel's occupants. The job of parking is to boost the rent of floorspace, and it is assumed to do so with diminishing marginal returns. Thus, we can write rent as a function $R(p)$, whose derivatives have the signs $\partial R/\partial p > 0$ and $\partial^2 R/\partial p^2 < 0$.

Landlords choose p to maximize profit. Since only floorspace earns money, and the amount of floorspace per unit area of land is $\beta/(1+p\beta)$, it follows that profit, π, per unit area of land, is

$$\pi(p) \equiv \frac{\beta}{1+p\beta} \cdot R(p). \tag{5.1}$$

This expression shows the landlord's fundamental trade-off when parking is bundled: adding parking (raising p) boosts the rent but at the cost of losing floorspace, which appears as an increase in the denominator. The derivatives of $R(p)$ give $\pi(p)$ the hump shape in Fig. 5.2. There is a unique, profit-maximizing parking ratio, p^*. As p grows ever larger, profit falls off to zero because there is no floorspace to rent out at all. In Fig. 5.2, the peak occurs at a non-zero p^*, but it is also possible that it is optimal to provide no parking at all.

In addition, the rent of floorspace depends—directly or indirectly—upon the neighborhood parking ratio $\bar{p}$. Depending on how it does so, a change in $\bar{p}$ may change the profit-maximizing parking ratio, p^*, that the landlord would like to have. Graphically, this means that a change in $\bar{p}$ shifts the peak of $\pi(p)$ to the right or left. If we calculate the optimal

[2]Since only returns—the margin of income over costs—matter to the developer's decision, treating real rents and costs would add complication without obtaining different results.

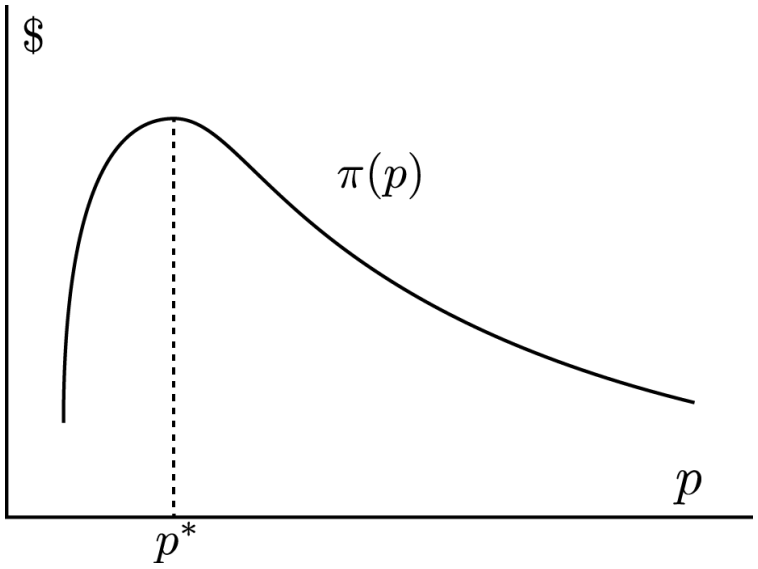

Figure 5.2: Profit function $\pi(p)$. There is one profit-maximizing value of the parking ratio.

parking ratio, p^*, for each aggregate ratio, $\bar{p}$, we wind up with a *best-response function* $p^*(\bar{p})$. The best-response function tells what parking ratio a landlord would like, given the aggregate ratio is $\bar{p}$. When there is positive feedback, $p^*(\bar{p})$ rises with p; the peak of $\pi(p)$ shifts rightward as $\bar{p}$ increases. When there is negative feedback, $p^*(\bar{p})$ falls with p; the peak of $\pi(p)$ shifts leftward as $\bar{p}$ increases.

5.2.2 Positive feedback assumptions

The story we will tell to obtain positive feedback is that, when a neighborhood is dense in destinations, rents become high enough—per unit area of floorspace though not necessarily per apartment or shop—that a lower parking ratio is optimal. Closing the loop is the fact that, when the neighborhood gains floorspace, it gains occupants who can financially support a richer ecosystem of destinations.

Suppose the count of walkable destinations in the neighborhood, w, rises with aggregate floorspace in the following way: for every unit of floorspace beyond a minimum area α, the neighborhood gains one destination. It is assumed that the destinations themselves—at least the ones who are drawn in by more occupants—do not consume a very significant fraction of the neighborhood's area. Let A be the area of developable land of the neighborhood, such that $\beta A/1+\bar{p}$ is the aggregate supply of floorspace. We may write a function

$$w(\bar{p}) \equiv \max\left\{0, \frac{\beta A}{1+\beta\bar{p}} - \alpha\right\}, \tag{5.2}$$

where the max operator precludes negative destinations.

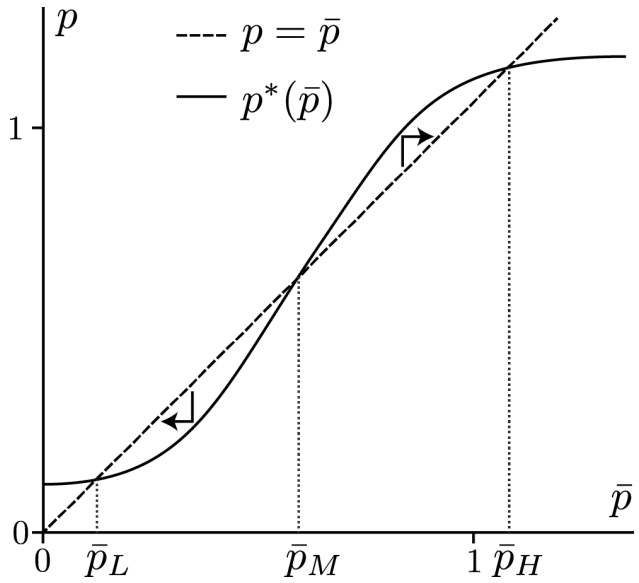

Figure 5.3: best-response curve with positive feedback;
$R_0 = 1.8$, $\alpha = 2$, $A = 4.7$, $\beta = 1$

Next, suppose rent is given by the function

$$R(p, w) \equiv R_0 - \frac{1 + \exp\left(-w^2\right)}{1 + \beta p}. \tag{5.3}$$

This function has been chosen to obtain illustrative results but is not especially meaningful in its own right. R_0 is a positive constant that can be thought of as a sort of "base rent." The second term rises in a diminishing way with p (so that tenants value parking), and it rises with w (so that tenants value access to destinations). To find the best-response function, we plug $R(p, w(\bar{p}))$ into (5.1) and seek the profit-maximizing p for every $\bar{p}$. For the right values of A, R_0, β and α, $p^*(\bar{p})$ is S-shaped. Such a curve appears in Figure 5.3 alongside the 45° line $p = \bar{p}$. Positive feedback is manifest in $p^*(\bar{p})$'s positive slope.

