Optimal Urban Transportation Policy: Evidence from Chicago[†]

Milena Almagro[*], Felipe Barbieri[§], Juan Camilo Castillo[†], Nathaniel Hickok[‡], Tobias Salz[¶]

MAY 22, 2024

Abstract

We characterize and quantify optimal urban transportation policies in the presence of congestion and environmental externalities. We formulate a framework in which a municipal government chooses among transportation equilibria through its choice of public transit policies—prices and frequencies—as well as road pricing. The government faces a budget constraint that introduces monopoly-like distortions and the potential need to cross-subsidize modes. We apply this framework to Chicago, for which we construct a new dataset that comprehensively captures transportation choices. We find that road pricing alone leads to large welfare gains by reducing externalities, but at the expense of travelers, whose surplus falls even if road pricing revenues are fully rebated. The optimal public transit price is near zero, with reduced bus and increased train frequencies. Combining transit policies with road pricing slackens the budget constraint, allowing for higher transit frequencies and lower prices, thereby increasing consumer surplus after rebates.

JEL classification: L91, L5, L13, H23, R41, R48.

Keywords: Urban transit policy, road pricing, spatial equilibrium, Ramsey pricing.

[†] We thank audiences at Booth, Harvard, MIT, Rice, Texas A&M, Yale, CEMFI, EIEF Junior Applied Micro Conference, Tinos IO Conference, LACEA-LAMES, NBER Market Design 2022 Meeting, Philadelphia Fed, UCLA IO/Spatial Conference, Berkeley, Toronto, Georgetown, JHU, Maryland, Wharton RE, UIUC, UNLP, WashU in St Louis, St Louis Fed, IIOC, LSE, Oxford, Warwick, IFS/LSE/UCL IO workshop, Cowles, SED, ERWIT/CURE, Sciences Po, Chicago-Princeton Spatial Conference, Stanford Cities Workshop, Duke, GWU, Northwestern Interactions Conference, PUC Chile, and PSE/ESSEC Workshop. We are grateful to Enrick Arnaud-Joufray, Panle Jia-Barwick, Isis Durrmeyer, Matthew Freedman, Clara Santamaria, and Daniel Sturm for their insighful discussions. We also thank Tomas Dominguez-Iino, Gilles Duranton, Jessie Handbury, Gabriel Kreindler, Jing Li, Jim Poterba, Liam Purkey, Chris Severen, and Chad Syverson for helpful comments. We are also grateful to Gilles Duranton and Prottoy Akbar for helping us obtain travel time data, and to Anson Steward, Jinhua Zhao, and Xiaotong Guo for helping us get access to CTA data. Melissa Carleton, Diego Gentile Passaro, and Shreya Mathur provided outstanding research assistantship. We acknowledge financial support from the NSF (Award No. 1559013, Supplement), through the NBER's transportation initiative, and from the John S. and James L. Knight Foundation through a grant to the University of Pennsylvania Center for Technology, Innovation & Competition.

[*] University of Chicago, Booth and NBER, email: milena.almagro@chicagobooth.edu

[§] University of Pennsylvania Economics, email: kupf@sas.upenn.edu

[†] University of Pennsylvania Economics and NBER, email: jccast@upenn.edu

[‡] MIT Economics, email: nhickok@mit.edu

[¶] MIT Economics and NBER, email: tsalz@mit.edu

1 Introduction

Since the 1950s, urban transportation in the U.S. has been characterized by the overwhelming use of cars. Meanwhile, despite generous subsidies, public transit accounts for only 3.4% of the 850 million daily urban trips in the country. This heavy reliance on cars poses significant challenges for cities: the costs of road congestion in the U.S. are estimated at $87 billion per year, and car usage causes major environmental impacts through emissions of carbon and other pollutants.[1]

Cities' efforts to mitigate their effect on the environment and reduce inequality have led to a renewed discussion about the right mix of urban transportation policies.[2] Some argue that public transit should be cheaper, and, indeed, several municipalities have recently introduced free public transit;[3] others suggest that cities should instead provide more frequent, higher-quality public transit.[4] Despite the potential benefits of these two proposals, it may not be feasible to pursue both because of stressed municipal budgets. In contrast, some cities have recently introduced charges to the private use of roads. For example, London enacted a £15 cordon tax during the daytime and New York recently approved a cordon tax below 60th Street in Manhattan.[5] A major argument used in favor of these taxes is the possibility of using the resulting revenue to subsidize public transit.[6] Given that we observe such varied approaches, what is the right combination of urban transportation policies? Should cities aim to increase the use of public transit, discourage the private use of roads, or some combination of the two?

In this paper, we characterize the optimal mix of urban transportation policies and measure their welfare and distributional effects. We argue that to do this, one must consider important interactions across modes. In addition to mode substitution on the demand side and technological interactions through road congestion, we show that budget constraints introduce important fiscal interactions

[1] See World Economic Forum— US Traffic Congestion Cost in 2018.
[2] See Brookings — U.S. Transportation policy and HKS — Free Public Transit.
[3] See NYT — "Should Public Transit Be Free? More Cities Say, Why Not?"
[4] See The Conversation — Low-cost, high-quality public transportation.
[5] For a comprehensive list of congestion pricing policies see DOT–Congestion Pricing.
[6] See Congestion Pricing's Billions to Pay for Nuts and Bolts of Subway System.

across modes. Due to prevailing difficulties to build new transit infrastructure in the US (Brooks and Liscow, 2023), we focus on some key alternative interventions: road pricing and changing the fares and service frequency of public transit.

We formulate a framework in which a municipal government maximizes welfare, accounting for the cost of congestion and environmental externalities. On the demand side, travelers choose between modes of transportation based on their prices and travel times. On the supply side, we model a transportation technology that determines travel times, taking into account congestion and the frequency of public transit. The government, which can be thought of as a multi-product seller, chooses the prices and qualities (in terms of frequencies) of modes, subject to a budget constraint that accounts for operational costs and revenues from fares and road pricing. Given these government choices, travelers in the market adjust and reach an equilibrium. We focus on adjustments keeping residents' and firms' locations fixed. Thus, our findings do not reflect additional welfare effects from relocation, which previous research suggests may be moderate.[7]

We find that an unconstrained social planner would set price minus marginal cost equal to the marginal externality (as in Pigou, 1932) plus a diversion term that accounts for mispricing of modes not under the planner's control. This is a second-best solution that arises in the multi-product context when the planner has fewer instruments than there are products. However, a budget constrained planner must raise revenue, which introduces two monopoly-like distortions (Ramsey, 1927). First, the planner charges markups that downwards-distort quantities. Second, quality (public transit frequency) is distorted towards the marginal consumer, as in Spence (1975). Cross-subsidization can completely eliminate these distortions. This observation emphasizes the importance of coordinated policies across modes and provides an efficiency rationale for the London and New York plans to use road pricing to cross-subsidize public transit.

Next, we move to an empirical application of this framework in Chicago, an ideal setting for our purposes. Both public and private transportation play an im-

[7] For road pricing, Herzog (2023) finds that endogenizing sorting and traffic congestion attenuates welfare effects by around 20%, whereas Barwick et al. (2021) find that it increases them by 18%.

portant role in this city. It has large economic disparities, which makes it important to measure the effects of transportation policies across different income levels. Furthermore, Chicago provides particularly good data availability. We combine several data sources to construct a high-resolution dataset of travel flows, travel times, and prices for all relevant modes. We have access to the near-universe of public transit trips through records from the Chicago Transit Authority (CTA) and the universe of ride-hailing and taxi trips, which are made public by the City of Chicago. One challenge we face is that there are no official records of car trips. To overcome this problem, we construct the total number of trips from individual cellphone location records and then recover the number of car trips by subtracting public transit, ride-hailing, and taxi trips from the total.

We then turn to estimating our demand model, which allows for heterogeneous substitution patterns across locations, income, and car ownership. The richness of our data allows us to define granular transportation markets—people traveling from one community area (CA) to another during a particular hour of the week—and still conduct our analysis with aggregate market shares (Berry et al., 1995). This approach has the advantage that we can use standard inversion techniques to address endogeneity concerns. While the cost of operating a car and public transit prices are invariant to demand shocks, both the travel times of road-based modes and the prices of ride-hailing are endogenous.

We instrument road travel times using free-flow travel times (i.e., without traffic congestion), arguing that these capture differences in infrastructure that are independent of within-day demand shocks. To address the endogeneity of ride-hailing prices, we exploit price variation due to a surcharge on ride-hailing trips that start or end downtown between 6 am and 10 pm.[8] Our estimates reveal substantial heterogeneity in the value of time across travelers, ranging from $4 to $28 per hour for travelers in the bottom and the top income quintile, respectively.[9]

We then estimate the road traffic congestion technology at a high resolution.

[8] Leccese (2022) also studies this policy variation.

[9] For comparison, the average hourly wage in Chicago's metropolitan area is $30. See Wage statistics from Bureau of Labor Statistics for the Chicago region.

We exploit hour-of-the-day variation in travel speeds and in the number of vehicles traveling between adjacent CAs, following Akbar and Duranton (2017) and Kreindler (2023). We find elasticities of car travel times with respect to traffic flows between 0.10 to 0.17, comparable to existing estimates in the literature (Akbar and Duranton, 2017; Couture et al., 2018). We model wait times for public transit as a function of their frequency and potential delays.

To understand the effect of different transit policies and their interactions, we simulate three main counterfactuals in which the government adjusts overall frequencies and prices. First, we separately explore scenarios in which the government only adjusts public transit prices and frequencies or only implements road pricing. To explore the interactions between these policies, we then compute a counterfactual where the government controls both.

A budget-constrained planner who only controls public transit would reduce bus prices and train prices by 40% and 23%, respectively. To stay on budget, bus frequencies are reduced by 14% and train frequencies by 3%. Welfare increases by $0.4 million per week, primarily from reduced externalities. Consumer surplus remains virtually unchanged, but there are progressive distributional effects: low-income travelers benefit from lower prices despite higher waiting times, while the highest-income travelers are worse off due to their higher value of time. These results are strongly influenced by the budget constraint. An unconstrained planner would reduce prices further, keep bus frequencies at the status quo, and increase train frequencies by 9.2%. In this case, travelers at all income levels would benefit, with weekly consumer welfare gains of $12.6 million ($4.6 per resident).

The optimal road tax, when used as the sole instrument by the planner, is 35 cents per kilometer, or roughly $13.4 per day for the average commuter. This leads to overall welfare gains of $4.57 million per week, nearly three times the gains from optimal transit policies alone. The majority of these gains come from a $3.6 million weekly reduction in externalities. However, without rebates, consumer surplus would decrease by $29.1 million per week, with middle-income consumers experiencing the greatest losses due to their dependence on cars. Even if the government

were to fully rebate resulting revenues, consumer surplus would decrease by $0.8 million per week.

When considering road pricing and public transit policies together, the government's budget constraint is no longer binding due to the revenue collected from road pricing. As a result, transit policies are similar to those implemented in isolation without a budget constraint. Public transit becomes nearly free, with optimal fares of $0.16 for buses and $0.26 for trains. Compared to a scenario where the budget-constrained planner only sets transit policies, train and bus frequencies are 12.6 and 11.5 percentage points higher, respectively. This combined policy increases overall welfare by $5.27 million per week. Similar to the effects with road pricing alone, consumer surplus experiences a substantial decrease, amounting to $18.54 million per week, or $6.76 per resident. If the government rebates its full surplus, on the other hand, combining road pricing and public transit increases consumer surplus by up to $0.5 million per week. These consumer surplus gains would be progressive under a flat rebate.

These findings highlight the importance of jointly considering public transit policies and road pricing: road pricing can help achieve large reductions in externalities, but it benefits consumers only if resulting revenues are used to cross-subsidize public transit and to rebate travelers. This way it can also help to undo inefficiencies that arise due to the government's need to balance the budge

We also investigate policies in which price and frequency adjustments are dependent on location and time. We find that there are almost no additional gains from adjusting prices in a more granular way. However, we find large gains from granular route frequency adjustments that redirect available capacity towards busy areas and times, implying misallocation in the status quo.

Related Literature Our work relates to several strands in the literature on transportation economics and industrial organization.

A growing literature analyzes transportation markets on the basis of spatial equilibrium models. These studies are closely linked to theoretical work by Arnott

(1996), which shows that taxis should be subsidized because of increasing returns to scale, and Lagos (2003), who formulates a spatial matching model of the New York taxi market. More recent empirical work has also studied the New York taxi market (Frechette et al., 2019; Buchholz, 2021), as well as the dry bulk shipping industry (Brancaccio et al., 2020) and ride-hailing platforms (Castillo, 2023; Rosaia, 2023; Gaineddenova, 2022; Buchholz et al., 2024). Kreindler (2023) studies the welfare effects of congestion taxes. Like Brancaccio et al. (2023), who derive optimal policies for transportation markets with matching frictions, we derive them for urban transportation markets with a budget-constrained social planner.

Within this strand of literature, Durrmeyer and Martínez (2023), Kreindler et al. (2023), and Barwick et al. (2021) are most closely related to our work. Durrmeyer and Martínez analyze an equilibrium model of mode substitution and assess the welfare impacts of private car restrictions and road pricing. Our study differs in two main ways. First, our research question focuses on the interaction between road pricing and public transit via mode substitution, congestion, and the planner's budget constraint. Second, by formulating the government's problem as that of a "monopoly seller," we are able to focus on the importance of distortions that are created by budget considerations and the resulting welfare gains created by cross-subsidization. Kreindler et al. (2023) study optimal transit policies but focus on the optimal network configuration for buses. While our policy simulations are less granular in terms of network planning, we incorporate the trade-off that the social planner faces when setting policies for both public transit and private modes of transportation. Barwick et al. (2021) jointly analyze transportation mode and residential location choices, and they also explore combinations of different transportation policies. They find that the combination of congestion pricing and subway expansion delivers the greatest congestion relief. We depart from this paper in two main ways. We characterize and decompose the optimal policy, and we highlight the interactions created by budget considerations.[10]

[10] Several papers investigate alternative margins of adjustment in response to transportation policy (Tsivanidis, 2023; Fajgelbaum and Schaal, 2020; Severen, 2023; Herzog, 2023; Brinkman and Lin, 2022; Allen and Arkolakis, 2022; Bordeu, 2023). We depart from their work by allowing for rich

We also build on a classic theoretical literature in transportation economics that develops models that capture the interaction between schedule constraints and congestion (Small, 1982; Arnott et al., 1990, 1993; Small et al., 2005). We enrich these models by combining a congestion model with the demand approach used in industrial organization (Berry, 1994; Berry et al., 1995), which allows us to model rider heterogeneity and account for the endogeneity of travel times and prices.

Finally, our work relates to the broader literature in transportation economics. Some works look at traffic congestion(Akbar and Duranton, 2017; Akbar et al., 2023; Couture et al., 2018; Kreindler, 2023) and different forms of road pricing (Hall, 2018; Cook and Li, 2023; Yang et al., 2020). These papers abstract away from mode substitution and the interaction between public and private transportation. Parry and Small (2009) is closely related to our work in that it also derives theoretical expressions for the optimal prices of public transit, which they then calibrate to aggregate data from three cities. We extend their results to account for the joint effect of prices and quality improvements and for the distortions introduced by budget considerations. Furthermore, we model the resulting equilibrium adjustments by taking into account the linkages across many markets.

2 Background and Data

2.1 Background

Chicago is the third largest city in the U.S. and its public transit system, which is operated by the Chicago Transit Authority (henceforth CTA), is one of the largest in the nation. It includes a bus network of 127 routes with almost two thousand buses, and a train rapid transit system—the "Chicago L"—that has eight routes and 145 stations. Full fares for bus and trains are $2.25 and $2.50, respectively.[11] The CTA has a history of budget shortfalls, making it important to account for

demand substitution patterns across modes and heterogeneity, which is crucial for understanding the distributional effects of transit policies.

[11] Reduced fares exist for students and seniors. There are also daily, 3-day, weekly, and monthly passes. See CTA Fares for additional details.

budget considerations.[12] Passengers can also travel by private for-hire-vehicles in the form of taxis and ride hailing. Taxis have a regulated fare of $\$2.25$ per mile or $\$0.2$ per 36 seconds, plus a $3.25 base fare.[13] Ride-hailing companies adjust prices dynamically according to market conditions.

2.2 Data description

We define a market as an origin-destination-time tuple. We use Chicago's Community Areas (CAs) as our spatial units. There are 77 CAs in Chicago, with an average size of three square miles and an average population of 36,000 people. We define a unit of time h as an hour of the day, distinguishing between weekdays and weekends. Thus, we have 48 time periods. Our main dataset consists of travel flows, prices, and travel times for every mode in every market during January 2020.

To construct this dataset, we rely on a variety of raw data sources. First, we use public transit microdata from the CTA. For both buses and trains (i.e., the Chicago L), we observe records of individual trips paid by fare card.[14] We observe the train station or bus stop of origin, the time when the passenger tapped in, and an inferred drop-off station or stop (Zhao et al., 2007).

The second data source, published by the City of Chicago, contains the universe of de-identified taxi and ride-hailing trips.[15] It includes prices, pickup and dropoff locations as well as trip length and duration.