5.2.3 Negative feedback assumptions

The story we will tell to obtain negative feedback is that on-street and off-street parking are substitutes, and that, as $\bar{p}$ declines (as the neighborhood becomes denser in floorspace and scarcer in parking), finding on-street parking involves more cruising and a longer walk. This crowding causes the rent to rise and fall with $\bar{p}$. Specifically, let rent be given by

$$R(p, \bar{p}) \equiv p^{1/2} + \bar{p}^{1/2}. \tag{5.4}$$

Each term captures the positive, but marginally diminishing, value of on-street and off-street parking. Figure 5.4 shows the best-response function $p^*(\bar{p})$ derived from (5.4), again when $\beta = 1$. Negative feeback is manifest in the negative slope of $p^*(\bar{p})$.

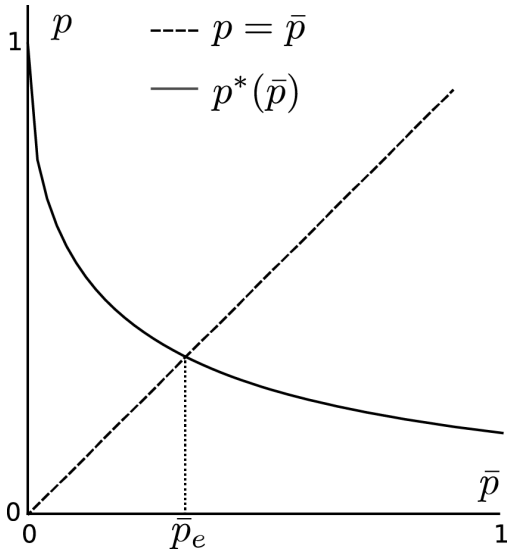

Figure 5.4: Best-response curve with negative feedback. Because $p^*(\bar{p})$ slopes downward, it can intersect $p = \bar{p}$ only once.

5.2.4 Neighborhood change

Since a neighborhood's characteristics arise from the sum of individual decisions, redevelopment opportunities lead to the possibility of neighborhood-level change. An advantage of our formal model is clarity about how this change may occur. Three ideas are discussed: (i) how to predict the direction of change, (ii) equilibrium and (iii) stability.

The process of change, of course, is liable to happen very slowly and possibly be subject to fits and starts. Buildings are very long-lived assets; they cannot be tailored in real-time to always fit what would be optimal to build today if the landlord were starting from scratch. But by the varied types of redevelopment we have proposed, it is possible to speak of general directions of change—densification, clearing out—without pretending that parking supply wavers like the stock of orange pulp in a warehouse.

5.2.4.1 The direction of change

To understand the direction of change, consider Figures 5.3 and 5.4 again. Whenever the line of the best-response function is below the 45° line $p = \bar{p}$, it is true that $p^*(\bar{p}) < \bar{p}$, which means the individual landlord chooses a below-average ratio. Granted that this is permitted, the redeveloping landlord will do so. And, just as the average height of a team falls when a shorter-than-average player joins, when one landlord chooses a less-than-average parking ratio, the aggregate ratio falls slightly. Therefore, if the neighborhood is ever found to have a $\bar{p}$ where $p^*(\bar{p}) < \bar{p}$, then the on-going process of redevelopment will slowly bleed the parking supply. And when $p^*(\bar{p}) > \bar{p}$, the neighborhood will take on parking, because landlords choose an above-average ratio.

A lesson is that whether the neighborhood gains or loses parking is not a matter of having "too much" or "too little" parking, relative to similarly-situated neighborhoods or to what intuition would suggest. Whether the neighborhood is liable to gain or lose parking depends on the forces at play for the particular parking ratio at which the neighborhood finds itself. Consider the neighborhood of Fig. 5.3, the positive feedback case. If the neighborhood has a ratio between $\bar{p}_L$ and $\bar{p}_M$, then some parking will disappear every few years. But if it has even *more* parking than this, if it has a $\bar{p}$ between $\bar{p}_M$ and $\bar{p}_H$, then new parking will spring up. Thus, between these two intervals, adding to the parking supply raises the demand for parking.

5.2.4.2 Equilibrium

An equilibrium is said to arise when no landlord wants to change her parking ratio. This may happen in the model under two circumstances: First, nearly every landlord has nearly the same parking ratio, in which case $p \approx \bar{p}$ for every landlord. Second, the universal parking ratio is a value at which at which $p^*(\bar{p}) = \bar{p}$. At ratios satisfying this condition, redevelopment does not change the aggregate.

Graphically, equilibrium parking ratios can be identified in our plots as the locations of intersection between the best-response curve and the 45° line. There are three in the positive feedback case of Fig. 5.3: p_L, p_M and p_H. There is only one, $\bar{p}_e$, in the negative feedback case of Fig. 5.4. Purely negative feedback can only allow one equilibrium, because, if the best-response curve always falls, then once it crosses the 45° line it can never rise to cross again. Positive feedback enables infinitely many crossings, depending on how the best-response curve wiggles as it rises, though the particular case presented here involves just three.

5.2.4.3 Stability

While there may be many equilibria, only those that can survive a very slight deviation—such as a few landlords not exactly following the rules of our model—are called "stable." Stable equilibria in our examples are $\bar{p}_L$ and $\bar{p}_H$, from the positive feedback case, and $\bar{p}_e$, from the negative one. These are cases where the best-response function crosses the 45° line from above, where $p^*(\bar{p})$ exceeds $\bar{p}$ to the left of the intersection and drops below $\bar{p}$ to the right. This sort of crossing means that, if the aggregate ratio $\bar{p}$ is slightly higher than $\bar{p}_H$, $\bar{p}_e$ or $\bar{p}_L$, then the neighborhood will lose parking and drop down to the equilibrium ratio; and if the aggregate ratio is slightly lower, then the neighborhood will gain parking and rise up to the equilibrium ratio.

Equilibria that unravel easily are called "unstable." The middle ratio $\bar{p}_M$ is unstable. If $\bar{p}$ winds up slightly higher than $\bar{p}_M$, then landlords will begin to choose greater-than-average parking ratios, and they continue to do so as long as $p^*(\bar{p}) > \bar{p}$. Change will continue until the stable equilibrium $\bar{p}_H$. Conversely, if $\bar{p}$ deviates just left of $\bar{p}_M$ (where $p^*(\bar{p}) < \bar{p}$), landlords will provide a less-than-average amount until the neighborhood levels out at $\bar{p}_L$.

In reality, no neighborhood will sit in a stable or unstable equilibrium, however conceived, for very long. Tastes, costs and larger market forces are in constant flux, and their irregular tides mean, for our model, constant shifts in the R and p^* curves. Nonetheless, the concept of stability can be a useful ideal. A stable equilibrium, even one in motion, suggests what direction the neighborhood is likely to move even if it never quite stops. And the possibility of multiple stable equilibria highlights to the importance of history-dependence—that the neighborhood is only in its current state because it was there yesterday, not because the status quo is the only way things could have practically turned out without planning.

5.2.5 Policies

So far, Sec. 5.2 has set up a neighborhood and asked how it might be expected to change and to stop changing in the absence of policy. Fortunately, the model can usefully adapted to the formal analysis of policy, and doing so yields several insights. Below, three policies will be examined: (i) parking requirements, (ii) parking taxes and (iii) direct coordination. Somewhat more space is devoted to the effects of policies in the positive feedback case. The reason is that results of positive feedback are more nuanced, not necessarily more realistic.

5.2.5.1 Parking requirements

While parking minimums are most common, some places—such as San Francisco and Manhattan—have maximums. These requirements can be treated as either floors or caps on the value of p that a landlord is allowed to choose. Real parking requirements in many cases are written exactly so, as a number of spaces that must be provided per 1000 sq. ft. of floorspace.

As is well known, just because there exists a requirement does not necessarily mean it has any meaningful effect on land-use. Requirements only bind—that is, the landlord only wishes to do something differently than he or she is allowed to—in certain cases. A minimum binds when landlords would like to supply less parking than allowed; a maximum, when they would like to supply more. Evidence of minimum parking requirements binding was mentioned in the introduction. The parking maximum in Manhattan must be binding for certain landlords, because they have organized to have it relaxed (Manville et al., 2013, p. 363).

Parking requirements are nested graphically in Fig. 5.5 and Fig. 5.6. A minimum $p_{\min}$ can bind if it falls where $p^*(\bar{p})$ is below the 45° line $p = \bar{p}$. In this case, if every parcel has a parking ratio near the minimum (so that $\bar{p} \approx p_{\min}$), then landlords will want smaller ratios than $p_{\min}$. Similarly, a maximum $p_{\max}$ can bind when $p^*(\bar{p})$ exceeds $p = \bar{p}$.

This way of looking at mandates has two main lessons. First, in order for a requirement to actively bind, the neighborhood must somehow reach the point where individual landlords wish to break it. In Figure 5(b), $p_{\min}$ is potentially binding, but if the neighborhood has settled in at the equilibrium $\bar{p}_H$ then landlords will gladly supply much more than $p_{\min}$. This is possible not only for this model but for any situation where there are multiple equilibria.

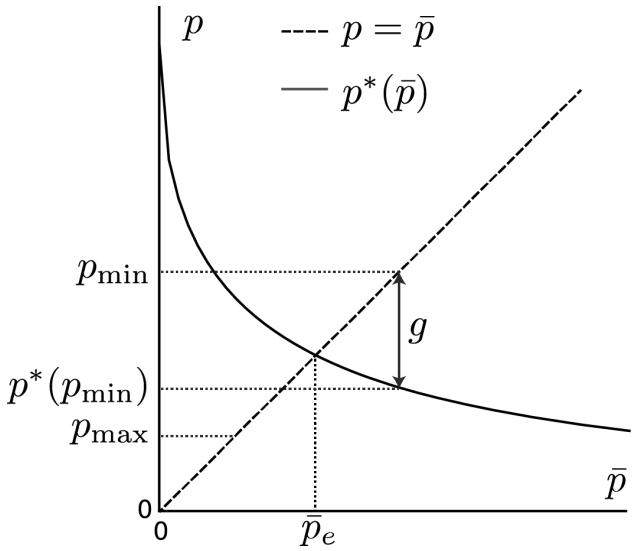

Figure 5.5: Binding parking requirement, negative feedback. Long-run effect is smaller than static effect.

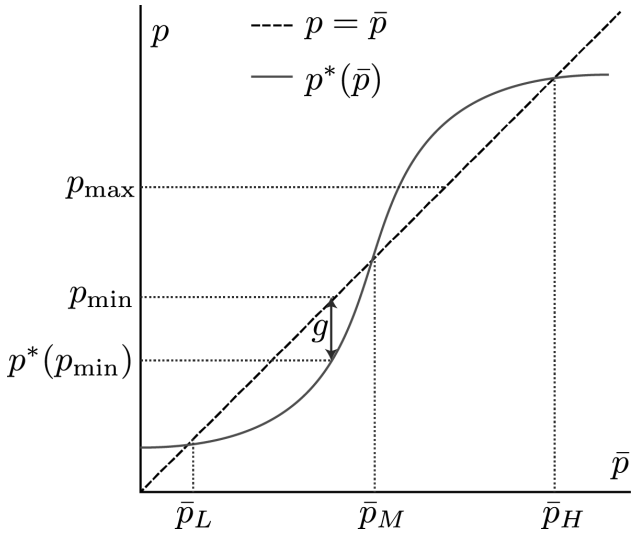

Figure 5.6: Binding parking requirement, positive feedback. Long-run effect is larger than static effect.