The third source is Veraset mobile-phone location data, which records a device ID and a sequence of GPS coordinates and timestamps for approximately 40% of active cellphone devices in the US.[16] We infer all motorized trips from the sequence of GPS coordinates for each individual device. Appendix A describes this

[12] See, for instance, CTA avoids service cuts, fare hikes under proposed $1.8 billion budget.
[13] See Chicago Taxi Fare Regulation.
[14] We do not observe the 12% of trips paid by cash, so we scale up trip counts to align with aggregate daily ridership (see Appendix S1). Additionally, Metra (commuter rail) trips—which account for less than 1% of trips (see My Daily Travel and Annual/Monthly Ridership)—are not included in our data because Metra is not managed by the CTA.
[15] Source: Chicago Data Portal, Transportation Network Providers - Trips
[16] Source: Veraset: Location Data Provider

process in detail. The frequency with which records are generated depends on the applications installed by the user, so we restrict our analysis to devices with frequent location information. As we explain at the end of this section, this restriction results in a sample of trips and travelers that is representative across many dimensions. We thus multiply the cellphone trips by a common inflation factor to arrive at the total number of trips implied by the 2019 Household Travel Survey from the Chicago Metropolitan Agency for Planning (CMAP).

We combine these data sources to construct the total number of trips across all modes: private car, taxi, ride-hailing, buses, and trains. While the CTA data allow us to observe the number of trips for buses, trains, taxis, and ride-hailing, we do not have official records of car trips. Given that the cellphone data covers all motorized trips, we recover car trips for any given market by subtracting public transit, taxi, and ride-hailing trips from the cellphone trips. Finally, we query Google Maps data to obtain travel times and routes by market for all modes, including those not chosen by travelers.

Since we only see motorized trips in our data, we do not observe travelers who decide not to travel, walk, or bike, which our model treats as the outside optio We thus set the overall size of each individual market to be twice the number of trips we observe. We arrive at this factor by comparing the number of morning commuters to the potential market size, which we take to be the number of residents (see Appendix A). Since our demand model includes a separate nest for the outside option, the choice of this factor does not have a large influence on the substitution patterns implied by our model.

Our comprehensive dataset has two advantages over survey data. First, survey data lack representativeness at granular spatial or temporal resolutions and their coverage diminishes at higher resolutions, leading to sparse data.[17] Our granular data allow us to estimate the relationship between vehicle flow and traffic speed throughout the city. Second, they provide econometric advantages, such as the ability to invert market shares (Berry, 1994; Berry et al., 1995) at a granular level,

[17] In the CMAP 2019 Household Travel Survey, over 60% of origin-destination CA pairs have zero trips, a problem that is exacerbated when the data are broken down by time period.

enabling the construction of moment conditions based on instruments to address the endogeneity of prices and travel times.

We add demographic information to our main dataset using the 2016-2020 American Community Survey (ACS). We match devices to census tracts by inferring a device user's home tract based on the modal GPS tract during night-time hours. We identify residents as those devices that spend at least three nights in their modal night location in a month, which we call *residents*, and denote the rest as *visitors*. Residents account for 93.3% of all cellphone trips. For residents, we impute their income and car ownership probability as the median income and car ownership rates of their home census tract. We are thus able to construct the distribution of travelers' income and car ownership for every market.

We validate our data in two ways. First, we compare the distribution of travel times and distances to those from the CMAP survey and see a large overlap between those distributions. Second, we show that the resulting data is representative across the distribution of inferred incomes. See Appendix A for more details.

2.3 Summary statistics and descriptive results

We present descriptive evidence in four parts. First, we explore the characteristics and usage of different motorized transportation modes. Next, we analyze rider characteristics and how they correlate with mode choices. We then document evidence of the low utilization rates of buses. Finally, we present evidence on the form of traffic congestion in the raw data.

Although Chicago has one of the most extensive public transit systems in the US, about 69% of trips are taken by car. Public transit accounts for 23% of trips, with buses taking slightly more than half of this share. Ride-hailing accounts for 6% of trips. Lastly, taxis account for only 0.9% of trips—and, hence, we exclude them from our analysis. The top panel of Figure 1 shows how trips of different modes are distributed across space. Bus and car trips are spread throughout the city. Ride hailing mostly accounts for short trips downtown or north of downtown and along the coast of Lake Michigan, as well as for trips to and from the two major

airports in Chicago: O'Hare to the northwest and Midway to the southwest.

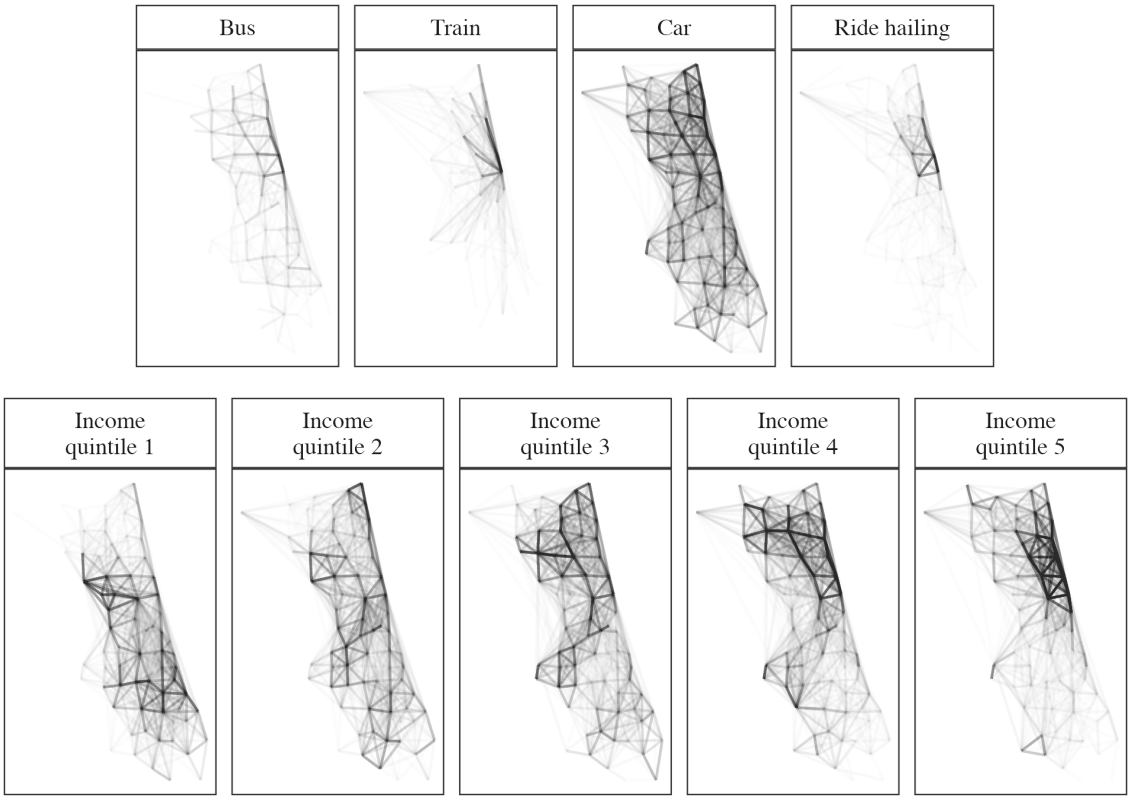


Figure 1: Trips by mode and by income

Notes: These figures show a random sample of 10,000 trips. A line connects the origin and destination of every trip. The panels at the top split trips by mode. The panels at the bottom split them by the income quintile of the traveler.

Chicago's stark income differences are reflected in distinct travel patterns. The bottom panel of Figure 1 shows that low income travelers mostly stay in the south and the west parts of the city. The highest income travelers mostly stay downtown and to the north, along the coast of Lake Michigan. Trips of intermediate income travelers are more evenly spread throughout the city.

Figure 2 shows the differences in speed, travel time, and prices across modes. The left panel shows the distribution of the speed of each mode relative to the speed of buses. Trains, on average, are 10% faster than buses, and cars and ride-hailing, on average, are almost twice as fast as buses. The right panel shows that

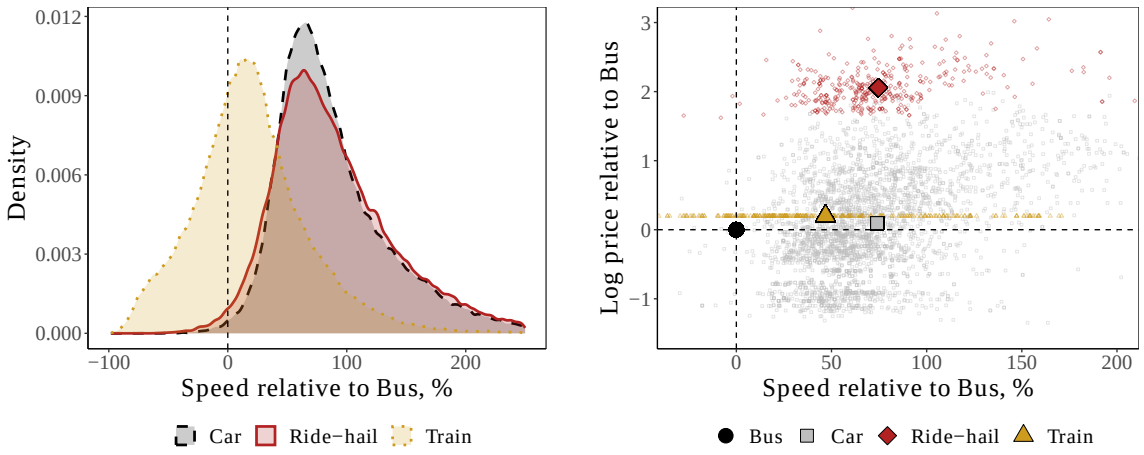


Figure 2: Speed and price differences across modes of transportation

Notes: The left panel shows the distribution of speed by mode of transportation. The right panel presents scatterplots of prices and speed by mode. The prices of public transit and ride-hail are the trip fares. Large dots indicate averages by mode. Observations are at the market level, weighted by the total number of trips in the market.

choosing a mode typically involves a tradeoff between prices and speed: faster modes tend to be more expensive. Cars are the main exception to this tradeoff. They dominate other modes because of their low price, which equals the sum of the marginal cost of fuel, depreciation, and maintenance.

Figure 3 shows patterns in car ownership across income levels. We observe an inverted U-shape: car ownership first increases in income and then declines at the top of the income distribution. Our demand estimation incorporates car ownership, as otherwise our estimates would conflate preferences for non-car modes of transportation with the actual possibility of travelling by car.

Figure 4 shows how mode choices vary by the average income of the origin CA. Lower income travelers are more likely to travel by bus. Perhaps surprisingly, higher income people are more likely to use trains than lower income people: as we can see in Figure 1, train trips are concentrated along L lines, which tend to be located in higher income areas. Consistent with car ownership patterns, car usage follows an inverted-U shape: middle income people are most likely to use cars—and, thus, they are likely to be affected the most by road pricing. Finally, ride-hailing is mainly used by the highest income people.

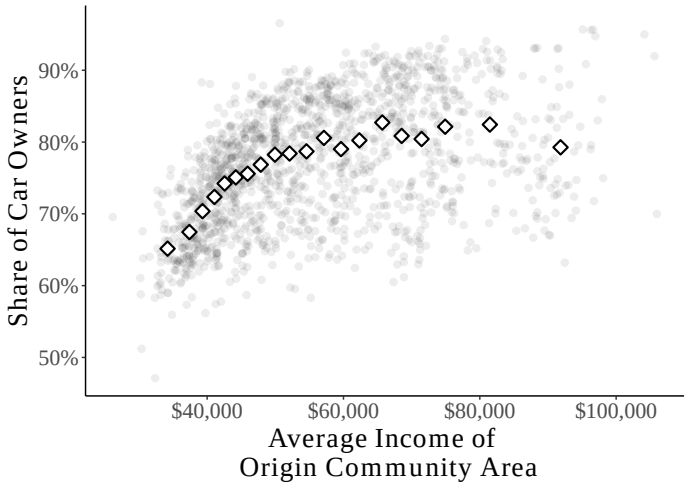


Figure 3: Car ownership by travelers' income

Notes: This figure plots a scatterplot and binscatter of car ownership against the average income of the origin CA.

We now show that, although buses are the cheapest mode, they generally travel almost empty. Figure 5 reveals that, even during the morning and afternoon rush hours, median utilization rates stay below 20%, and less than 10% of buses are at a utilization above 75%. Moreover, buses reach full capacity very rarely; less than 5% of buses are full during the busiest times. The low utilization of buses results in large average costs per passenger, which will be a key ingredient behind some

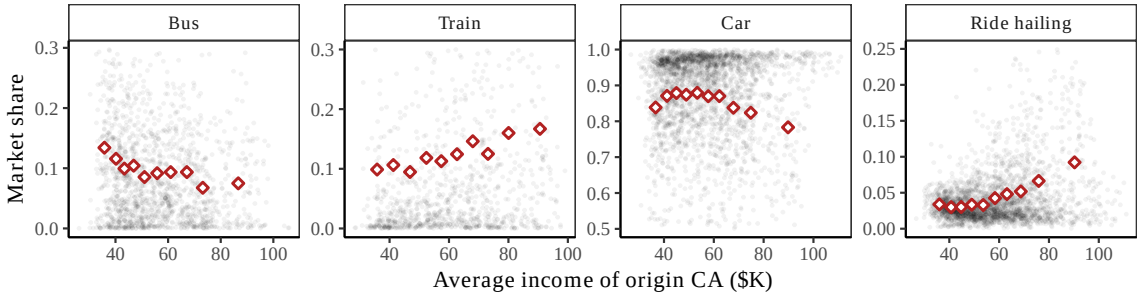


Figure 4: Mode market shares by travelers' income

Notes: Each one of these panels presents a scatterplot and a binscatter of market shares against average income for each mode. Each observation represents trips going from an origin CA to a destination CA. Note that the vertical scale varies by mode.

of our counterfactual results.

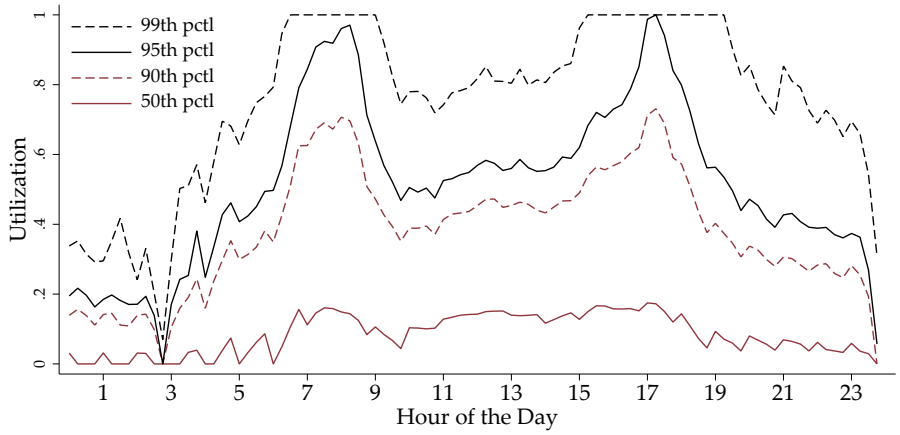


Figure 5: Bus utilization rates

Notes: This figure shows the 50th, 90th, 95th, and 99th percentile bus utilization rate over the course of the day, restricting to weekdays. We measure utilization for each bus every fifteen minutes by taking the number of riders on the bus divided by the capacity of the bus. We conservatively assume each bus has a capacity of 53, which is the smaller of the two bus sizes used by the CTA. If the number of observed riders is greater than the assumed capacity we set the utilization rate to 1.

Lastly, we present raw data patterns that show how traffic congestion impacts travel times. The left panel of Figure 6 shows the relationship between the number of vehicles moving between pairs of adjacent CAs and travel times, after residualizing on CA pair fixed effects. The data exhibit a "hockey-stick" pattern: travel times are constant at lower vehicle counts; however, beyond a certain traffic volume, they begin to rise at a rate close to log-linear. This suggests that additional vehicles reduce travel speeds with an approximately constant elasticity. The right panel shows that this pattern is also clear for buses and when focusing on two arbitrary markets with different levels of infrastructure: market 1 corresponds to two CAs connected by a highway, whereas market 2 lacks a highway connection.

Under the hockey-stick functional form that we see in these figures, travel times in the flat region—which we refer to as *free-flow* times—give us a measure of persistent differences in road speed that is unrelated to demand-induced variation. Figure 7 shows that infrastructure is an important driver of free-flow times. For cars, free-flow times between pairs of CAs are substantially lower when those CAs

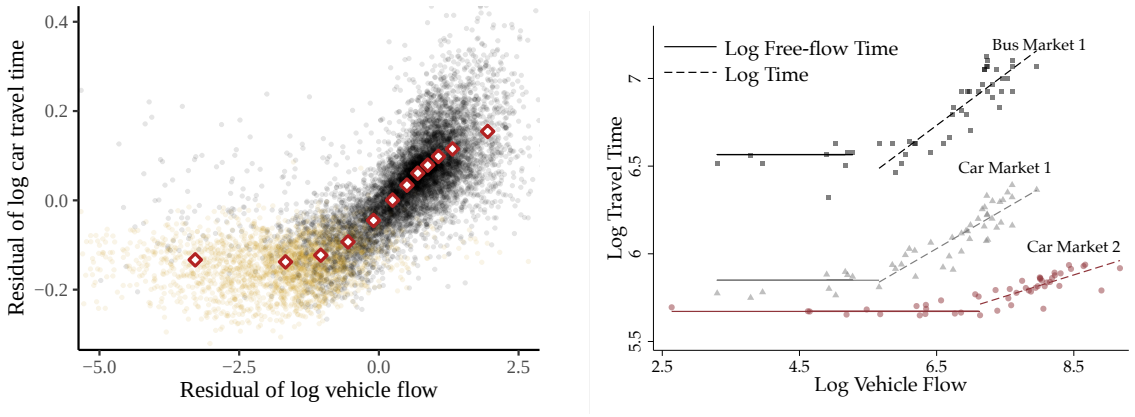


Figure 6: Relationship between flow of vehicles and travel times

Notes: These figures present the relationship between the flow of vehicles and travel times, using data at the level of a pair of adjacent CAs during an hour of the week. The left panel shows log car travel times against the logarithm of vehicles on the road, where both variables are residualized by market fixed effects. Yellow points represent observations between midnight and 5 am, which we use to define free-flow travel times. The right panel shows that the same pattern holds for two arbitrary markets with different levels of infrastructure, as well as for buses.

are connected by highways. For the speed of buses, on the other hand, highway connections make little difference. This variation makes different modes more or less attractive in different markets. We thus use it as an instrument in our demand estimation.