What might be called a requirement's "bindingness," then, is not an essential feature of the requirement itself; it has power only in a specific history of a particular location.

Second, the long-run effect of a requirement may be substantially stronger or weaker than what the individual landlord's current point-of-view would suggest. The span g on Figures 5.5 and 5.6 measures the gap between $p^*(p_{\min})$ and $p_{\min}$: how much parking a landlord provides to meet the minimum when the neighborhood has settled at the minimum. In 5.5, landlords would rather supply much less parking (large g). But were the minimum abolished, the neighborhood would converge not to $p^*(p_{\min})$ but to $\bar{p}_e$, which is much nearer to $p_{\min}$ than is $p^*(p_{\min})$. Thus, the long-run effect with negative feedback is weaker than g in the status quo. Some of the parking that one landlord decides not to provide will be provided by a different landlord. The long-run effect with positive feedback, on the other hand, may be much stronger than g. Figure 5.6 shows that, if the minimum were abolished, the conversion of parking to floorspace would compound until the neighborhood reached $\bar{p}_L$. As an example, suppose that, after the parking minimum in some neighborhood is weakened, some developers construct apartment complexes with many residents and little parking, and that the extra population draws a major grocery store. Once the grocery store is in place, future developers will be even less inclined to waste floorspace on parking, because rents will have risen somewhat, and customers with fewer cars will more often take an interest in the apartments.

5.2.5.2 Parking tax

A tax on parking area is advocated in Feitelson and Rotem (2004) as a means of internalizing driving externalities and the heat absorption and polluting runoff of parking surface. To include a tax in our models, let τ be the amount of the tax. Since the tax applies per unit area of parking, the developer pays a tax on the parking fraction, $\beta p/(1+\beta p)$ of the parcel. Therefore, profit is

$$\pi = R \cdot \frac{\beta}{1+\beta p} - \tau \cdot \frac{\beta p}{1+\beta p}. \tag{5.5}$$

The best-response function $p^*(\bar{p})$ is then found in the usual way: by finding the profit-maximizing level of p for every $\bar{p}$.

For both R above, the tax shrinks p^*, which is probably unsurprising. More noteworthy is the fact that, with multiple stable equilibria, the size of the effect of the tax may jump when the tax rises only slightly. There can be a "threshold effect" or a "tipping point." Examples of neighborhood thresholds for other variables are reviewed in Quercia and Galster (2000).

To illustrate a threshold effect, Fig. 5.7 show the best-response curve for three levels of the tax in the positive feedback case. The neighborhood begins at the equilibrium $\bar{p}_0$ and with $\tau = 0$, for which $p_0^*(\bar{p})$ is the best-response curve. When the tax rises slightly to τ_1, the best-response curve falls to p_1^*, and, over time, the parking ratio falls to $\bar{p}_1$. The result is a somewhat, but not substantially, denser and more walkable neighborhood. When the tax rises again to τ_2, the best-response curve shifts downward to $p_2^*(\bar{p})$. Remarkably, even if $p_2^*(\bar{p})$ is only slightly lower than $p_1^*(\bar{p})$, the move produces a seismic change: there is only one

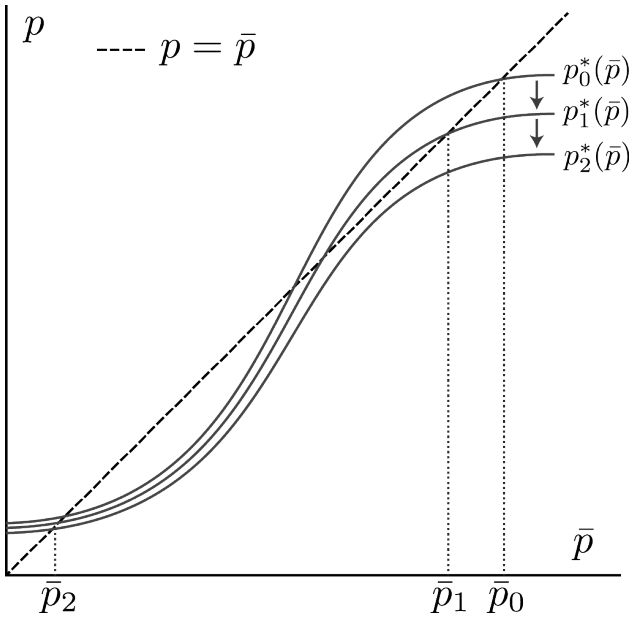

Figure 5.7: Effect of parking tax on best-response. In moving from tax level 1 to 2, the high-parking equilibrium ceases to exist, and the neighborhood will leap to equilibrium $\bar{p}_2$.

equilibrium, at $\bar{p}_2$, and it involves only a fraction as much parking as $\bar{p}_1$. The reason is that, in moving from $p_1^*(\bar{p})$ to $p_2^*(\bar{p})$, the upper part of the curve fell below $p = \bar{p}$, not because the tax increase from τ_1 to τ_2 was especially large.

In a regime with negative feedback, on the other hand, there can be no such jumps, because only one equilibrium an possibly arise for any given level of the tax.

5.2.5.3 Coordination

The above two policies influence landlords as individuals. It is also possible, in the case of multiple equilibria, for an authority to directly coordinate all parcels toward a desired equilibrium. The most straightforward way to do so is to develop a large number of empty parcels, all at once, to achieve an aggregate parking ratio in the vicinity of a chosen stable equilibrium. Planning movements such as New Urbanism could be be thought of as attempts to begin at a walkable equilibrium.

Planners could also attempt to shift an established neighborhood to a new equilibrium by providing *sunspots*. This term from macroeconomics refers to "phenomena that do not affect tastes, endowments or production possibilities" but that nonetheless have real effects by synchronizing beliefs (Cass and Shell, 1983, p. 193). For example, an upbeat government report on the economy may induce businesses to invest and hire, even when it contains no original news. Businesses invest because they believe others will do so.