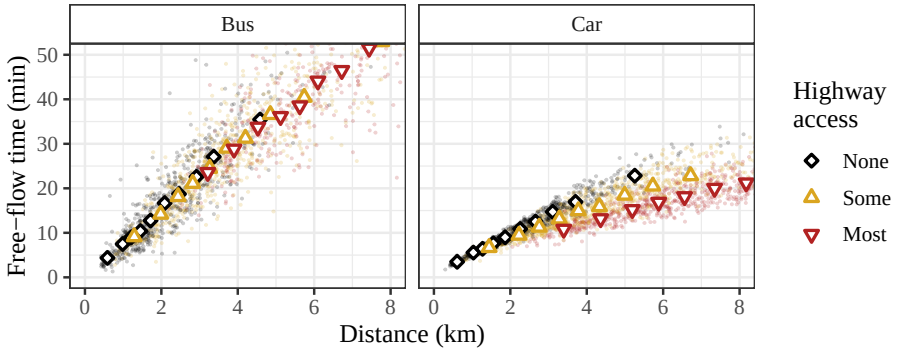


Figure 7: Impact of infrastructure on free-flow times

Notes: This figure shows scatterplots and binscatters of car and bus free-flow times as a function of distance. We focus on weekdays between 10 pm and midnight, when there is little traffic congestion. Observations are at the origin by destination CA. Observations are classified by whether the route suggested by Google Maps does not use highways ("None"), somewhat uses highways ("Some", <50% of the distance), or mostly uses highways ("Most", >50% of the distance).

3 Model

Our model consists of three parts. First, travelers, who have fixed origins and destinations, choose either one of the available modes or not to travel at all. Second, the transportation technology captures the relationship between the number of people who use each mode and travel times. Third, a social planner maximizes welfare, subject to a budget constraint.

Section 3.1 presents a simple version of our model that focuses on only one market. In Section 3.2, this simplified model is used to derive theoretical results about the main forces in the social planner's problem. Section 3.3 presents the empirical version of our model, which accounts for temporal and spatial variation as well as for the spatial linkages across markets.

3.1 Setup and Equilibrium Definition

The first component of our model consists of travelers. They differ in their type $\theta \in \mathbb{R}^n$, with density $f(\cdot)$, which captures their preferences for modes and whether they own a car.

A traveler of type θ decides which transportation mode j to take to her destination. She can choose among the set $\mathcal{J}(\theta)$, which varies depending on whether public transit is easily accessible and whether she owns a car. She can also choose the outside option (walking, biking, or not taking a trip), which we denote by $j = 0$. The traveler gets utility $u_j(t_j, \theta) - p_j$ if she takes transportation mode j, where p_j is the price and t_j is the travel time. This travel time includes the in-vehicle time, the waiting time before the trip starts, and—for public transit—the walking time to the station or stop.[18] We normalize the utility of the outside option to zero. The traveler chooses the mode in her choice set $\mathcal{J}(\theta) \cup \{0\}$ that maximizes utility:

$$j^*(\theta) = \underset{j \in \mathcal{J}(\theta) \cup \{0\}}{\operatorname{argmax}} \; u_j(t_j, \theta) - p_j \tag{1}$$

[18] This model can also incorporate heterogeneity in price sensitivity with utility $u_j(t_j, \theta) - \theta_p \cdot p_j$. Rescaling utility as $1/\theta_p \cdot u_j(t_j, \theta) - p_j$ leads to the same optimal choices.

Given vectors of prices $\mathbf{p}$ and total trip times $\mathbf{t}$ for all modes, demand for mode j is given by

$$q_j = q_j(\mathbf{p}, \mathbf{t}) = \int_{\Theta_j(\mathbf{p},\mathbf{t})} f(\theta)\, d\theta, \tag{2}$$

where $\Theta_j(\mathbf{p}, \mathbf{t})$ is the set of traveler types who choose mode j at $(\mathbf{p}, \mathbf{t})$. We refer to the vector $\mathbf{q}$ as trips.

Gross consumer utility and consumer surplus are given by

$$U(\mathbf{p}, \mathbf{t}) = \sum_j \int_{\Theta_j(\mathbf{p},\mathbf{t})} u(t_j, \theta) f(\theta)\, d\theta \quad \text{and} \quad CS(\mathbf{p}, \mathbf{t}) = \sum_j \int_{\Theta_j(\mathbf{p},\mathbf{t})} (u(t_j, \theta) - p_j) f(\theta)\, d\theta.$$

Travel times are determined by a transportation technology that depends on the number of travelers choosing each mode as well as on the overall capacity of the fleet for each mode. The fleet size for public transit is a policy choice and determines the frequency at which buses and trains run. For ride-hailing, the fleet size is determined by the number of drivers. The transportation technology also captures the fact that the in-vehicle time for road-based modes of transportation depends on the degree of road congestion. Accounting for all these considerations, we can compactly write the vector $\mathbf{t}$ of travel times for all modes as

$$\mathbf{t} = T(\mathbf{q}, \mathbf{k}), \tag{3}$$

where $\mathbf{k}$ is the vector of fleet sizes for all modes.

For each mode j there is a cost $C_j(q_j, k_j)$ to supply q_j rides with fleet size k_j. This cost function includes both labor costs and physical costs, such as fuel and vehicle depreciation. Additionally, society bears an environmental externality $E_j(q_j, k_j)$. We also define total costs and externalities $C(\mathbf{q}, \mathbf{k}) = \sum_j C_j(q_j, k_j)$ and $E(\mathbf{q}, \mathbf{k}) = \sum_j E_j(q_j, k_j)$. With this notation we can now define an equilibrium.

Definition 1 (Transportation equilibrium)**.** *Given prices* $\mathbf{p}$ *and fleet sizes* $\mathbf{k}$*, an equilibrium is a vector of trips* $\mathbf{q}^*(\mathbf{p}, \mathbf{k})$ *and travel times* $\mathbf{t}^*(\mathbf{p}, \mathbf{k})$ *such that* (2) *and* (3) *hold.*

In this model, for any given fleet size and prices, travel times and quantities adjust to bring the market to equilibrium.

3.2 The Social Planner's Problem

The city government's goal is to maximize welfare subject to a budget constraint. The choice variables are the prices and fleet sizes of buses and trains as well as the road price. Modes controlled by the government are denoted $\mathcal{J}_G$. We denote their prices and fleet sizes by where $\mathbf{p}_G$ and $\mathbf{k}_G$.

To define welfare and the government's budget, it will be easier to think of the allocation $(\mathbf{q}, \mathbf{k})$ that arises in equilibrium. The government's revenue is equal to the payments it obtains from travelers minus its costs:

$$\Pi(\mathbf{q}, \mathbf{k}) = \sum_{j \in \mathcal{J}_G} \left[p_j(\mathbf{q}, T(\mathbf{q}, \mathbf{k}))q_j - C_j(q_j, k_j) \right].$$

This revenue cannot fall below $-B$, where B is the transportation budget. Welfare is equal to gross consumer utility minus the cost of transportation provision and externalities:

$$W(\mathbf{q}, \mathbf{k}) = U(\mathbf{q}, T(\mathbf{q}, \mathbf{k})) - C(\mathbf{q}, \mathbf{k}) - E(\mathbf{q}, \mathbf{k}).$$

The government's optimization problem is thus:

$$\max_{\mathbf{p}_G, \mathbf{k}_G} \quad U(\mathbf{q}^*, T(\mathbf{q}^*, \mathbf{k})) - C(\mathbf{q}^*, \mathbf{k}) - E(\mathbf{q}^*, \mathbf{k}) \qquad \text{s.t.} \qquad \Pi(\mathbf{q}^*, \mathbf{k}) \geq -B. \tag{4}$$

To simplify notation, we omit the arguments of $\mathbf{q}^*(\mathbf{p}, \mathbf{k})$.

To state optimality conditions for this problem, we introduce notation for the derivatives of costs, externalities, travel times and utilities with respect to some quantity x. To do this, we use superscripts. For example, C_j^q denotes the derivative of the cost with respect to the number of rides of mode j. Also, let Ω_{lj} represent elements of the inverse Jacobian of $\mathbf{q}(\mathbf{p}, \mathbf{t})$ with respect to p. And, finally, define $D_{lj} = \frac{\partial q_l}{\partial p_j} / \frac{\partial q_j}{\partial p_j}$ as the diversion ratio from j to l of a price increase for mode j.

With this notation in place we can now state the following proposition:

Proposition 1. *Prices under the solution of the social planner's problem* (4) *are given by:*

$$p_j = \overbrace{C_j^q + E_j^q}^{\text{Mg. cost and env. externality}} - \overbrace{\sum_l u_l^T \cdot T_{lj}^q}^{\text{Network effects}} + \overbrace{M_j^q}^{\text{Diversion}} + \frac{\lambda}{1+\lambda} \cdot \Bigg(\underbrace{\sum_{k \in \mathcal{J}_G} q_k \cdot \Omega_{kj}}_{\text{Market power markup}} - E_j^q - \underbrace{\sum_l (\tilde{u}_l^T - u_l^T) \cdot T_{lj}^q}_{\text{Spence distortion}} + \underbrace{\tilde{M}_j^q - M_j^q}_{\text{Diversion distortion}} \Bigg) \tag{5}$$

where λ is the Lagrange multiplier for the budget constraint, $\tilde{u}_j^T$ is a weighted sum of the derivative of gross utility among marginal travelers with respect to mode-j travel time, and M_j^q and $\tilde{M}_j^q$ are defined as:

$$M_j^q \equiv \sum_{k \neq j} D_{kj} \left(C_k^q + E_k^q - \sum_l u_l^T \cdot T_{lk}^q - p_k \right) \tag{6}$$

$$\tilde{M}_j^q \equiv \sum_{l \in \mathcal{J}_G \setminus j} D_{lj} \left(C_l^q + \sum_{k \in \mathcal{J}_G} q_k \cdot \Omega_{kj} - \sum_m \tilde{u}_m^T \cdot T_{ml}^q - p_l \right). \tag{7}$$

Proof. See Appendix C.2 □

In Appendix C.1, we derive a similar expression that decomposes optimal travel times into analogous terms.

To understand Equation 5, first imagine an unconstrained social planner with $\lambda = 0$. Prices are equal to the Pigouvian solution: in addition to the direct costs, they are set to internalize environmental externalities, network effects, and an additional term that we call the *diversion term*. We now describe the terms arising from network effects and mode diversion.

First, network effects are equal to the sum over modes of the product of u_k^T, the derivative of gross utility with respect to mode k travel time, and T_{kj}^q, the change in that time given an additional trip using mode j. If j and k are road-based modes, T_{kj}^q is positive due to traffic congestion, and so these terms lead to a Pigouvian tax. T_{jj}^q is also positive for ride hailing because more trips deplete streets of idle drivers,

increasing pickup times.

Second, the diversion term M_j^q (Equation 6) captures the extent to which change in the price of j induces substitution towards mispriced modes not under the planner's control. For instance, without road taxes, traveling by car may be underpriced. As a second best, the government would want to lower the price of public transit to induce substitution away from cars. M_j^q is a sum over modes of two components: a diversion ratio D_{kj} and a term in brackets that equals deviations of prices from a standard Pigouvian solution (marginal costs plus marginal externalities and network effects). This term is therefore the diversion-ratio weighted sum of these deviations for all other modes, and it is zero whenever all other modes are already priced at the Pigouvian solution.

We now account for budget constraints, which make the social planner behave like a monopolist because of the need to raise revenue. This introduces a market power markup in Equation 5, and the social planner now under-weights environmental externalities. There is also a Spence distortion: while the government internalizes effects on other travelers' utility, it does so imperfectly by accounting for changes in the utility of marginal travelers rather than that of all travelers. Finally, the planner is now concerned with the revenue implications of diverting travelers to other modes. As a result, the diversion term is distorted towards its revenue-motivated equivalent $\tilde{M}_j^q$, which captures whether price changes induce substitution towards modes that are chosen by too few travelers to maximize revenue, rather than social welfare. As $\lambda \to \infty$, the social planner becomes purely revenue maximizing: terms related to environmental externalities cancel out, there is a full markup and a full Spence distortion, and the planner only cares about the revenue-motivated diversion term.

In our counterfactual analysis of Section 5, we come back to these results, empirically teasing out how different sources of externalities contribute to optimal policy. We use a version of this decomposition that applies to a multiple-markets setting, which we derive in Appendix C.3.

3.3 Empirical Model

Here we describe the empirical implementation of our demand model and of the transportation technology. We divide the city into CAs a and time into hours h.

3.3.1 Demand

First, we define a market $m = (a, a', h)$ as the collection of people who make travel decisions from CA a to CA a' at a particular time h.[19] In each market, there is an exogenous number of potential travelers N_m. They decide which mode $j \in \mathcal{J}_m^i \cup \{0\}$ to use, where the outside option $j = 0$ corresponds to walking, biking, or staying put. To make that decision, they solve the following problem:

$$\max_{j \in \mathcal{J}_m^i \cup \{0\}} \quad \xi_{mj} + \alpha_T \cdot T_{mj} + \alpha_p^i \cdot p_{mj} + \epsilon_{mj}^i, \tag{8}$$

where T_{mj} denotes the travel time for mode j (including walk and wait times), p_{mj} is the price for mode j, α_T is the preference parameter over travel times, α_p^i is the person i-specific price coefficient, and ϵ_{mj}^i is an idiosyncratic taste shock. The value of time (VOT) is α_T/α_p^i. Motivated by the disparities in mode choice across the income distribution from Section 2.3, we allow α_p^i to vary across income levels. This captures differences in the marginal utility of money by income, which leads to heterogeneity in the trade-off between time and money.

Given that some modes might be unavailable to some individuals, we allow the choice set to vary across markets and consumers. For instance, some CAs cannot be reached by train and some consumers do not own a car. Cars are in the choice set $\mathcal{J}_m^i$ with a probability equal to the empirical fraction of car owners among consumers of type i in market m.

The taste shock ϵ_{mj}^i is specific to mode j. The joint distribution of the shocks for all modes follows the standard form for a nested logit model with two nests, one consisting solely of the outside option and one consisting of all inside goods $\mathcal{J}_m^i$.

[19] We aggregate across days, so traveling decisions should be thought as the choice for an average hour h rather than choices stemming from short-run shocks, such as special occasions.

This allows for stronger substitution among inside goods, which is mediated by a parameter $\rho \in [0, 1]$. A higher value of ρ indicates stronger substitution among inside goods. Concretely, the taste shock takes the form $\epsilon^i_{mj} = \varsigma^i_{mg(j)} + (1-\rho)\eta^i_{mj}$, where η^i_{mj} is specific to mode j and is distributed Type 1 Extreme Value. The term $\varsigma^i_{mg(j)}$ is common to all goods in group $g(j)$ and follows the unique distribution such that $\varsigma^i_{mg(j)} + (1-\rho)\eta^i_{mj}$ is also distributed Type 1 Extreme Value.

Under our distributional assumptions, the probability that person i chooses mode j in market m is therefore given by:

$$\mathbb{P}^i_{mj} = \frac{\exp\left(\frac{\delta^i_{mj}}{1-\rho}\right)}{\left[\sum_{j \in g} \exp\left(\frac{\delta^i_{mj}}{1-\rho}\right)\right]^{\rho} \cdot \left(\sum_g \left[\sum_{j \in g} \exp\left(\frac{\delta^i_{mj}}{1-\rho}\right)\right]^{(1-\rho)}\right)}. \tag{9}$$

where $\delta^i_{mj} = \xi_{mj} + \alpha_T \cdot T_{mj} + \alpha^i_p \cdot p_{mj}$.

Integrating over α^i_p, mode shares and trips for mode j in market m are:

$$\mathbb{P}_{mj} = \int \mathbb{P}^i_{mj} \, dF_m(\alpha^i_p), \qquad q_{mj} = N_m \cdot \mathbb{P}_{mj}. \tag{10}$$

3.3.2 Transportation Technology

Our transportation technology determines travel times as a function of trips and fleet sizes (i.e., frequencies). We model the total travel time as the sum of three components that vary by mode—walk time, wait time, and in-vehicle time:

$$T_{mj} = \gamma \cdot \left(T^{\text{walk}}_{mj} + T^{\text{wait}}_{mj}\right) + T^{\text{vehicle}}_{mj},$$

where γ is the relative distaste for time spent walking or waiting relative to in-vehicle time.[20]

We model in-vehicle times T^{vehicle}_{mj} as a function of road traffic. To do this, we represent the city by a directed graph, where each node represents a CA and edges connect neighboring CAs. Edge $e = (a, a')$, for instance, connects CAs a and a'. We

[20] We set $\gamma = 2$ following Small (2012). For ride hailing and cars, walk times are zero; for cars, wait times are zero. We take walk times from Google Maps and we assume they are exogenous.

assume that routes are exogenous: travelers take the route suggested by Google Maps. If a traveler uses mode j in market $m = (a, a', h)$, she follows a directed path $P_{mj} = ((a, a_1), (a_1, a_2), \ldots, (a_n, a'))$ over edges that connects a with a'.

During hour h, we define the total vehicle flow on edge e as:

$$F_{eh} = \sum_j w_j \cdot f_{ehj}, \tag{11}$$

where f_{ehj} is the total number of vehicles of mode j going through e. Weights w_j capture the fact that cars and buses may have different effects on congestion. For cars, the number of vehicles is a function of trips $f_{ehj} \equiv \sum_{m \in \mathcal{M}^e_{hj}} q_{mj}$, where $\mathcal{M}^e_{hj}$ is the set of all markets in which travelers take a route that goes through edge e.[21] For buses, the number of vehicles is a function of frequencies $f_{ehj} \equiv \sum_{r \in \mathcal{R}^e_j} k_{rj}$, where $\mathcal{R}^e_{hj}$ is the set of bus routes that go through e.

For road-based modes, the travel time over edge e at time h for mode j is given by:[22]

$$T^{\text{vehicle}}_{ehj} = \max\{T^0_{ej}, A_{ehj} \cdot F^{\beta_j}_{eh}\}. \tag{12}$$

This functional form is directly motivated by the empirical patterns in Figure 6. For every pair of neighboring CAs, there is a range with low vehicle flows for which the travel time is independent of vehicle flows. Travel time is then equal to an edge-mode specific *free-flow time* T^0_{ej} that captures road infrastructure and geography (including distance). The second term inside the maximum represents the range in which travel times increase with vehicle flows. Over that range, we assume a constant elasticity β_j of travel times to vehicle flows. A_{ehj} is an edge-mode specific scale factor that captures geography and road infrastructure.

We define the in-vehicle time in market m as the sum of the travel times over

[21] To account for the fact that there are often multiple travelers in the same car, we scale down the number of trips by the average occupancy by mode to obtain flows. See Appendix D.4.

[22] For trains, we assume in-vehicle times are constant. We take the expected times given by Google Maps.

all edges in the path P_{mj}:

$$T_{mj}^{\text{vehicle}} = \sum_{e \in P_{mj}} T_{ehj}^{\text{vehicle}}. \tag{13}$$

Next, we define how public transit wait times are determined. Travelers that choose public transit mode j in market m take an exogenous bus or train route r_m—the one suggested by Google Maps.[23] The frequency of a route is given by its fleet-size $k_{r_{mj}}$. The average time between two vehicles is $1/k_{r_{mj}}$.

However, from a traveler's perspective, the wait time until a vehicle arrives is not the same as the time between vehicles. Importantly, passengers need to wait longer the more unpredictable arrivals are due to schedule violations. If buses are always on schedule, for instance, the time between two buses is always exactly $1/k_{r_{mj}}$. The average passenger arrives halfway between two buses, so the expected wait time is $1/(2 \cdot k_{r_{mj}})$. At the opposite extreme, bus arrivals are a Poisson process—the arrival rate does not depend on whether the last bus arrived recently—and the expected wait time is $1/k_{r_{mj}}$.

We estimate a model that nests both extremes (for details see Appendix D.1). The model allows the time between two vehicles to follow an arbitrary distribution with mean $1/k_{r_{mj}}$. The expected wait time for passengers is given by

$$T_{mj}^{wait} = \frac{1+\omega^2}{2k_{r_{mj}}},$$

where ω is the coefficient of variation of the time between vehicles. We observe the deviations from schedules, which we use to estimate ω. Trains follow schedules closely, so we set $\omega = 0$. For buses, we find that $\hat{\omega}^2 = 0.194$, which means that there is substantially more variability than for trains.[24]

For ride-hailing, wait time T_{mj}^{wait} depends on three main factors: (1) lower waiting times when many drivers are working due to a large number of idle drivers; (2) higher waiting times when demand for ride-hailing trips is high, depleting avail-

[23] If Google Maps suggests a route with transfers, the wait time is the sum of individual wait times.
[24] To estimate this number, we compute actual times between buses and divide them by their average at the hour by route level. The variance of this ratio is our estimate $\hat{\omega}^2 = 0.194$.

able drivers; and (3) lower waiting times in areas with more idle drivers. We set up a model of driver movements that accounts for the higher concentration of idle drivers in neighborhoods with net trip inflows as well drivers' tendency to relocate towards areas with higher earnings opportunities. Appendix D.2 presents the details of this driver movement model.

3.4 Costs and environmental externalities

We assume that costs and environmental externalities are proportional to the number of vehicle miles driven. For cars and ride hailing, the number of miles driven depends on how many passengers choose to travel using these modes. For buses and trains, the number of miles driven depends on their frequency; hence, the marginal cost and externality of an additional passenger is zero. This is a good assumption as long as occupancy rates do not reach capacity; Figure 5 shows that that is the case for buses.

For all modes, the cost per mile accounts for fuel or energy, vehicle depreciation, and maintenance. For buses, trains, and ride hailing, it also includes labor costs. Environmental externalities account for the social cost of carbon, for which we use latest EPA proposal of $190 per tonne as the baseline number, as well as for the social cost of local pollutants, which we obtain from Holland et al. (2016).[25] Appendix D.4 describes in detail the numbers that we use for all costs and externalities. When we present our counterfactual results, we also conduct sensitivity analyses across a range of alternative values for each of these inputs.

[25] See EPA Issues Supplemental Proposal to Reduce Methane and Other Harmful Pollution from Oil and Natural Gas Operations.

4 Estimation and Computation

4.1 Demand model

In this section, we explain how to estimate parameters α_T and α_p^i. Recall that the utility of traveler i for taking mode j in market m is given by:

$$U_{mj}^i = \xi_{mj} + \alpha_T \cdot T_{mj} + \alpha_p^i \cdot p_{mj} + \epsilon_{mj}^i.$$

We define the outside option $j = 0$ as staying put, walking, or biking. We assume that the price coefficient takes the form $\alpha_p^i = \alpha_p / y_i^{1-\alpha_{py}}$, so that α_{py} captures the extent to which the price coefficient varies with income y_i.[26] To estimate α_T and α_p^i we need to address two important endogeneity concerns.

The first concern relates to prices. The prices of cars and public transit are fixed and are not affected by time-varying demand shocks. Therefore, we assume they are orthogonal to unobservable demand shocks, following a similar argument made by DellaVigna and Gentzkow (2019). On the other hand, ride-hailing prices adjust to demand conditions and thus are potentially correlated with demand shocks ξ_{mj}. To address this concern, we use price variation introduced by a ride-hailing surcharge that applies to weekday trips that either originate or end in a downtown zone between 6 am and 10 pm. In Appendix B we compute a difference-in-difference estimate from this policy, which implies an own-price elasticity of -1.42. We use this estimate to help identify our price coefficient by constructing a moment that equals the difference between the estimated and the model-implied demand response.

The second endogeneity issue concerns travel times, as these are an equilibrium object: positive demand shocks ξ_{jm} for road-based modes lead to more travel, inducing congestion and higher travel times. This type of endogeneity biases the travel time coefficient towards zero, as it would for prices in standard demand and supply models. To address this concern, we exploit the fact that travel times

[26] This functional form corresponds to a first-order Maclaurin series approximation of a utility function that follows a Box-Cox transformation, as in Miravete et al. (2023).

are not affected by congestion when there are few cars on roads, as shown in Figure 6. We instrument travel times T_{mj} with free-flow times T^0_{mj}, which do not depend on vehicle flows and, therefore, are not affected by within-day demand shocks. The variation in free-flow times is instead driven by permanent differences in infrastructure. For instance, between two CAs that are only connected by small roads with traffic lights, traffic will flow slower than if they were connected by a highway—and, importantly, these speed differences are different for cars and for buses (as we show in Figure 7). We include mode and market fixed effects to ensure that we only exploit the extent to which infrastructure differentially affects modes.

We implement the estimation using two-step GMM with two sets of moments. The first set of moments are indirect inference moments:

$$\mathbb{E}[(\hat{\eta}_{mj} - \tilde{\eta}_{mj})\mathbb{1}\{j = \text{ride-hail}, m \in \mathcal{M}_\tau\}] = \mathbf{0},$$

where $\hat{\eta}_{mj}$ is the ride-hailing elasticity from our differences-in-differences estimate (which we describe in Appendix B), $\tilde{\eta}_{mj}$ is the model-implied elasticity, and $\mathcal{M}_\tau$ are the markets affected by the surcharge policy.

The second set of moments take the form of orthogonality conditions of instruments $\mathbf{Z}_{mj}$, $\mathbb{E}[\mathbf{Z}_{mj}\xi_{mj}] = \mathbf{0}$. Our first instrument are free-flow times T^0_{mj}, as described above. We use public transit and car prices directly as instruments. We also interact those prices with income quintiles π^y_m in each market to identify heterogeneous sensitivity to prices.

To identify the parameter ρ that governs substitution towards the outside option, we follow Gandhi and Houde (2019) to construct local and quadratic differentiation instruments based on free-flow times, which measure the similarity of inside-modes in terms of travel time:

$$Z^{\text{local}}_{mj} = \sum_{j' \neq j} \mathbb{1}\{|T^0_{mj'} - T^0_{mj}| < SD_{T^0}\} \qquad \text{and} \qquad Z^{\text{quad}}_{mj} = \sum_{j' \neq j} (T^0_{mj'} - T^0_{mj})^2.$$

We expect differentiation instruments to affect ride-hailing prices—the only en-

dogenous prices—in a way that differs from other prices. Thus, we also interact these instruments with a ride-hailing indicator.

We follow the nested fixed-point algorithm outlined in Berry et al. (1995) to minimize the GMM objective function

$$J(\theta) = \hat{g}(\theta)' \cdot W \cdot \hat{g}(\theta),$$

where $\hat{g}(\theta)$ is the stacked vector of moment conditions.

Table 1: Demand estimation results

	Pooled					Peak/Off-Peak	
	(1)	(2)	(3)	(4)	(5)	(6)	
						Peak	Off-peak
α_T	-1.068	-1.692	-2.345	-2.415	-1.928	-1.824	-1.872
	(0.011)	(0.022)	(0.023)	(0.023)	(0.018)	(0.022)	(0.027)
α_p	-0.058	-0.155	-8.461	-3.416	-2.078	-2.388	-1.657
	(0.001)	(0.002)	(0.492)	(0.111)	(0.09)	(0.225)	(0.068)
α_{py}	·	·	-1.262	-0.588	-0.414	-0.696	-0.152
			(0.039)	(0.02)	(0.022)	(0.048)	(0.022)
ρ	·	·	·	·	0.262	0.376	0.162
					(0.012)	(0.017)	(0.017)
Estimator	OLS	IV	GMM	GMM	GMM	GMM	
Policy Moment			✓	✓	✓	✓	
Car Ownership				✓	✓	✓	
Nest					✓	✓	
Avg. VOT	18.41	10.89	23.88	14.65	13.47	19.47	9.81
VOT (Bot. Quintile)	·	·	2.44	3.26	3.62	3.9	3.42
VOT (Top Quintile)	·	·	64.24	32.36	27.94	45.32	18.09
Avg. Price Elast.	-0.2	-0.53	-0.5	-0.61	-0.65	-0.55	-0.72
Avg. Time Elast.	-0.58	-0.91	-1.26	-1.27	-1.29	-1.44	-1.07
M	92,284	92,284	91,908	91,561	91,561	42,989	48,572
N	281,755	281,755	281,042	280,185	280,185	136,337	143,848

Notes: This table presents demand estimation results from the specifications outlined in section 4.1. We obtain the average VOT by first computing the within market average VOT as the weighted average of α_T/α_p^i and then averaging across markets, with weights given by market size. Average elasticities are computed as the weighted average of own-price and own-time elasticities across all mode-market observations, with weights given by market size. We drop markets without income information in specifications with income heterogeneity.

Table 1 shows estimates for several specifications of our model, gradually build-

ing up from a simple logit model to the main specification that we outline in Section 3.3. The first two columns show estimates from a model without a nest or heterogeneity across consumers—a standard logit model. Column (1) presents OLS estimates, and column (2) presents IV estimates in which the instruments are free-flow times, differentiation instruments, and the prices of all modes except for ride-hailing. These specifications do not yet use the indirect-inference moment that makes use of the ride-hail surcharge. As expected, the IV specification substantially increases the sensitivity to travel times and prices.

Specification (3) allows for heterogeneity in price sensitivity, and it also includes the indirect-inference moment as well as our full set of instruments. We obtain overall larger price and travel-time sensitivities, and the value of time (VOT) increases. In specification (4), we account for variation in car ownership by allowing for random choice set variation across travelers based on census car ownership data. Specification (5) also introduces the nested structure of taste shocks.

Model (6) is our main specification. It is the same model as specification (5), but we allow for different parameters during peak hours—those times when the ride-hailing surcharge is active—and off-peak hours. We find substantial variation in the VOT: it ranges from \$3.90 to \$45.30 during peak hours (for the lowest and highest income quintiles, respectively), and from \$3.40 to \$18 during off-peak hours. The city-wide average VOT is about \$15. Overall, the VOT is quite stable across different specifications, ranging from \$10.89 to \$23.88. Appendix G.1 presents a number of additional robustness checks across which the VOT is still quite stable, ranging from \$8.30 to \$12.78.

We now present the main spatial patterns implied by our estimates. Figure 8 shows the VOT by CA. It tends to be higher in the North Side, which is characterized by higher incomes. In the South Side, we mostly observe low values, with a few exceptions: Midway airport (center left) as well as the neighborhoods of Beverly, Mount Greenwood, and Morgan Park (bottom left), which were popular white-flight destinations during the 1950s and 1960s. These patterns correlate closely with the movement patterns in the bottom panel of Figure 1.

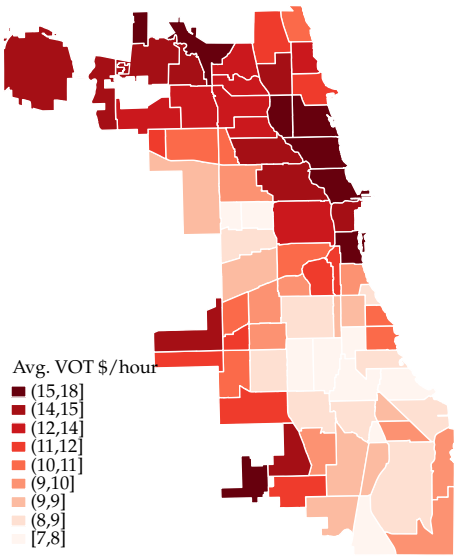


Figure 8: Value of Time across space

Notes: This figure shows the average VOT by CA implied by our main demand specification, column (6) of Table 1. Income heterogeneity is driven by differences in the price coefficient.

Table 2 presents substitution patterns in the form of diversion ratios. There is substantially more substitution to the outside option among travelers who do not own cars. Among car owners, substitution towards the outside option is twice as important from cars as it is from the remaining modes. High-income travelers substitute more often to cars and ride hailing than low-income travelers. By contrast, low income travelers are more likely to substitute towards buses or the outside option.

4.2 Traffic congestion

In this section, we estimate the traffic congestion model from Section 3.3.2, which models the in-vehicle time for edge e during hour h for mode j as:

$$T_{ehj}^{\text{vehicle}} = \max\{T_{ej}^{0}, A_{ehj} \cdot F_{eh}^{\beta_j}\},$$

Table 2: Diversion ratios

(a) Overall

From \ To	Bus	Car	RideH	Train	Outside
Bus	·	0.29	0.06	0.10	0.56
Car	0.10	·	0.05	0.08	0.77
RideH	0.14	0.31	·	0.09	0.46
Train	0.16	0.25	0.07	·	0.52

(b) Car

From \ To	Bus	Car	RideH	Train	Outside
Bus	·	0.57	0.03	0.06	0.35
Car	0.10	·	0.04	0.08	0.78
RideH	0.06	0.59	·	0.05	0.30
Train	0.07	0.54	0.04	·	0.36

(c) No Car

From \ To	Bus	Car	RideH	Train	Outside
Bus	·	·	0.10	0.14	0.76
Car	·	·	·	·	·
RideH	0.23	·	·	0.13	0.64
Train	0.23	·	0.10	·	0.66

(d) Bottom Income Quintile

From \ To	Bus	Car	RideH	Train	Outside
Bus	·	0.28	0.01	0.10	0.62
Car	0.10	·	0.00	0.07	0.83
RideH	0.15	0.25	·	0.09	0.51
Train	0.16	0.25	0.01	·	0.58

(e) Top Income Quintile

From \ To	Bus	Car	RideH	Train	Outside
Bus	·	0.43	0.13	0.07	0.36
Car	0.10	·	0.16	0.07	0.67
RideH	0.10	0.45	·	0.08	0.38
Train	0.10	0.39	0.15	·	0.36

Notes: These tables present the average diversion ratios implied by our main demand specification, column (6) of Table 1, for various consumer types. Individual diversion ratios are averaged across markets, weighted by market size.

where $F_{eh} = \sum_j w_j f_{ehj}$. We take $w_{car} = w_{ride-hail} = 1$ and $w_{bus} = 2$, implying that the marginal effect of a bus on congestion is twice the effect of a standard vehicle.[27]

As we can see in Figure 6, observations between 12 am and 5 am overwhelmingly lie in the region where travel times do not depend on traffic. For that reason, we define the free-flow time T^0_{ej} for cars to be the average travel time during these early morning hours. For buses, we take the average between 10pm and 12am, which avoids issues that arise becuase ofunusual early-morning schedules.