Similarly, municipal policies could act as sunspots, coordinating landlord's beliefs about the future state of the neighborhood. Consider the construction of a streetcar. Even if it fails

to provide a good substitute for car travel, its presence may induce landlords to provide less parking by fostering a general understanding that the neighborhood is headed in a walkable direction. Landlords will thus provide somewhat less parking in expectation of this state of affairs, and in doing so fulfill the prophecy of a walkable neighborhood. It is not even essential that landlords have faith in the streetcar as a viable transportation option, only that they believe other landlords associate the streetcar with a walkable land-use pattern. Of course, a streetcar alone could never shoulder such a large burden, but the point is that its contribution, as part of a policy package, could be its power as a coordinator of private expectations.

5.3 Discussion

This chapter has drawn attention to the possibility of a feedback loop in the provision of off-street parking for neighborhoods where development is on-going and off-street parking may substantially alter urban form. After an introduction, Sec. 5.1 discussed feedback qualitatively, offering ways it could arise and distinguishing positive from negative feedback. We showed that walkability externalities can foster positive feedback, and crowding externalities negative feedback. Section 5.2 set up a stylized economic model of bundled parking, showed how individuals decide how much parking to provide in the model and inserted several policies.

The examples given can certainly not be expected to apply everywhere, and the actual process of neighborhood change rests on considerably more than the parking supply. Still, if academics and planners are trying to predict the course of a neighborhood or to explain the course one has already taken, the feedback potential of the competition between parking and floorspace could be worth considering. Parking is a major part of the cityscape, and developers do not make their decisions about parking in a vaccum. The examples have tried to make clear, in a concrete way, how the neighborhood's character and the small choices that decide it may be inexorably and plausibly tied together.

While the chapter has mainly formalized policies, not judged them, the analysis should serve to sound a note of epistemic caution. Under simple assumptions, policies have been shown to have consequences that are difficult to predict from a survey of static conditions and interests. This point may be read as favorable to intervention in some cases and hostile in others. Against intervention is the argument that, though they may target real externalities like cruising for parking or spillover into private lots, parking regulations that carry even a slight danger of unwraveling or precluding a good land-use equilibrium merit scrutiny. In favor of intervention is the argument that past policy failures do not always foreshadow future results: a policy—such as a parking tax—may have very modest results until some threshold is breached.

In closing, an interesting question presents itself: under what conditions does each externality—walkability or crowding—dominate and lead to the sort of feedback we have assigned to it? It is possible for both to impact development decisions at the same time, but

one force may easily swamp the other. Intuitively, we can deduce at least a few qualities of a neighborhood that are likely to privelege or mute one of the two effects.

First, some neighorhoods have policy in place that nullify the externalities. Neighborhoods—such as many subdivisions—where land-use regulations or private covenants make the supply of "destinations" highly inelastic, where it is impossible to establish many new cafes and shops among the houses, will not exhibit positive feedback—at least not by way of the mechanism envisioned above; in such a place, the streets will simply become more crowded as developers forgo off-street parking. Similarly, if a city were to ration on-street parking by limiting the number of resident passes granted or pricing on-street parking competitively, the crowding mechanism could not be said to apply.

Second, it is probably that case that the externality itself depends on the existing density of development in a non-linear way. For example, extremely dense urban neighborhoods, such as Lower Manhattan, where unpriced on-street parking is rare and thoroughly crowded, are unlikely to be affected by the crowding externality of additional development; there, residents are very unlikely to own cars or, if they do own cars, to plan to park their cars on the street. Thus, a rise in the demand for on-street parking will not affect rents much. By the opposite token, very sparse neighborhoods—such as a rural community—are unlikely to be affected by either walkability or on-street crowding, because few people park on the street at all and the population is spread too thinly for even a major increase to support more destinations. The most likely setting for our effects, then, is a neighborhood that falls in a Goldilocks zone of density—not too crowded but not too sparse—where both externalities have a substantial role to play.

Chapter 6

Conclusion

Transportation engineering must necessarily involve analysis at both low and high scales. Its topics range from the measurement of cracking in pavement, to the optimal frequency of bus service on a route, to the long-range plans of cities and regions, to the allocation of the national transportation budget. For the higher scales, to make sense of interconnected phenomena, an aggregated perspective is necessary. Models taking an aggregate perspective are often *aspatial*, in the sense that they do not take into account individual interactions among vehicles using the infrastructure at the same time. Rather, they rely on summary measures. This dissertation has shown the value of aspatial modelling to understanding two policies: zone pricing (downtown congestion pricing) and parking policies, of which the former topic consumed the majority of the dissertation.

The dissertation has made several contributions to the zone pricing literature. Chapter 2 gave the most recent in-depth review of real schemes. There has been a need for a new review of zone pricing, since nearly all of the systems in operation have come online in the past fifteen years. Chapter 4 proposed an original, static model of travel into a downtown zone with pricing. This model has been the first study to incorporate the following elements: probabalistic demand, variable trip length, Macroscopic Fundamental Diagram traffic physics and an agent-based model that tracks traveler's progress through the zone. Somewhat remarkably, in spite of these considerable complications, the gist of the model could be captured in readable diagrams—one being the traditional supply/demand diagram of transportation economics and the other an original way to capture mode split between car and transit. Another important theoretical result is that an agent-based model of a dynamic rush with constant arrival rates converges to a steady state which is also the equilibrium of a simpler static traffic model, in spite of the randomness and variability involved. In addition to these theoretical points, Chapter 4 also contains practical lessons for the design of real zone pricing schemes. A toll that prices vehicles for how far they travel (a "distance toll") is shown to improve total economic welfare more than can one that prices all vehicles in the same way (a "trip toll"). However, it turns out that opting for a distance over a trip toll may not actually be desirable for a particular city. Because while the distance toll improves social welfare more, a great deal of those welfare gains go to the government in the form of toll

revenues; the trip toll, by contrast, leaves more of the welfare gains in the hands of drivers. Consequently, whether a city should adopt distance pricing depends on current taxes, the availability of public projects with good benefit-cost ratios and the public's trust in the government to undertake those projects competently. Moreover, in numerical simulation the trip toll proved to improve total welfare by almost as much as the distance toll.