When $T^{\text{vehicle}}_{ehj} \geq T^0_{ej}$, our model becomes $T^{\text{vehicle}}_{ehj} = A_{ehj} \cdot F^{\beta_j}_{eh}$. To estimate A_{ehj} and β_j, we focus on observations in which the time T^{vehicle}_{ehj} is above 110% of the free-flow time, which account for 70% of our sample. Assuming that $a_{ehj} = \log A_{ehj} =$

[27] These values follow London's Traffic Modelling Guidelines.

$a_e + \varepsilon_{ehj}$, our estimation equation becomes:

$$\log T_{ehj}^{\text{vehicle}} = a_e + \beta_j \log F_{eh} + \varepsilon_{ehj}. \tag{14}$$

The edge fixed effect a_e captures any edge-specific differences in geography or infrastructure that determine travel times. The remaining error ε_{ehj} captures unobservable shocks that vary across hours of the week h within edge e.

Table 3: Traffic congestion estimation results

Dependent Variable: Log travel time in traffic

	Bus			Car		
	(1)	(2)	(3)	(4)	(5)	(6)
Log Flow	0.092***	0.059***	0.101***	0.128***	0.100***	0.168***
	(0.006)	(0.006)	(0.008)	(0.005)	(0.005)	(0.004)
Edge FE	✓	✓	✓	✓	✓	✓
Weather controls		✓	✓		✓	✓
IV			✓			✓
within R^2	0.093	0.129	0.116	0.411	0.529	0.443
First-stage F			2767.096			4487.258
Observations	7962	7962	7962	11739	11739	11739

Notes: This table shows the regression estimates for the elastic portion of the congestion function for buses, columns (1)-(3), and cars, columns (4)-(6). The unit of observation is an edge. The dependent variable is the log of travel times for the corresponding mode, while the independent variable is the log vehicle flows. Specifications (1) and (4) control for edge fixed effects, specifications (2) and (5) add weather controls (temperature, visibility, and precipitation), and specifications (3) and (6) use the potential market size as an instrument for vehicle flows. Robust standard errors are in parentheses. $^{*}\ p < 0.10$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$.

Table 3 presents the estimates of Equation 14 for buses and cars. All specifications include edge fixed effects, so we only use within-edge variation across hours. Columns (1) and (4) present estimates without any additional controls. The identification assumption is that, within an edge, shocks to the traffic congestion technology are uncorrelated with the number of vehicles. Since we aggregate data at the hour of the week level, the only threat to identification are shocks that repeat themselves every week, such as weather patterns or changes in visibility due to sunlight (as discussed by Akbar and Duranton, 2017). We control for such variables in columns (2) and (5).

A remaining concern is that travelers may re-optimize their route choices in response to unobservable but expected traffic shocks, such as planned construction during certain hours of the day. To address those concerns, we follow the strategy used by Kreindler (2023). In columns (3) and (6), we instrument traffic flows with the city-wide number of travelers by hour. This strategy is valid as long as the city-wide demand for travel is driven by daily patterns—commuting to work in the morning, going out for dinner in the afternoon, etc.—rather than by unobserved traffic shocks, like anticipated construction or public transit disruptions.

We find elasticities of travel time with respect to traffic flows between 0.06 and 0.1 for buses and between 0.10 to 0.17 for cars; both are comparable with existing estimates in the literature (Akbar and Duranton, 2017; Couture et al., 2018). The main elasticities that we use for our model are those in columns (3) and (6).[28]

4.3 Solving for Equilibrium and the Planner's Problem

Before we move on to describe our counterfactuals, we restate our equilibrium Definition 1 in the context of our empiricial model. Then we explain how we compute equilibria and how we solve the planner's problem.

Definition 2. *A transportation equilibrium is a vector of trips* $\{q_{mj}\}_{mj}$ *and a vector of travel times* $\{T_{mj}\}_{mj}$ *such that:*

1. *Trips are determined from the demand model:* $q_{mj} = N_m \cdot \mathbb{P}_{mj}(T_{mj})$ *given by equation 10,* $\forall j \in \mathcal{J}, m \in \mathcal{M}$.

2. *Total travel times* $\forall j \in \mathcal{J}, m \in \mathcal{M}$ *are the sum of three components* $T_{mj} = \gamma(T_{mj}^{walk} + T_{mj}^{wait}) + T_{mj}^{vehicle}$, *where:*

   (a) *Time in vehicle for road-based modes result from the congestion function:* $T_{mj}^{vehicle} = \sum_{e \in r_{mj}} T_{ehj}^{vehicle} = \sum_{e \in r_{mj}} \max\{T_{ej}^0, A_{ehj} \cdot F_{eh}^{\beta_j}\}$ *as determined by equation 12.*

[28] Our estimated model predicts the in-vehicle times of short trips well, but it systematically overestimates in-vehicle times for long trips. This is likely because long trips tend to take highways. We therefore scale down travel times by a factor that depends on distance. See Appendix D.3.

Trips and fleet sizes are aggregated to obtain vehicle flows $F_{eh}^{\beta_j}$ as described in Section 3.3.2.

(b) *Wait times for buses and trains are given by $T_{mj}^{wait} = \frac{1+\omega^2}{2 \cdot k_{mj}}$, where k_{mj} is the number of buses and trains running at the hour. Wait time for cars are zero.*

(c) *Walk times are fixed for all modes.*

Point 1 represents the demand condition, equation (2): travelers take prices and travel times as given when making their travel decisions. Point 2 represents the supply condition, equation (3): these decisions generate vehicle flows across locations, which feed into congestion of edges and in turn determine travel times. The city is in equilibrium when both conditions hold simultaneously.

To find an equilibrium, we write the equilibrium conditions as a fixed point. Let $f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) \equiv q(\mathbf{p}, T(\mathbf{q}, \mathbf{k}))$. If travelers believe that that the number of trips will be $\mathbf{q}$—and, hence, they believe that travel times will be $T(\mathbf{q}, \mathbf{k})$—this function gives the number of trips that will actually occur. An equilibrium is a fixed point of $f^{\mathbf{p},\mathbf{k}}$: travelers' beliefs must be consistent with the realized number of trips.[29]

Naive algorithms to find fixed points—such as simple or dampened fixed-point iteration—often diverge. We rewrite this problem as a root-finding problem ($f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) - \mathbf{q} = 0$) and use a limited-memory version of Broyden's method to solve it. Appendix D.5 describes in detail the algorithm we use. Once we find an equilibrium, we can compute all quantities that go into the city government's objective function. To reduce the computational burden, we only simulate the market during six representative hours of the week, which we aggregate as a weighted sum to obtain outcomes for one whole week.[30] Appendix E shows that our estimated model fits the data well.

To solve the social planner's problem—which involves a budget constraint—we follow the augmented Lagrangian method (Nocedal and Wright, 2006), where we iteratively maximize problems that approximate the Lagrangian of the main

[29] To see why this is consistent with Definition 1, plug in $\mathbf{t} = T(\mathbf{q}, \mathbf{k})$ into $\mathbf{q} = q(\mathbf{p}, \mathbf{t})$ to obtain $\mathbf{q} = q(\mathbf{p}, T(\mathbf{q}, \mathbf{k})) = f^{\mathbf{p},\mathbf{k}}(\mathbf{q})$.

[30] Those representative hours are weekdays at 3 am, 8 am, 12 pm, and 5 pm as well as weekends at 3 pm and 10 pm. We give them weights 50, 20, 25, 25, 16, and 32, respectively.

problem until convergence. Every evaluation of the Lagrangian requires solving for the transportation equilibrium. Further details are provided in Appendix D.6.

5 Optimal Policy Design

In what follows, we explore counterfactual policy designs based on the optimality conditions we describe in Equation 4. We start by analyzing coarse policies that change overall prices and frequencies for all markets. Section 5.3 analyzes more granular policies.

We first analyze the case of an unconstrained planner who only sets public transit prices and frequencies, which we call *Transit*. To quantify the additional distortions that are caused by budget considerations, we also consider a budget constrained planner that cannot exceed the current public transit deficit of Chicago (*Transit, Budget*). We then separately analyze the effect of *Road Pricing*. To explore the interactions of these policies, we next analyze the case where the planner can use them simultaneously (*Transit + Road Pricing*).

We now discuss each of these counterfactuals in detail. Throughout this discussion, we refer to Table 4, which presents our main counterfactual results, including the optimal policy levers and their welfare effects. Each column represents one counterfactual policy, and we report results relative to the *Status Quo* (column 1). We also refer to Figures 9 and 10, which decompose the forces that give rise to the optimal policies for buses and cars, as in our theoretical results from Section 3.2 (the expressions that we use for this higher-dimensional problem, which include spillovers across markets, are derived in Appendix C.3).[31] In these decompositions, red bars represent effects that the planner should correct through higher prices and times; yellow bars represent effects that should be corrected with lower prices and times.

In *Transit*, the planner would want to set slightly negative prices for buses and trains. The reason for this is, first, that the components of the optimal price cor-

[31] Appendix G.2 presents figures for trains, which are almost identical to those for buses.

Table 4: Counterfactual results

		Status Quo (1)	Transit (2)	Transit, Budget (3)	Road Pricing (4)	Transit + Road Pricing (5)
Panel A: Prices						
Avg. Price ($)	Bus	1.09	-0.33	0.65	1.09	0.16
	Train	1.33	-0.37	1.02	1.33	0.26
Road Tax ($/km)		0	0	0	0.35	0.32
Panel B: Wait Times and Frequencies						
Avg. Wait (min)	Bus	7.06	7.15	8.19	7.06	7.15
	Train	4.37	4.05	4.55	4.37	4.06
Δ Frequency	Bus	0%	-1.33%	-13.88%	0%	-1.24%
	Train	0%	9.21%	-2.58%	0%	8.91%
Panel C: Trips						
Number Of Trips (M/week)	Bus	3.7	5.0	3.8	4.2	5.0
	Train	2.7	3.5	2.8	2.9	3.5
	Ride-hailing	3.0	2.9	3.0	3.1	3.1
	Car	21.3	20.5	21.2	17.6	17.4
	Total	30.6	31.9	30.7	27.8	28.9
Panel D: Welfare						
Δ Welfare ($M/week)		0	1.54	0.39	4.57	5.27
Δ CS ($M/week)		0	12.65	0.03	-29.11	-18.54
Δ City Surplus ($M/week)		0	-10.99	0	28.31	18.96
Δ Transit Surplus ($M/week)		0	-10.99	0	0.82	-6.32
Road Taxes ($M/week)		0	0	0	27.49	25.28
Δ Externalities ($M/week)		0	-0.62	-0.35	-3.59	-3.69

Notes: This table presents the changes in prices, frequencies, trips, and welfare relative to the *Status Quo* (column 1) across different counterfactual scenarios. Column 2, *Transit*, changes public transit prices and frequencies without budget considerations. Column 3 (*Transit, Budget*) repeats the same exercise subject to a budget constraint. Column 4 uses *Road Pricing*. Column 5 combines both, *Transit + Road Pricing*.

responding to marginal costs, environmental externalities, and network effects are all zero: an additional passenger has zero social cost in our model, which is a good approximation given the low utilization of buses (Figure 5). The only other com-

ponent of the optimal price is the diversion term, which is negative because the planner wants to divert travelers away from socially underpriced cars.

The optimal wait times for buses are slightly higher than in the status quo. The marginal costs, environmental externalities, and network externalities (primarily congestion) of buses are all positive, leading to fewer buses and, therefore, higher wait times. The effects from environmental externalities and congestion are small relative to the direct cost effect. The only force against fewer buses is again the planner's motive to divert passengers away from underpriced cars. For trains, on the other hand, the planner wants to increase frequency relative to the status quo.

These changes in *Transit* increase welfare by $1.54 million per week relative to the status quo. This improvement partly stems from a reduction in the cost of environmental externalities of $0.62 million per week. Additionally, consumer surplus increases substantially—by $12.65 million per week, or $4.7 per resident—because both prices as well as train wait times are reduced. However, these changes also increase the planner's deficit by $11 million per week.

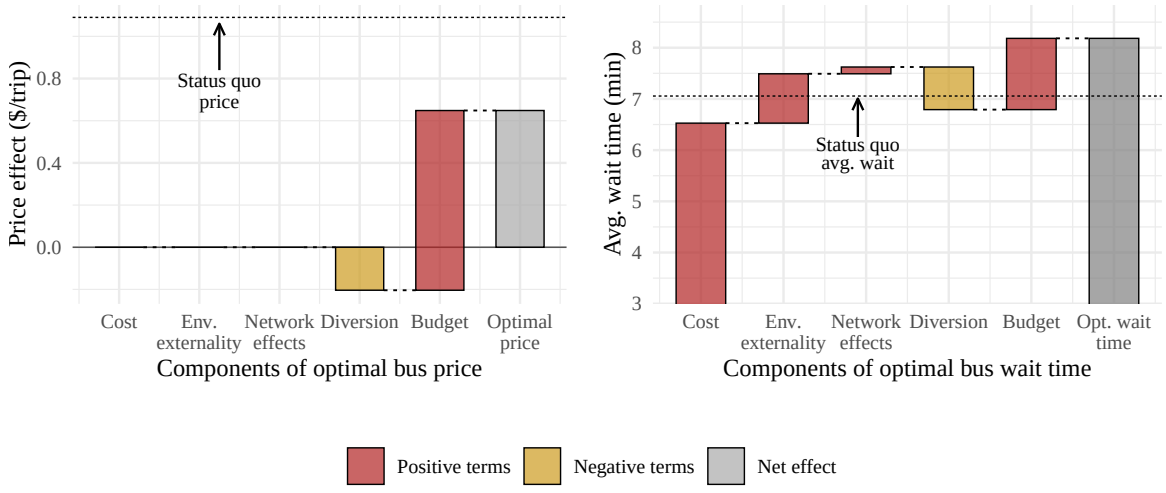


Figure 9: Optimal bus price and wait time decomposition for *Transit + Budget*

Notes: This graph shows a decomposition of the optimal prices and travel times for buses corresponding to our theoretical decomposition in Section 3.2. Red bars indicate terms that lead prices and travel times to be higher and yellow bars indicate terms that lead them to be lower.

We next introduce budget considerations (*Transit, Budget*), which have a large impact on the welfare effect of optimal transit policies. Welfare increases only by \$0.39 million per week relative to the status quo. The smaller welfare effect is partly driven by the fact that the reduction of environmental externalities is only \$0.35 million per week. However, the most significant difference is in the surplus of travelers, which remains almost the same as in the *Status Quo*. Travelers no longer benefit because a binding budget constraint leads the government to make two adjustments that hurt travelers: it increases fares, and it reduces the frequency of both buses and trains by around 12%. The budget terms in Figure 9 represent these budget effects. These figures do not distinguish between the different components of budget distortions. For prices, we find that the markup term is the most important source of the distortion. In addition, due to large differences in the value of time across travelers, the Spence distortion also contributes to larger wait times for public transit. Despite the reduction in frequencies, the number of public transit trips increases by 0.2 million per week due to the reduced prices.

A comparison between the *Status Quo* and the optimal transit policies in *Transit, Budget* reveals the extent to which the current prices and frequencies deviate from the optimum. In *Transit, Budget*, both bus and train fares are substantially lower. Frequencies are also reduced, especially for buses, with a decrease of 13.88%. This indicates that the current frequency of buses is excessively high, which contributes to the low bus utilization that we find in the data (Figure 5). One possible reason why the CTA allocates a large amount of resources towards buses is its desire to target low-income neighborhoods, which are disproportionately reliant on buses.[32] However, we show below that low income travelers benefit from the combination of price and frequency adjustments in *Transit, Budget*.

We now turn to *Road Pricing*.[33] When it is the only lever available to the gov-

[32] We explore which welfare weights for different income groups rationalize the observed frequencies under the current prices. The government must weigh the welfare of travelers in the lowest income quintile 3.93 times more than more than that of travelers in the highest quintile.

[33] In our *Road Pricing* counterfactuals, ride hailing trips do not pay road taxes. In a different counterfactual, we find that it is optimal to decrease the price of ride hailing by around 20%: the markup charged by ride-hailing companies is larger than the optimal Pigouvian tax.

ernment, the optimal per-km tax is 35 cents, or $13.3 per day for the average car commuter.[34] Figure 10 shows that this roughly doubles the status quo price of driving a car, which is simply the marginal cost. About 40% of the tax is due to environmental externalities and another 40% is due to congestion externalities. The remaining portion of the tax is intended to divert travelers to public transit. The budget term is zero because road tax revenue generates a fiscal surplus, so the budget constraint becomes nonbinding.