Regarding parking policy, Chapter 5 has made several original points. Most generally, it has brought formal modelling to bear on a topic—parking and land-use—that has attracted little algebraic modelling but enjoys an enormous planning literature of a descriptive or statistical nature. More specifically, it has provided an explanation (feedback) for why neighborhoods exist for long periods of time with utterly different amounts of parking per household, shown that the widely-acknowledge forces of walkability and the crowding of on-street parking can beget feedback loops, and that extant policies, even apparently modest regulations and taxes, can under certain circumstances have long-term effects much larger than a survey of present conditions would suggest to the local planning department.

Bibliography

Anas, A. and Lindsey, R. (2011). Reducing urban road transportation externalities: Road pricing in theory and in practice. *Review of Environmental Economics and Policy*, 5(1):66–88.

Ardekani, S. and Herman, R. (1987). Urban Network-Wide Traffic Variables and Their Relations. *Transportation Science*, 21(1):1–16.

Arnott, R. (2013). A bathtub model of downtown traffic congestion. *Journal of Urban Economics*, 76:110–121.

Arnott, R., Rave, T., and Schöb, R. (2005). *Alleviating Urban Traffic Congestion*, volume 1.

Barter, P. A. (2010). Off-Street Parking Policy without Parking Requirements: a Need for Market Fostering and Regulation? *Transport Reviews*, 30(5):571–588.

Beckmann, M., McGuire, C., and Winsten, C. B. (1956). *Studies in the economics of transportation*. Published for the Cowles Commission for Research in Economics by Yale University Press, New Haven.

Benabou, R. (1993). Workings of a city: location, education, and production. *The Quarterly Journal of Economics*, 108(3):619–652.

Borins, S. F. (1988). Electronic road pricing: An idea whose time may never come. *Transportation Research Part A: General*, 22(1):37–44.

Börjesson, M. and Kristoffersson, I. (2015). The Gothenburg congestion charge. Effects, design and politics. *Transportation Research Part A: Policy and Practice*, 75:134–146.

Branch, E. (1979). Keeping Hong Kong Moving: the White Paper on Internal Transport Policy. Technical report, Hong Kong.

Cass, D. and Shell, K. (1983). Do Sunspots Matter? *Journal of Political Economy*, 91(2):193–227.

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

Chin, K.-K. (2010). The Singapore experience: The evolution of technologies, costs and benefits, and lessons learnt. *OECD/ITF Joint Transport Research Centre Discussion Papers,*, (01).

Chu, S. (2015). Car restraint policies and mileage in Singapore. *Transportation Research Part A: Policy and Practice*, 77:404–412.

Cutter, W. B. and Franco, S. F. (2012). Do parking requirements significantly increase the area dedicated to parking? A test of the effect of parking requirements values in Los Angeles County. *Transportation Research Part A: Policy and Practice*, 46(6):901–925.

Daganzo, C. F. (1994). The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B*, 28(4):269–287.

Daganzo, C. F. (2007). Urban gridlock: Macroscopic modeling and mitigation approaches. *Transportation Research Part B: Methodological*, 41(1):49–62.

Daganzo, C. F. and Geroliminis, N. (2008). An analytical approximation for the macroscopic fundamental diagram of urban traffic. *Transportation Research Part B: Methodological*, 42(9):771–781.

Daganzo, C. F. and Lehe, L. J. (2015). Distance-dependent congestion pricing for downtown zones. *Transportation Research Part B: Methodological*, 75:89–99.

Daganzo, C. F. and Lehe, L. J. (2016). Traffic flow on signalized streets. *Transportation Research: Part B: Methodological*, 90(C):56–69.

Danielis, R., Rotaris, L., Marcucci, E., and Massiani, J. (2011). An economic, environmental and transport evaluation of the Ecopass scheme in Milan: three years later.

Dawson, J. and Catling, I. (1986). Electronic road pricing in Hong Kong. *Transportation Research Part A: General*, 20(2):129–134.

Eliasson, J. (2008). Lessons from the Stockholm congestion charging trial. *Transport Policy*, 15(6):395–404.

Eliasson, J., Börjesson, M., van Amelsfort, D., Brundell-Freij, K., and Engelson, L. (2013). Accuracy of congestion pricing forecasts. *Transportation Research Part A: Policy and Practice*, 52:34–46.

Feitelson, E. and Rotem, O. (2004). The case for taxing surface parking. *Transportation Research Part D: Transport and Environment*, 9(4):319–333.

Ferguson, E. (2004). Zoning for Parking as Policy Process: A Historical Review. *Transport Reviews*, 24(May):177–194.

Franklin, J. P., Eliasson, J., and Karlstrom, A. (2009). Traveller Responses to the Stockholm Congestion Pricing Trial: Who Changed, Where Did They Go, and What Did It Cost Them? In Saleh, W. and Sammer, G., editors, *Demand Management and Road User Pricing: Success, Failure and Feasibility*, pages 215–238. Ashgate Publications., Surrey, England.

Geroliminis, N. and 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. and Sun, J. (2011). Properties of a well-defined macroscopic fundamental diagram for urban traffic. *Transportation Research Part B: Methodological*, 45(3):605–617.

Gibson, M. and Carnovale, M. (2015). The effects of road pricing on driver behavior and air pollution. *Journal of Urban Economics*.

Godfrey, J. W. (1969). The Mechanism of a Road Network. *Traffic Engineering and Control*, 11(7):323–327.

Gómez-Ibáñez, J. A. and Small, K. A. (1994). *Road pricing for congestion management : a survey of international practice*. National Academy Press, Washington D.C.

Gonzales, E. J. (2015). Coordinated pricing for cars and transit in cities with hypercongestion. *Economics of Transportation*, 4(1-2):64–81.

Granovetter, M. (1978). Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6):1420.

Greenshields, B. (1935). A study of traffic capacity. *Highway Research Board Proceedings*, pages 448–477.

Harstad, R. (2005). William S. Vickrey.

Hau, T. D. (1990). Electronic Road Pricing: Developments in Hong Kong 1983-1989. *Journal of Transport Economics and Policy*, 24(2):203–214.

Hau, T. D. (1992). Congestion Charging Mechanisms for Roads: An Evaluation of Current Practice. Technical Report 9, World Bank.

Herman, R. and Ardekani, S. (1984). Characterizing Traffic Conditions in Urban Areas. *Transportation Science*, 18(2):101.