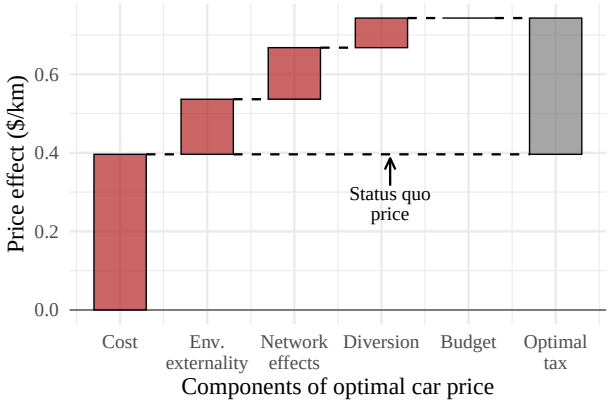


Figure 10: Optimal car price decomposition in the *Road Pricing* scenario

Notes: This figure shows the price decomposition for cars, following our theoretical derivations in Section 3.2. Red bars indicate terms that lead optimal car prices to be higher.

The overall welfare gains from *Road Pricing*, $4.57 million per week, are much larger than from transit policies alone. However, these gains predominantly result from a reduction in environmental and pollution externalities, while travelers are worse off. In the absence of rebates, consumer surplus decreases by a $29.11 million per week, or $10.78 per resident per week. Even if the government fully rebated the revenue it collected from road taxes, consumers would lose $0.8 million in weekly surplus.[35] The total number of trips goes down by 2.8 million per week.

Finally, when transit polices and road pricing are combined (*Transit + Road Pricing*) the planner can achieve large welfare gains by reducing externalities while

[34] If, instead of a per-km tax, the government sets a cordon price for cars entering downtown Chicago, the optimal level is $8.28.

[35] *Road Pricing* also impacts commuters who enter and exit the City of Chicago. Assuming these travelers are perfectly inelastic, $41.9M in tax revenue would be collected from them.

also increasing consumer surplus. By setting road taxes at \$0.32 per km, the government runs a surplus, resulting in a slack budget constraint. This eliminates resulting distortions and allows for public transit prices and frequencies that closely resemble those under *Transit*, in which the government faces no budget constraint. Specifically, the government offers virtually free public transit, with bus fares at \$0.16 and train fares at \$0.26. These policies lead to a welfare gain of \$5.27 million per week and a decrease of 1.7 million in weekly trips. As with *Road Pricing*, consumer surplus decreases without rebates—by \$18.54 million per week. On the other hand, if the government rebates its surplus, consumer surplus can increase by as much as \$0.48 million per week. Combining efficient road pricing and cross-subsidizing public transit ends up benefiting travelers.

5.1 Distributional Effects

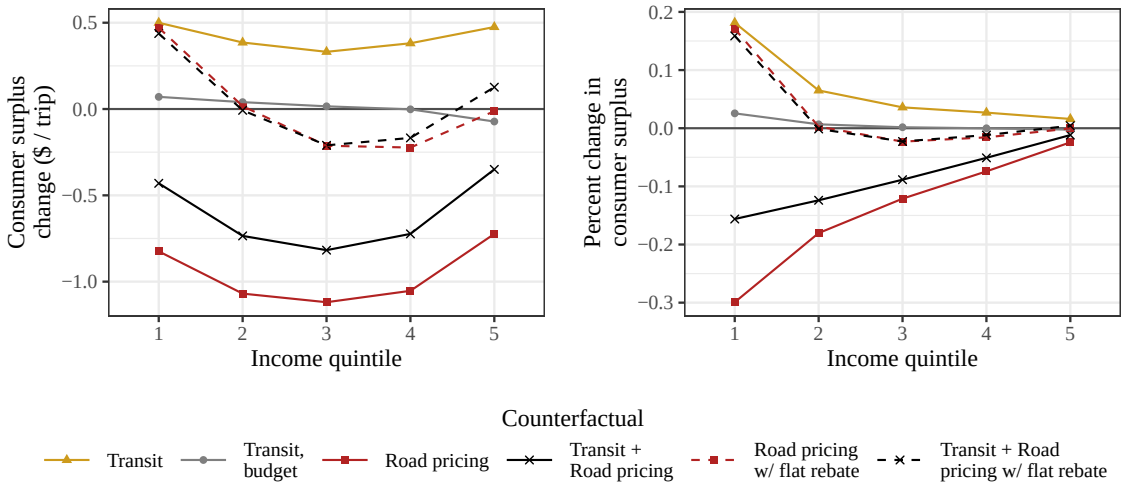


Figure 11: Change in consumer surplus across income quintiles

Notes: This figure presents changes in consumer surplus by income quintile relative to the *Status Quo* under optimal policies across different counterfactual scenarios. Panel (a) displays net changes in dollars per trip. Panel (b) displays percent changes in consumer surplus. Solid lines represent our main counterfactuals in Table 4. Dashed lines represent scenarios in which road pricing revenue is rebated back to residents as a flat refund.

Figure 11 shows the effects of our counterfactuals on consumers across income quintiles. The left panel measures changes in consumer surplus per trip. All trav-

elers but the highest income group benefit from public transit interventions in isolation. Under road pricing without rebates (solid lines), all income groups are worse off. Losses are U-shaped because middle income consumers are the most reliant on cars, as we have shown in Figure 4. When we measure those losses as a percentage of consumer surplus, on the other hand, we find that road pricing is highly regressive. Policies that deliver the largest efficiency gains are also the ones that hurt low-income consumers the most relative to their income. However, the government can undo this regressivity by rebating revenue back to residents as a flat refund (dashed lines). Low-income travelers benefit the most, both in absolute and relative terms. Only middle-income travelers are worse off than in the status quo.

5.2 Sensitivity Analysis

We explore the extent to which the optimal policies from Table 4 are sensitive to changes in key model parameters (costs, externalities, and demand sensitivity to prices and times). We find that the optimal prices are very robust: 10% changes in parameter values change prices by less than two cents. On the other hand, optimal frequencies are somewhat more sensitive, but still within 5% of our main results. Additional details on these computations are provided in Appendix F.

5.3 More Granular Policies

We explore the additional gains that can be achieved by setting different prices and frequencies across different times of day, location, and baseline utilization rates of bus routes. Table 5 presents results from several granular policies that we consider.

Setting different road taxes for the city center, across different times of the day, or both, increases the welfare gains from road pricing by at most 1.3%. The additional gains are small because the optimal road taxes are relatively homogeneous, in part because environmental externalities are invariant to space or time of day. Furthermore, differences in the remaining two components of the optimal tax—

network effects and the diversion term—tend to offset each other.

More granular frequency adjustments, on the other hand, result in large welfare gains. The government would increase frequencies during rush hour and decrease them at other times. It would also increase the frequency of high utilization routes by more than 50% while decreasing the frequency of low utilization routes by almost 40%. These adjustments result in welfare gains that are four to six times larger than those from uniform frequency adjustments. However, these improvements achieve less than half of the welfare gains from combined road pricing and transit policies.

Table 5: Granular counterfactual results

Panel A: Road Pricing	Uniform (1)	Time Heterogeneity (2)	Spatial Heterogeneity (3)	Time + Spatial Heterogeneity (4)
Base ($/km)	0.35	0.33	0.32	0.30
Rush Hour ($/km)	·	0.37	·	0.34
CBD ($/km)	·	·	0.56	0.51
Rush Hour × CBD ($/km)	·	·	·	0.61
Δ Welfare ($M/week)	4.57	4.59	4.62	4.63

Panel B: Transit, Budget		Uniform (1)	Time Heterogeneity (2)	Utilization Heterogeneity (3)
Δ Bus Frequency (%)	Base	-13.88	-33.69	-37.98
	Rush Hour	·	6.62	·
	High Utilization	·	·	52.29
Δ Train Frequency (%)	Base	-2.58	-32.31	-1.97
	Rush Hour	·	32.16	·
	Bus Price ($)	0.65	0.57	0.58
	Train Price ($)	1.02	0.90	0.89
	Δ Welfare ($M/week)	0.39	1.76	2.33

Notes: This table presents changes in prices, frequencies, and welfare relative to the status quo across different counterfactual scenarios. Panel A considers different road pricing scenarios: a uniform price (column 1), a time differentiated price (column 2), a spatially differentiated price (column 3), and a time and spatially differentiated price (column 4). Panel B considers different scenarios for adjusting transit prices and frequencies: a uniform adjustment (column 1), a time differentiated adjustment (column 2), and a utilization differentiated adjustment (column 3).

6 Discussion

We now discuss some of the simplifying assumptions that keep our model tractable. First, our model does not account for intertemporal substitution directly. Instead, it captures it indirectly as substitution towards the outside option. This approach allows us to model elasticities to own prices and own travel times correctly; however, the downside is that we are not able to capture spillover effects of policies across different hours of the week. Kreindler (2023) finds that intertemporal choices are rather inelastic and peak-spreading policies have a limited impact, suggesting that allowing for inter-temporal substitution would not have a large effect on our findings.

Second, although travelers often decide the mode of transportation for outbound and return trips jointly, we only model individual trips. This choice arises from a data limitation: we are not able to link together two trips from the same rider. Once again, the main challenge this brings to our model is that we are not able to capture spillover effects between different hours of the week.

As noted in the introduction, our model does not capture how residents and firms relocate in response to transportation policies. Prior research finds that long-run adjustments to transportation policy are limited. Herzog (2023) finds that sorting attenuates the welfare effects of time savings due to road pricing by around 20%. Barwick et al. (2021), on the other hand, show that residential sorting increases the overall welfare effects of road pricing by 18%.

A final simplifying assumption is that travelers do not adjust routes in counterfactuals. Google Maps data provides us with good routing data for the status quo. Absent infrastructure investments, reoptimizing routes relative to these routes is unlikely to have large impacts.

7 Conclusion

In this paper, we measure the welfare effects of urban transportation policies and explore how a budget-constrained planner should choose among a portfolio of policies. Based on a theoretical framework, we derive expressions for optimal policies that show that budget considerations introduce inefficiencies. We then quantify empirically the welfare effects of such policies in Chicago by constructing a dataset that captures granular mode choices across the city. Our results show that the government can undo the "monopoly" distortions that arise due to budget considerations by using road pricing revenues to cross-subsidize public transit. Indeed, recent transit policies in London and New York explicitly designate the revenues from road pricing to fund public transit. Our results highlight that such combined policy approaches can eliminate inefficiencies.

References

Akbar, P., Couture, V., Duranton, G. and Storeygard, A. (2023). Mobility and congestion in urban india. *American Economic Review* 113(4):1083–1111.

Akbar, P. and Duranton, G. (2017). Measuring the cost of congestion in highly congested city: Bogotá. CAF - Working paper N° 2017/04, CAF.

Allen, T. and Arkolakis, C. (2022). The welfare effects of transportation infrastructure improvements. *The Review of Economic Studies* 89(6):2911–2957.

Arnott, R. (1996). Taxi travel should be subsidized. *Journal of Urban Economics* 40(3):316–333.

Arnott, R., De Palma, A. and Lindsey, R. (1990). Economics of a bottleneck. *Journal of Urban Economics* 27(1):111–130.

Arnott, R., De Palma, A. and Lindsey, R. (1993). A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *American Economic Review* pp. 161–179.

Barwick, P.J., Li, S., Waxman, A.R., Wu, J. and Xia, T. (2021). Efficiency and equity impacts of urban transportation policies with equilibrium sorting. Working Paper 29012, National Bureau of Economic Research.

Berry, S., Levinsohn, J. and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica* 63(4):841–890.

Berry, S.T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics* pp. 242–262.

Bordeu, O. (2023). Commuting infrastructure in fragmented cities. Working paper, University of Chicago Booth School of Business.

Brancaccio, G., Kalouptsidi, M. and Papageorgiou, T. (2020). Geography, transportation, and endogenous trade costs. *Econometrica* 88(2):657–691.

Brancaccio, G., Kalouptsidi, M., Papageorgiou, T. and Rosaia, N. (2023). Search Frictions and Efficiency in Decentralized Transport Markets. *The Quarterly Journal of Economics* 138(4):2451–2503.

Brinkman, J. and Lin, J. (2022). Freeway revolts! the quality of life effects of highways. *The Review of Economics and Statistics* pp. 1–45.

Brooks, L. and Liscow, Z. (2023). Infrastructure costs. *American Economic Journal: Applied Economics* 15(2):1–30.

Buchholz, N. (2021). Spatial Equilibrium, Search Frictions, and Dynamic Efficiency in the Taxi Industry. *The Review of Economic Studies* 89(2):556–591.

Buchholz, N., Doval, L., Kastl, J., Matjka, F. and Salz, T. (2024). The value of time: Evidence from auctioned cab rides. Working Paper 27087, National Bureau of Economic Research.

Castillo, J.C. (2023). Who benefits from surge pricing? Working paper, University of Pennsylvania.

Cook, C. and Li, P.Z. (2023). Value pricing or lexus lanes? the distributional effects of dynamic tolling. Working paper, Stanford University.

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

DellaVigna, S. and Gentzkow, M. (2019). Uniform pricing in us retail chains. *The Quarterly Journal of Economics* 134(4):2011–2084.

Durrmeyer, I. and Martínez, N. (2023). Dp18332 the welfare consequences of urban traffic regulations. CEPR Discussion Paper No. 18332, CEPR.

Fajgelbaum, P.D. and Schaal, E. (2020). Optimal transport networks in spatial equilibrium. *Econometrica* 88(4):1411–1452.

Frechette, G.R., Lizzeri, A. and Salz, T. (2019). Frictions in a competitive, regulated market: Evidence from taxis. *American Economic Review* 109(8):2954–92.

Gaineddenova, R. (2022). Pricing and efficiency in a decentralized ride-hailing platform. Working paper, University of Wisconsin-Madison.

Gandhi, A. and Houde, J.F. (2019). Measuring substitution patterns in differentiated-products industries. Working Paper 26375, National Bureau of Economic Research.

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

Herzog, I. (2023). The city-wide effects of tolling downtown drivers: Evidence from london's congestion charge. Working paper, Huron University College.

Holland, S.P., Mansur, E.T., Muller, N.Z. and Yates, A.J. (2016). Are there environmental bene-

fits from driving electric vehicles? the importance of local factors. *American Economic Review* 106(12):3700–3729.

Kreindler, G. (2023). Peak-hour road congestion pricing: Experimental evidence and equilibrium implications. Working Paper 30903, National Bureau of Economic Research.

Kreindler, G., Gaduh, A., Graff, T., Hanna, R. and Olken, B.A. (2023). Optimal public transportation networks: Evidence from the world's largest bus rapid transit system in jakarta. Working Paper 31369, National Bureau of Economic Research.

Lagos, R. (2003). An analysis of the market for taxicab rides in new york city. *International Economic Review* 44(2):423–434.

Leccese, M. (2022). Asymmetric taxation, pass-through and market competition: Evidence from ride-sharing and taxis. In: *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, p. 371–372. New York, NY, USA: Association for Computing Machinery.

Miravete, E., Seim, K. and Thurk, J. (2023). Elasticity and curvature of discrete choice demand models. CEPR Discussion Paper No. 18310, CEPR.

Nocedal, J. and Wright, S.J. (2006). *Numerical Optimization*. Springer New York, NY.

Parry, I.W.H. and Small, K.A. (2009). Should urban transit subsidies be reduced? *American Economic Review* 99(3):700–724.

Pigou, A. (1932). *The Economics of Welfare*. Macmillan.

Ramsey, F.P. (1927). A contribution to the theory of taxation. *The Economic Journal* 37(145):47–61.

Rosaia, N. (2023). Who benefits from surge pricing? Working paper, Columbia Business School.

Severen, C. (2023). Commuting, Labor, and Housing Market Effects of Mass Transportation: Welfare and Identification. *The Review of Economics and Statistics* 105(5):1073–1091.

Small, K.A. (1982). The scheduling of consumer activities: work trips. *American Economic Review* 72(3):467–479.

Small, K.A. (2012). Valuation of travel time. *Economics of Transportation* 1(1):2–14.

Small, K.A., Winston, C. and Yan, J. (2005). Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica* 73(4):1367–1382.

Spence, A.M. (1975). Monopoly, quality, and regulation. *The Bell Journal of Economics* pp. 417–429.

Tsivanidis, N. (2023). Evaluating the impact of urban transit infrastructure: Evidence from bogotá's transmilenio. Working paper, University of California, Berkeley.

Yang, J., Purevjav, A.O. and Li, S. (2020). The marginal cost of traffic congestion and road pricing: evidence from a natural experiment in beijing. *American Economic Journal: Economic Policy* 12(1):418–53.

Zhao, J., Rahbee, A. and Wilson, N.H. (2007). Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering* 22(5):376–387.