Herman, R. and Prigogine, I. (1979). A two-fluid approach to town traffic. *Science (New York, N.Y.)*, 204(4389):148–151.

Inci, E. (2014). A review of the economics of parking. *Economics of Transportation*.

Jacobs, J. (1961). *The Death and Life of Great American Cities*. Random House, New York.

Jakle, J. A. and Sculle, K. A. (2004). *Lots of Parking: Land Use in a Car Culture*. University of Virginia Press, Charlottesville.

Karlström, A. and Franklin, J. P. (2009). Behavioral adjustments and equity effects of congestion pricing: Analysis of morning commutes during the Stockholm Trial. *Transportation Research Part A: Policy and Practice*, 43(3):283–296.

Krugman, P. (1990). Increasing returns and economic geography. *Jounral of Political Economy*, 99(3):483–499.

Krugman, P. R. (2011). The new economic geography, now middle-aged. *Regional Studies*, 45(1):1–7.

Land Transport Authority (2016). Electronic Road Pricing (ERP).

Leape, J. (2006). The London Congestion Charge. *Journal of Economic Perspectives*, 20(4):157–176.

Lehe, L. (2017). Feedback and the use of land for parking. *Journal of Transport and Land Use*.

Li, F. and Guo, Z. (2014). Do parking standards matter? Evaluating the London parking reform with a matched-pair approach. *Transportation Research Part A: Policy and Practice*, 67:352–365.

Li, M. Z. F. (1999). Estimating congestion toll by using traffic count data: Singapore's area licensing scheme. *Transportation Research Part E: Logistics and Transportation Review*, 35:1–10.

Lighthill, M. and Whitham, G. (1955). On kinematic waves. *Proceedings of the Royal Society of London*, 229(1178):281–316.

Mahmassani, H. S., Williams, J. C., and Herman, R. (1984). Investigation of Network-Level Traffic Relationships: Some Simulation Results. *Transportation Research Record*, 971:121–140.

Manville, M. (2013). Parking Requirements and Housing Development. *Journal of the American Planning Association*, 79(1):49–66.

Manville, M., Beata, A., and Shoup, D. (2013). Turning Housing Into Driving: Parking Requirements and Density in Los Angeles and New York. *Housing Policy Debate*, 23(2):350–375.

Manville, M. and Shoup, D. (2005). Parking, People, and Cities. *Journal of Urban Planning and Development*, 131(4):233–245.

Mattioli, G., Boffi, M., Colleoni, M., and Sociale, R. (2012). Milan's pollution charge: sustainable transport and the politics of evidence. In *Human Dimensions of Global Environmental Change*, number October, pages 1–16.

May, A. D. (1975). Supplementary Licensing: An Evaluation. *Traffic Engineering and Control*, 16:162–167.

May, A. D. and Gardner, K. E. (1989). Transport policy for London in 2001 The case for an integrated approach. *Transportation*, 16(3):257–277.

McCahill, C. and Garrick, N. (2014). Parking Supply and Urban Impacts. In Ison, S. and Mulley, C., editors, *Parking: Issues and Policies*, volume 5, pages 33–55. Emerald.

McCarthy, P. and Tay, R. (1993). Economic Efficiency vs Traffic Restraint: a Note on SIngapore's Area License Scheme. *Journal of Urban Economics*, 34(1):96–100.

McDonnell, S., Madar, J., and Been, V. (2011). Minimum parking requirements and housing affordability in New York City. *Housing Policy Debate*, 21(November 2014):45–68.

Menon, A. G. and Chin, K.-K. (2004). Erp in Singapore-What'S Been Learnt From Five Years of Operation?

Menon, a. P. G. (2000). ERP in Singapore - a perspective one year on.

Ministry of Transport (1964). *Road Pricing: The Economic and Technical Possibilities*. Her Majesty's Stationary Office, London.

Ministry of Transport (1967). *Better Towns with Less Traffic*. HMSO, London.

Mukhija, V. and Shoup, D. (2006). Quantity versus Quality in Off-Street Parking Requirements. *Journal of the American Planning Association*, 72(3):296–307.

Muncipality of Milan (2015). Area C.

MVA Consultancy (1995). London Congestion Charging Research Programme: Final Report. Technical report, Her Majesty's Stationary Office, London.

Newell, G. F. (2002). A simplified car-following theory: A lower order model. *Transportation Research Part B: Methodological*, 36(3):195–205.

Olszewski, P. and Xie, L. (2005). Modelling the effects of road pricing on traffic in Singapore. *Transportation Research Part A: Policy and Practice*, 39(7-9):755–772.

Percoco, M. (2014). The effect of road pricing on traffic composition: Evidence from a natural experiment in Milan, Italy. *Transport Policy*, 31:55–60.

Phang, S.-Y. and Toh, R. S. (1997). From manual to electronic road congestion pricing: The Singapore experience and experiment. *Transportation Research Part E: Logistics and Transportation Review*, 33(2):97–106.

Phang, S.-Y. and Toh, R. S. (2004). Road Congestion Pricing in Singapore : 1975 to 2003. *Transportation Journal*, 43(2):16–25.

Pigou, A. (1920). *The Economics of Welfare*. Macmillan and Co., London.

Quercia, R. G. and Galster, G. C. (2000). Threshold Effects and Neighborhood Change. *Journal of Planning Education and Research*, 20:146–162.

Ramjerdi, F., Minken, H., and Østmoe, K. (2004). 10. Norwegian Urban Tolls. *Research in Transportation Economics*, 9(04):237–249.

Richards, M., Gilliam, C., and Larkinson, J. (1996). The London Congestion Charging Research Programme - 1: the programme in overview. *Traffic Engineering & Control*, 37(2).

Richards, P. I. (1956). Shock Waves on the Highway. *Operations Research*, 4(1):42–51.

ROCOL (2000). *Road Charging Options for London: A Technical Assessment*. Her Majesty's Stationary Office, London.

Rotaris, L., Danielis, R., Marcucci, E., and Massiani, J. (2010). The urban road pricing scheme to curb pollution in Milan, Italy: Description, impacts and preliminary cost benefit analysis assessment. *Transportation Research Part A: Policy and Practice*, 44(5):359–375.

Santos, G., Li, W. W., and Koh, W. T. H. (2004). Transport Policies in Singapore. *Research in Transportation Economics*, 9(February):209–235.

Schelling, T. C. (1971). Dynamic models of segregation. *The Journal of Mathematical Sociology*, 1(2):143–186.

Shoup, D. (2005). *The High Cost of Free Parking*. Planners Press, American Planning Association, Chicago.

Small, K. A. (2004). 6. Road Pricing and Public Transport. *Research in Transportation Economics*, 9(04):133–158.

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

Small, K. A. and Gomez-Ibanez, J. A. (1998). Road Pricing for Congestion Management: The Transition from Theory to Policy. (391).

Small, K. A. and Rosen, H. S. (1981). Applied Welfare Economics with Discrete Choice Models. *Econometrica*, 49:105–130.

Smeed, R. (1966). Road Capacity of City Centres. *Traffic Engineering and Control*, 8(7):455–458.

TfL (2003). Impacts Monitoring. First annual report. Technical report.

TfL (2007a). Central London Congestion Charging Scheme: ex-post evaluation of the quantified impacts of the original scheme. Technical report.

TfL (2007b). Impacts monitoring. Fifth annual report. Technical Report July.

TfL (2008). Non-statutory consultation on the future of the Western Extension of the Congestion Charging Zone: Report to the Mayor. Technical report, Transport for London, London.

Thomson, J. M. (1967). Speeds and Flows of Traffic in Central London: 1. Sunday Traffic Survey. *Traffic Engineering and Control*, Vol. 8(No. 11):672–676.

Transportstyrelsen (2013). Statistics for congestion in Gothenburg 2013.

Transportstyrelsen (2015a). Congestion taxes in Stockholm and Gothenburg.

Transportstyrelsen (2015b). Frequently asked questions about congestion tax.

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.

Vickrey, W. (1959). Statement on the Pricing of Urban Street Use. *Hearings: U.S. Congress, Joint Committee on Metropolitan Washington, D.C. Problems,*, pages 466–477.

Vickrey, W. S. (1969). Congestion theory and transport investment. *American Economic Review*, 59(2):10.

Walters, A. A. (1961). The Theory and Measurement of Private and Social Cost of Highway Congestion. *Econometrica*, 29(4):676–699.

Wardrop, J. (1968). Journey speed and flow in central urban areas. *Traffic Engineering and Control*, 9(11):528–532.

Wardrop, J. G. (1952). Some theoretical aspects of road traffic research. *ICE Proceedings: Engineering Divisions*, 1(3):325–362.

Watson, P. L. and Holland, E. P. (1978). Relieving Traffic Congestion: the Singapore Area License Scheme.

Weinberger, R., Kaehny, J., and Rufo, M. (2010). U . S . Parking Policies : an Overview of Management Strategies. Technical report, Institute for Transportation and Development Policy.

Williams, J. C., Mahmassani, H. S., and Herman, R. (1987). Urban Traffic Network Flow Models. *Transportation Research Record*, (1112):p. 78–88.

Willson, R. W. (1995). Suburban Parking Requirements: A Tacit Policy for Automobile Use and Sprawl. *Journal of the American Planning Association*, 61(1):29–42.

Wilson, P. W. (1988). Welfare effects of congestion pricing in Singapore. *Transportation*, 15(3):191–210.

Zahavi, Y. (1972). Traffic performance evaluation of road networks by the α-relationship. *Traffic Engineering and Control*, 14(5-6).

Appendix A

Agent-based model

An agent-based, discrete-time simulation has been constructed along the lines of the one used in Daganzo and Lehe (2015). At a high level, the simulation works as follows: it iterates over a large number of virtual "Days," each with two stages: a "Rush" and a "Reflection." In the Rush, agent's mode choices are used to simulate a day of traffic. In the Reflection, agents use information from that Day's Rush to decide what to do on the next Day. Agents have two properties: a permanent Trip Length property and a Distance Traveled property.

Rush

During the Rush, time is divided into very small intervals of constant length equal to one "minute." There is an evolving list, called "Traffic," that consists of all driving travelers who have arrived but not exited. At the start of each minute, three things happen: (i) the drivers who finished their trips in the prior interval are plucked from Traffic; (ii) new drivers are added to Traffic; and (iii) a pace, p, is calculated based on the density (i.e., the length of Traffic). Next, to capture travel, the distance p (since pace is given in km per minute and the length of the interval is one minute) is added to a "distance-traveled" property of each driver in Traffic. If a driver's distance-traveled meets or exceeds her trip length, she will be slated for removal during step (i) of the subsequent minute.

Reflection

After the Rush stage comes the Reflection stage. During the Reflection, a randomly-chosen subset of the population decides what mode to take on the next Day. To facilitate these decisions, two functions are built from the data of the Rush to let travelers decide whether to drive or take transit. The function $L(t)$ gives the cumulative distance that would have been traveled at time-of-day t by a driver constantly cruising in the network, or

$$L(t) := \int_0^t 1/p[k(t)] \cdot dt, \tag{A.1}$$

where $k(t)$ is time-path of density from the Day's Rush. The second function is $L(t)$'s inverse, called $T(L)$, which gives the time when the cruising driver would have traveled L km.

Travelers use $L(t)$ and $T(L)$ to predict what the durations of their car trips will be in the next Rush. Let $\bar{w}(t, l)$ be the *predicted* duration of a trip taken by car of length l beginning at time-of-day t. It will be defined

$$\bar{w}(t, l) := T[L(t) + l] - t. \tag{A.2}$$

In other words, the traveler asks herself, "How long would it have taken the cruiser to travel an additional l km from the time when my trip begins?" The answer is her prospective trip duration. After a random cost ε_i, drawn from a logistic distribution with scale parameter α, is added to her prospective net utility from driving, she decides whether to drive or take transit on the next day. We also tested a version of the simulation in which each traveler kept her ε_i draw from day to day, but having a different draw each day aided stability and did not produce very different aggregate outcomes. Using different draws for each also seems permissible under the interpretation that each traveler is actually the discrete representative of a very large population with similar trip lengths.