Online Appendix

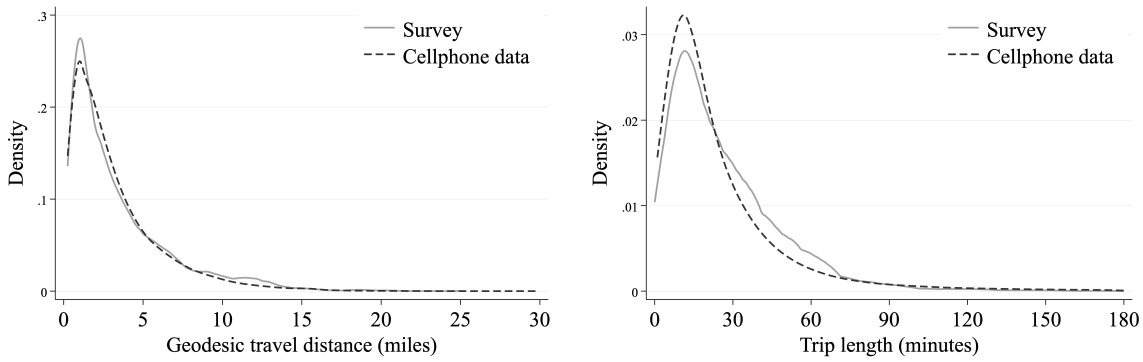
A Data Construction and Validation

This section provides an overview of how we construct our sample of trips based on the raw cellphone data. Supplementary Appendix S1 provides a detailed description. The raw data are composed of pings with timestamps, latitudes, longitudes, and device identifiers. We subset these data to a rectangle corresponding to the Chicago Metropolitan Agency for Planning (CMAP) region[36] and to January 2020. We drop noisy pings and identify movement using distance, time, and speed. Stays are defined as ping sequences without movement. Trips are defined as movement streams that start and end with a stay, with a minimum total distance of 0.25 miles.

We determine device home locations by assigning pings to census blocks. Pings during night hours are scored based on the likelihood of being at home. We label the highest-scoring census block for each device as the home location if it appears on at least 3 nights during the month of our data. Devices without an assigned home location are considered visitors. For devices with a home location, we impute the census tract median household income.

We validate our data in two ways. First, we show that our cellphone data accurately represents travel patterns. To do so, we plot the distributions of the travel time and geodesic distance between the origin and destination, for both cellphone and survey data. Figure A1 presents a high degree of overlap and similarity. Second, Figure A2 shows the share of the tract population covered by the cellphone data. We order tracts by income percentiles. Our coverage is fairly constant around 5% for all percentiles of the income distribution, suggesting that our cellphone location records cover a representative sample of the population in terms of income.

[36] Specifically, our subsample of pings is restricted to those with latitudes between 41.11512 and 42.494693, and longitudes between -88.706994 and -87.52717. This corresponds to the seven counties (Cook, DuPage, Kane, Kendall, Lake, McHenry and Will) of the Chicago Metropolitan Agency for Planning (CMAP) region.



(a) Travel distance (b) Travel time

Figure A1: Representativeness of travel patterns

Notes: This figure plots kernel densities of the distribution of travel distances (Panel a) and travel times (Panel b) using trips in the survey data as well as in the cellphone data. Our level of observation is a trip. Trips in the cellphone data are constructed following the steps in Appendix S1.1. Trips in the survey data do not include walking, biking or multi-modal trips.

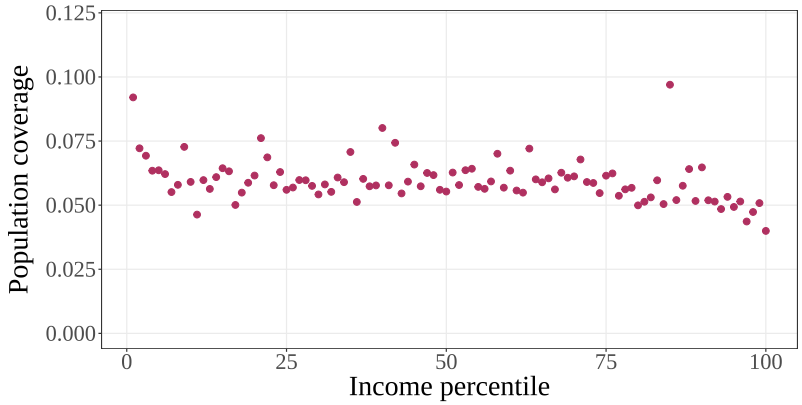


Figure A2: Representativeness across income groups

Notes: This figure plots a binscatter of the fraction of the population in each income percentile covered by the mobile phone data. We define the census-tract specific population coverage as the ratio between (i) the number of cellphones whose home location is assigned to that specific census tract, and (ii) the he number of inhabitants of the census tract according to the 2010 Census data. Income percentiles are defined by the census tract median household income.

B Downtown Surcharge

Effective January 6, 2020, the City of Chicago implemented a new Ground Transportation Tax structure for ride-hailing trips.[37] The Downtown Zone Surcharge applies to any trip that starts or ends within a specified Downtown Zone Area during peak times, which are weekdays between 6 am and 10 pm. For single ride-hailing trips, the tax is $1.25 without a Downtown Zone Surcharge and $3.00 with the surcharge. Before January 6, 2020, the surcharge was $0.72 for all rides, at all times and in all areas.[38] This implies a $0.53 basic increase in single rides and extra charges in the surcharge zone at peak hours of $1.75.

Figure A3: Downtown TNC surcharge area

Notes: This figure shows the downtown surcharge zone. The surcharge of $3 applies to any trip that starts or ends within this zone on weekdays between 6 am and 10 pm.

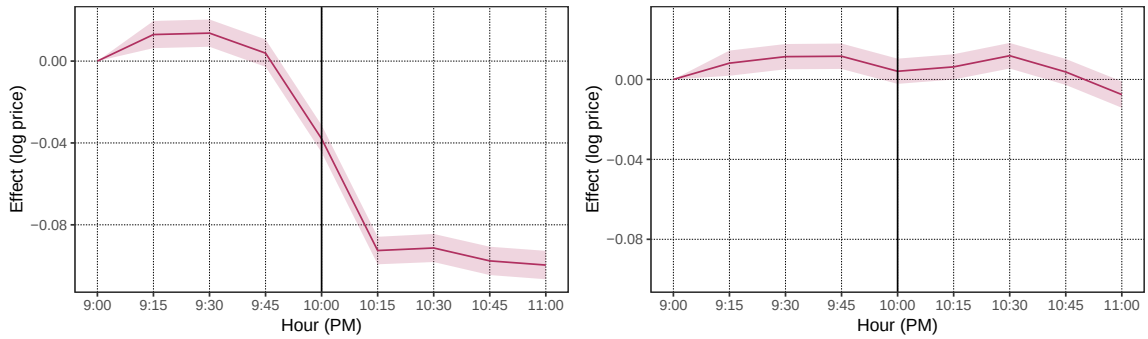
We use the policy to identify the average price elasticity of travelers by comparing trips that originate or end in the zone to those that originate from or end in adjacent, non-treated areas around 10PM, when the surcharge is no longer active. Concretely, our specification is

[37] See on the website of the City of Chicago.
[38] See https://abc7chicago.com/uber-lyft-chicago-congestion-tax-taxes/5818233/

$$y_{o,d,t} = \mu_{o,d} + \alpha_t + \beta_t \cdot treat_{o,d} + \epsilon_{o,d,t},$$

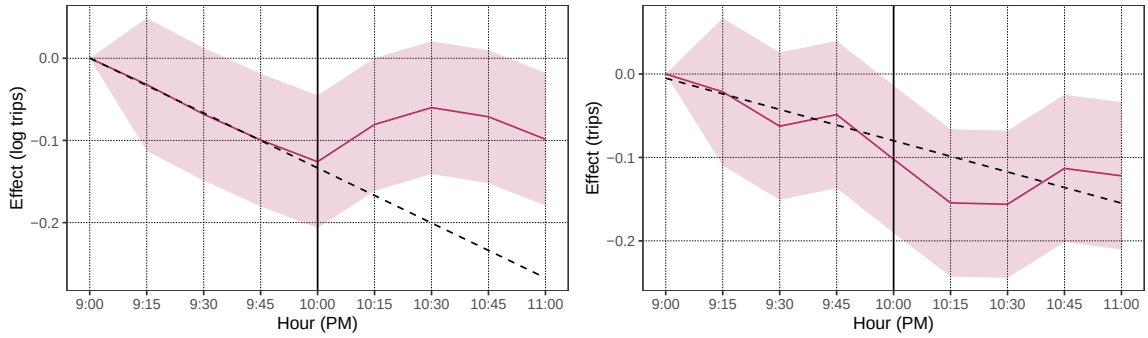
where $y_{o,d,t}$ is either log price or log trips, o, d refers to origin/destination CA. Time t is measured in 15-min intervals. $treat_{o,d}$ refers to all trips $o \to d$ trips subject to surcharge. We plot the coefficients of these treatment effects in Figure A4 and Figure A5. Taking both estimates, we recover an implied price elasticity of -1.42.

Figure A4: Evening price response, 2020 (left) and 2019 (right)



Notes: The top panel shows ride-hailing prices of areas affected by the surcharge of $3 relative to unaffected adjacent areas around 10 pm, after which the surcharge no longer applies. The bottom panel shows the same Figure in 2019, when the surcharge policy was not in place yet.

Figure A5: Evening quantity response, 2020 (left) and 2019 (right)



Notes: The top panel shows how ride hail trips of areas affected by the ride hail surcharge of $3 relative to unaffected adjacent areas around 10 pm, after which the surcharge no longer applies. One can see an increase relative to a downwards trend. The bottom panel shows the same Figure in 2019, when the surcharge policy was not in place yet and we see that the downwards trend continues.

C Proofs and Additional Theoretical Results

We first introduce some new notation. We decompose the derivatives of travel times with respect to fleet sizes as

$$T^k_{jk} = \check{T}^k_{jk} + \tilde{T}^k_{jk}.$$

The first term $\check{T}^k_{jk}$ accounts for effects on waiting times. This term is zero when $k \neq j$. The second term accounts for effects due to travel times.

C.1 Optimality condition for fleet size

Proposition 2. *The first order conditions for the social planner's problem* (4) *with respect to fleet sizes can be written as*

$$\overbrace{u_j^T \check{T}^k_{jj}}^{\textit{Direct benefit of fleet size}} = \overbrace{C^k_j}^{\textit{Mg. cost of fleet size}} + \overbrace{E^k_j}^{\textit{Mg. env. externality of fleet size}} - \overbrace{\sum_l u_l^T \cdot \tilde{T}^k_{lj}}^{\textit{Network effects}} + \overbrace{M^k_j}^{\textit{Diversion}} + \frac{\lambda}{1+\lambda}\left(E^k_j + \underbrace{\sum_k (\tilde{u}^T_k - u^T_k) \cdot T^k_{lj}}_{\textit{Spence distortion}} + \underbrace{\tilde{M}^k_j - M^k_j}_{\textit{Diversion distortion}} \right), \quad (15)$$

where M^k_j and $\tilde{M}^k_j$ are defined as:

$$M^k_j = \sum_l \frac{\partial q_l}{\partial k_j} \left(C^q_l + E^q_l - \sum_m u^T_m \cdot T^q_{ml} - p_l \right)$$

$$\tilde{M}^k_j = \sum_{l \in \mathcal{J}_G} \frac{\partial q_l}{\partial k_j} \left(C^q_l + \sum_{k \in \mathcal{J}_G} q_k \cdot \Omega_{kj} - \sum_m \tilde{u}^T_m \cdot T^q_{ml} - p_l \right).$$

Proof. See Appendix C.2 □

This result takes a very similar form to equation (5). The left hand side is the direct benefit of an increase in the fleet size—on those riders taking that mode—instead of the price (which can be thought of as the direct benefit of an additional

trip). The marginal cost, marginal externality, and network effects terms are almost identical, except that they are derivatives with respect to fleet sizes.

The diversion terms follow a similar intuition to those for equation (5). They are, once again, weighted sums of deviations from Pigouvian prices, but the weights are now given by the increase in mode-l trips caused by a change in k_j. This can be thought of as the mode substitution caused by an increase in mode-j capacity.

Finally, the budget causes two monopoly-like distortions: underweighting the environmental externality and a Spence distortion.

C.2 Proof of Propositions 1 and 2

Proof. The Lagrangian for the social planner's problem is:

$$U(\mathbf{q}, T(\mathbf{q}, \mathbf{k})) - C(\mathbf{q}, \mathbf{k}) - E(\mathbf{q}, \mathbf{k}) - \lambda \left(\sum_{j \in \mathcal{J}_G} [C_j(q_j, k_j) - p_j(\mathbf{q}, T(\mathbf{q}, \mathbf{k}))q_j] - B \right).$$

In this expressions, $\mathbf{q}$ is a function of $(\mathbf{p}, \mathbf{k})$ given by market equilibria.

The first order condition for p_j is:

$$\sum_l \frac{\partial q_l}{\partial p_j} \left[\frac{\partial U}{\partial q_l} + \sum_m u_m^T T_{ml}^q - C_l^q - E_l^q + \lambda \left(p_l + \sum_m q_m \frac{dp_m}{dq_l} - C_l^q \right) \right] = 0. \tag{16}$$

The first order condition for k_j is:

$$\sum_m u_m^T T_{ml}^k - C_l^k - E_l^k + \lambda \left(\sum_m q_m \frac{dp_m}{dk_j} - C_l^k \right) +$$
$$\sum_l \frac{\partial q_l}{\partial k_j} \left[\frac{\partial U}{\partial q_l} + \sum_m u_m^T T_{lk}^q - C_l^q - E_l^q + \lambda \left(p_l + \sum_m q_m \frac{dp_m}{dq_l} - C_l^q \right) \right] = 0. \tag{17}$$

We now show that $\partial U / \partial q_j = p_j$. Let $\partial \Theta_j(p,t)$ be the boundary between $\Theta_j(p,t)$ and $\Theta_0(p,t)$, and let $\partial \Theta_{jk}(p,t)$ be the boundary between $\Theta_j(p,t)$ and $\Theta_k(p,t)$. Gross utility can be written as $U(q,t) = \int_{\Theta_j(q,t)} u(t_j, \theta) f(\theta)\, d\theta$. Using the Leibniz integral

rule, we get that

$$\frac{\partial}{\partial q_j}U(q,t) = \sum_k \int_{\partial\Theta_k(q,t)} u_k(t_k,\theta)e_k(\theta)f(\theta)\,d\theta + \sum_{kl} \int_{\partial\Theta_{kl}(q,t)} u_k(t_j,\theta)e_k(\theta)f(\theta)\,d\theta,$$

(the interior term from the integral rule is zero because t is fixed), where $e_k(\theta)$ denotes by how much $\Theta_k(q,t)$ expands at θ as q_j increases. This also equals:

$$\sum_k \int_{\partial\Theta_k(q,t)} u_k(t_k,\theta)e_k(\theta)f(\theta)\,d\theta + \sum_{k,l>k} \int_{\partial\Theta_{kl}(q,t)} (u_k(t_k,\theta) - u_{lk}(t_l,\theta))e_l(\theta)f(\theta)\,d\theta.$$

Since agents in the boundaries are indifferent between two choices, $u_k(t_k,\theta) = p_k$ for the first sum and $u_k(t_k,\theta) - u_l(t_l,\theta) = p_k - p_l$ for the second sum. After substituting and rearranging terms, our main expression can be written as:

$$\sum_k p_k \left(\int_{\partial\Theta_k(q,t)} e_k(\theta)f(\theta)\,d\theta + \sum_l \int_{\partial\Theta_{kl}(q,t)} e_k(\theta)f(\theta)\,d\theta \right).$$

The term in parentheses is how much $\Theta_k(p,t)$ expands in total into all other regions, so it is equal to $\partial q_k/\partial q_j$. It is thus equal to 1 for j and 0 for $k \neq j$. We can thus conclude that $\partial U(q,t)/\partial q_j = p_j$.

The term $\sum_m q_m dp_m/dq_l$ can simply be written as $\sum_m (q_m \partial p_m/\partial q_l + q_m \cdot \partial p_m/\partial t_m \cdot \partial T_m/\partial q_l) = \sum_m (q_m \Omega_{ml} + \sum_n q_m \cdot \partial p_m/\partial t_n \cdot \partial T_n/\partial q_l)$. Similarly, $\sum_m q_m dp_m/dk_l = \sum_m q_m \cdot \partial p_m/\partial t_m \cdot \partial T_m/\partial q_l$ by a simple application of the chain rule. We now show that $\sum_{k'} q_{k'} \cdot \partial p_{k'}/\partial T_k$ can be written as a weighted average of the change in gross utility among marginal travelers, which we denote by $\tilde{u}_j^T$. Given that definition, we can rewrite

$$\sum_m q_m \frac{dp_m}{dq_l} = \sum_m (q_m \Omega_{ml} + \tilde{u}_m^T T_{ml}^q) \quad \text{and} \quad \sum_m q_m \frac{dp_m}{dk_l} = \sum_m \tilde{u}_m^T T_{ml}^k.$$

First, by Leibniz's integral rule,

$$\frac{\partial q_j}{\partial p_j} = -W_j(p,t) - \sum_{k \neq j} W_{jk}(p,t),$$

where $W_j(p,t) = \int_{\partial\Theta_j(p,t)} v_j(\theta) \cdot \hat{n}_j(p,t,\theta) f(\theta)\, d\theta$ and $W_{jk}(p,t) = \int_{\partial\Theta_{jk}(p,t)} v_{jk}(\theta) \cdot \hat{n}_{jk}(p,t,\theta) f(\theta)\, d\theta$ are integrals over boundaries $\partial\Theta_j(p,t)$ and $\partial\Theta_{jk}(p,t)$, where the integrand is the density of riders that are willing to switch modes in response to an increase in utility. That density is given by the dot product of $v_{jk}(\theta)$, the vector whose elements are the inverse of $\partial u_j/\partial\theta - \partial u_k/\partial\theta$ (and the inverse of $\partial u_j/\partial\theta$ for $v_j(\theta)$), and $\hat{n}_x(p,t,\theta)$, the unit normal component of the boundary $\partial\Theta_x(p,t)$ at θ.

Also by Leibniz's integral rule,

$$\frac{\partial q_j}{\partial t_j} = V_j(p,t) + \sum_{k \neq j} V_{jk}(p,t),$$

where $V_j(p,t) = \int_{\partial\Theta_j(p,t)} \frac{\partial u_j(t,\theta)}{\partial t} v_j(\theta) \cdot \hat{n}_j(p,t,\theta) f(\theta)\, d\theta$ and $V_{jk}(p,t) = \int_{\partial\Theta_{jk}(p,t)} \frac{\partial u_j(t,\theta)}{\partial t} v_{jk}(\theta) \cdot \hat{n}_{jk}(p,t,\theta) f(\theta)\, d\theta$. These are similar integrals as before, only that the integrand is the density of riders that are willing to switch modes in response to an increase in pickup times.

Let $\Lambda(p,t)$ be the matrix whose j-th diagonal element is $W_j(p,t) + \sum_k W_{jk}(p,t)$, and whose non-diagonal element (j,k) is $W_{jk}(p,t)$. Let $\Sigma(p,t)$ be a matrix that is defined similarly, but whose elements arise from $V_{jk}(p,t)$ instead of $W_{jk}(p,t)$. Then, by the implicit function theorem, the matrix of derivatives $\partial p_j/\partial t_k$ is given by

$$\Psi(p,t) = \Lambda^{-1}(p,t)\Sigma(p,t).$$

From the definition of W and V, it is clear that this is a weighted average of $\partial u_j(t,\theta)/\partial t$ over sets of marginal agents. We define

$$\tilde{u}_j^T \equiv \sum_l q_l \frac{\partial p_l}{\partial t_j} = \sum_l q_l \Psi_{lj},$$

the sum of such weighted averages, weighted by the number of agents in each market.

Substituting $\partial U/\partial q_j = p_j$, $\partial p_j/\partial k_l = T_{jl}$, $\sum_m q_m \cdot dp_m/dq_l = \sum_m (q_m \Omega_{ml} + u_m^T T_{ml}^q)$, and $\sum_m q_m \cdot dp_m/dk_l = \sum_m \tilde{u}_m^T T_{ml}^k$ into equations (16) and (17), and then isolating

p_j and $u_j^T T_{jj}^k$ yields expressions (5) and (15). □

C.3 Generalizing Propositions 1 and 2 to multiple markets

Consider a city government that faces many markets m—people going between different geographical areas at different times. The market can still be described as in Section 3.2, where the vectors $\mathbf{q}$, $\mathbf{p}$, and $\mathbf{t}$ represent quantities, prices, and times for all modes j and markets m. The vector of capacities $\mathbf{k}$ can represent the capacities of different bus or train routes at different times. We index it by r.

In this setting, it may not be realistic to think of a government that sets a separate price and frequency for every mode in every market. We therefore consider coarser policy levers, such as the price of buses for the whole city, the price of trains during rush hour, a per km carbon tax for the whole city, the frequency of one bus route, or an overall factor for the frequency with which all trains run.

Consider one such policy lever, which we represent by some parameter σ. The government chooses the level that maximizes its objective function subject to the budget constraint, which can be written as

$$\max_{\sigma} U(\mathbf{q}(\sigma), T(\mathbf{q}(\sigma), \mathbf{k}(\sigma))) - C(\mathbf{q}(\sigma), \mathbf{k}(\sigma)) - E(\mathbf{q}(\sigma), \mathbf{k}(\sigma)) + \lambda \left[\sum_{mj} p_{mj}(\sigma) q_{mj}(\sigma) - C(\mathbf{q}(\sigma), \mathbf{k}(\sigma)) \right], \quad (18)$$

where $\mathbf{q}(\sigma)$ is taken to be the equilibrium vector of trips.

The first-order condition for this Lagrangian can be written out as

$$0 = \sum_{mj} p_{mj} \frac{dq_{mj}}{d\sigma} + \sum_{nkmj} \frac{\partial U}{\partial t_{nk}} \frac{\partial t_{nk}}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} + \sum_{nkr} \frac{\partial U}{\partial t_{nk}} \frac{\partial t_{nk}}{\partial k_r} \frac{dk_r}{d\sigma} - \sum_{mj} \frac{\partial C}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} - \sum_{mj} \frac{\partial E}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} - \sum_{r} \frac{\partial C}{\partial k_r} \frac{dk_r}{d\sigma} - \sum_{r} \frac{\partial E}{\partial k_r} \frac{dk_r}{d\sigma}$$
$$+\lambda \Bigg\{ \sum_{mjnk} q_{nk} \frac{\partial p_{nk}}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} + \sum_{mjnkol} q_{nk} \frac{\partial p_{nk}}{\partial t_{ol}} \frac{\partial t_{ol}}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} + \sum_{mj} p_{mj} \frac{dq_{mj}}{d\sigma} - \sum_{mj} \frac{\partial C}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} - \sum_{r} \frac{\partial C}{\partial k_r} \frac{dk_r}{d\sigma} \Bigg\}. \tag{19}$$

Suppose that σ is a price instrument, in which case $\frac{dk_r}{d\sigma}$ is equal to zero for all r. Then, after some algebra, this first order condition can be written as

$$p_j^\sigma = C_j^\sigma + E_j^\sigma - U_j^{A,\sigma} + M_j^{W,\sigma} + \frac{\lambda}{1+\lambda} \left\{ \mu_j^\sigma - E_j^\sigma - \Delta U_j^\sigma + \Delta M_j^\sigma \right\}, \tag{20}$$

where

$$w_{mj}^\sigma = \frac{\frac{dq_{mj}}{d\sigma}}{\sum_n \frac{dq_{nj}}{d\sigma}} \qquad D_{kj}^\sigma = \frac{\sum_m \frac{dq_{mk}}{d\sigma}}{\sum_m \frac{dq_{mj}}{d\sigma}}$$
$$p_j^\sigma = \sum_m p_{mj} w_{mj}^\sigma \qquad C_j^\sigma = \sum_m \frac{\partial C}{\partial q_{mj}} w_{mj}^\sigma \qquad E_j^\sigma = \sum_m \frac{\partial E}{\partial q_{mj}} w_{mj}^\sigma$$
$$U_j^{A,\sigma} = \sum_{nkm} \frac{\partial U}{\partial t_{nk}} \frac{\partial t_{nk}}{\partial q_{mj}} w_{mj}^\sigma \qquad U_j^{M,J,\sigma} = \sum_{mnkol} q_{nk} \Phi_{nkol}^{J,\sigma} \frac{\partial t_{ol}}{\partial q_{mj}} w_{mj}^\sigma \qquad \mu_j^\sigma = \sum_{nkm} q_{nk} \Omega_{nkmj}^{J,\sigma} w_{mj}^\sigma$$
$$M_j^{W,\sigma} = \sum_{k \neq j} D_{kj}^\sigma (C_k^\sigma + E_k^\sigma - U_k^{A,\sigma} - p_k^\sigma) \qquad M_j^{\Pi,\sigma} = \sum_{k \in J \setminus \{j\}} D_{kj}^\sigma (C_k^\sigma + \mu_k^\sigma - U_k^{M,\sigma} - p_k^\sigma)$$
$$\Delta U_j^\sigma = U_j^{M,J,\sigma} - U_j^{A,\sigma} \qquad \Delta M_j^\sigma = M_j^{\Pi,\sigma} - M_j^{W,\sigma}.$$

This equation resembles very closely equation (5). When generalizing it to this expression, the key insight is that the relevant price, marginal cost, marginal externality, network effects, and diversion ratios are weighted averages of individual-market quantities across markets. The weight given to market m is $w_{mj}^\sigma = \frac{\frac{dq_{mj}}{d\sigma}}{\sum_n \frac{dq_{nj}}{d\sigma}}$: to what extent does a change in σ affect the number of trips in market m.

For a per-km road tax, one can find an explicit expression for the tax. The price faced by travelers taking the taxed mode is given by $p_{mj} = \frac{\partial C}{\partial q_{mj}} + r_{mj}\tau$, where r_{mj} is the trip distance and τ is the per km tax. One can substitute this expression on the above FOC, isolate τ, and do some algebra to write it as:

$$\tau = \frac{1}{r_j^\sigma}\left(E_j^\sigma - U_j^{A,\sigma} + M_j^{W,\sigma}\right), \tag{21}$$

where $r_j^\sigma = \sum_m r_{mj} w_{mj}^\sigma$ is the average distance per trip. This expression takes a standard Pigouvian form, where the optimal price is equal to the average per-km externality plus the average per-km network effects and a misallocation term. This expression does not account for the budget constraint because it is unlikely to be binding after charging a road tax.

If we now consider a policy lever that does affect k, we can rewrite the first-order condition as

$$-\tilde{U}^{A,k,\sigma} = C^{k,\sigma} + E^{k,\sigma} - U^{A,k,\sigma} + M^{W,k,\sigma} + \frac{\lambda}{1+\lambda}\left\{-E^{k,\sigma} - \Delta U^{k,\sigma} + \Delta M^{k,\sigma}\right\}, \tag{22}$$

where

$$C^{k,\sigma} = \sum_r \frac{\partial C}{\partial k_r}\frac{dk_r}{d\sigma} \qquad E^{k,\sigma} = \sum_r \frac{\partial E}{\partial k_r}\frac{dk_r}{d\sigma} \qquad U^{A,k,\sigma} = \sum_{nkr} \frac{\partial U}{\partial t_{nk}}\frac{\partial t_{nk}}{\partial k_r}\frac{dk_r}{d\sigma}$$

$$\Delta q_k^\sigma = \sum_m \frac{dq_{mk}}{d\sigma} \qquad \tilde{U}^{M,k,J,\sigma} + U^{M,k,J,\sigma} = \sum_{nkolr} q_{nk}\Phi_{nkol}^{J,\sigma}\frac{\partial t_{ol}}{\partial k_r}\frac{dk_r}{d\sigma}$$

$$M^{W,\sigma} = \sum_{k,m} \Delta q_k^\sigma (C_k^\sigma + E_k^\sigma - U_k^{A,\sigma} - p_k^\sigma) \qquad M^{\Pi,\sigma} = \sum_{k\in J,m} \Delta q_k^\sigma (C_k^\sigma + \mu_k^\sigma - U_k^{M,\sigma} - p_k^\sigma)$$

$$\Delta U^{k,\sigma} = U_j^{M,k,J,\sigma} - U_j^{A,k,\sigma} \qquad \Delta M^{k,\sigma} = M_j^{\Pi,k,\sigma} - M_j^{W,k,\sigma},$$

and all other terms are defined as before. We decompose the effects of k on times into an effect due to waiting $\check{U}^{M,k,J,\sigma}$ and an effect due to in-vehicle time $U^{M,k,J,\sigma}$.

Once again, this equation resembles equation (15) very closely. Quantities are also aggregated across markets through a weighted average in which the weight given to market m is $w_{mj}^\sigma = \frac{\frac{dq_{mj}}{d\sigma}}{\sum_n \frac{dq_{nj}}{d\sigma}}$.

D Model Details

D.1 Model of Waiting Times for Public Transit

We assume that the time between vehicles follows some distribution with density $\phi(\cdot)$ that has mean $1/k_{mj}$ and variance ω^2/k_{mj}^2. We also assume that travelers arrive to the stop or station at times that are uniformly distributed.

The density of travelers arriving between two subsequent vehicles with a time difference of t is $t \cdot k_{mj} \cdot \phi(t)$: the density $\phi(t)$ is multiplied by $t \cdot k_{mj}$ because the longer the gap between vehicles, the more riders arrive between them. If the time difference is t, a rider arriving between two vehicles needs to wait $t/2$ in expectation. Therefore, the expected waiting time is given by

$$T_{mj}^{wait} = \int \frac{1}{2} t \cdot (t \cdot k_{mj} \cdot \phi(t))\, dt = \frac{1+\omega^2}{2k_{mj}}.$$

D.2 Model of Waiting Times for Ride-hailing and Taxis

Consider mode j (either taxi or ride hailing). Let q_{ahj} be the number of mode-j trips with origin a during hour h, and let I_{ahj} be the number of drivers working for mode j that are idle in location a during this time. We assume that there is a matching technology such that the expected waiting time for riders before their trip starts is given by

$$T_{ahj}^{W} = A_{aj}^{W} I_{ahj}^{-\phi_j}. \tag{23}$$

A_{aj}^{W} is a scale factor that measures the overall matching inefficiency for mode j in location a. The parameter ϕ_j is an elasticity that determines how quickly waiting times decrease with the number of idle drivers. This flexible specification nests simple models of matching in taxi and ride-hailing markets.[39]

To determine the number of idle drivers in every location, we assume that the distribution of drivers across the city arises from a parsimonious model that cap-

[39] In the taxi model in Lagos (2003), for instance, $\phi_j = 1$. In the simplest ride-hailing model described by Castillo et al. (2024), $\phi_j = 1/n$ in n-dimensional space.

tures the spatial dynamics of the market. Let L_{hj} be the total number of drivers working for mode j during hour h. The number of drivers that are busy is given by $B_{hj} = \sum_{od} T^{\text{vehicle}}_{odh} q_{odhj}$, where T^{vehicle}_{odh} are the travel times from the traffic congestion model, and q_{odhj} is the number of people taking mode j from o to d. The total number of idle drivers is given by $I_{hj} = L_{hj} - B_{hj}$.

We assume that the probability that an idle driver is in location a during hour h is given by

$$\frac{\exp(\mu_a + \sum_b B_{ab} F_{hb})}{\sum_{a'} \exp(\mu_{a'} + \sum_b B_{a'b} F_{hb})}, \tag{24}$$

where $F_{ha} = \sum_b (q_{bahj} - q_{abhj})$ represents the net inflow of mode-j trips into location a, $B_{ab} = \lambda r_{ab}^{-\rho}$ is a factor for each pair of locations a and b that decays with the distance r_{ab} between them. This probability takes the form of a multinomial logit model that depends on two terms. First, μ_a, which are fixed effects that capture the fact that drivers tend to work in certain locations of the city. Second, $\sum_b B_{ab} F_b$, which models the extent to which idle drivers are more likely to be located near areas where net inflows are high. The latter term is driven by two opposing forces: a high net inflow of trips induces a high net inflow of drivers, so those areas tend to have many idle drivers; however, these areas have an oversupply of drivers so earnings go down, and drivers will try to move away from them.

Putting all these pieces together, the number of idle drivers in every location is given by

$$I_{ahj} = (L_{hj} - B_{hj}) \frac{\exp(\mu_a + \sum_b B_{ab} F_{hb})}{\sum_{a'} \exp(\mu_{a'} + \sum_b B_{a'b} F_{hb})}. \tag{25}$$

This expression, coupled with equation (23), determines the waiting times for taxis and ride hailing.

Estimation We first estimate the parameters A^W_{ahj} and ϕ_j that map the number of idle drivers into waiting times. Consider CA a. We make the simple assumption that the I_{ahj} available drivers are distributed homogeneously across a and that the pickup time conditional on distance is $t(x) = M_{aj} x^{c_j}$. That implies that the pickup

time has a distribution whose expectation is[40]

$$T^W_{ahj} = M_{aj}\Gamma\left(1+\frac{c_j}{2}\right)\left(\frac{1}{\pi I_{ahj}}\right)^{\frac{c_j}{2}}. \tag{26}$$

This takes the desired form $A^W_{aj} I^{-\phi_j}_{ahj}$, where $A^W_{aj} = M_{aj}\Gamma\left(1+\frac{c_j}{2}\right)\left(\frac{1}{\pi}\right)^{\frac{c_j}{2}}$ and $\phi_j = \frac{c_j}{2}$.

We obtain M_{aj} and c_j from a regression of the log of the travel time on the log of the travel distance for all car trips in our Google Maps dataset originating and ending within the same CA, where we include CA fixed effects. The main coefficient from this regression is $c_j = 0.730$ (s.e.=0.0022), and M_{aj} are the fixed effects that we estimate. Based on those results, we can conclude that $\phi_j = \frac{c_j}{2} = 0.365$, and we back out A^W_{ahj} from the expression above.

We then move on to estimate the parameters of the driver location model (μ_a, λ, and ρ). We do not observe drivers directly, but we use Uber data for the average waiting time at the CA by hour of the week level—i.e., T^W_{ahj}. Inverting equation (26) allows us to compute all values of I_{ahj}. We can then estimate (μ_a, λ, and ρ) by maximum likelihood, based on equation (24). Maximizing this likelihood is not a simple problem since the vector of μ_a has 77 elements. We simplify the task by splitting the problem into an inner loop that computes the optimal vector of μ_a given λ and ρ using a contraction mapping, as in Berry et al. (1995), and an outer loop that maximizes over λ and ρ. Table A1 presents our main estimates.

Table A1: Driver Movement Estimates

	Coefficient	Standard Error
λ	0.0419	0.00007
ρ	-0.1312	0.0101

Notes: Standard errors are computed using a sandwich estimator.

[40] With a density of idle drivers I_{ahj}, the pdf of the distance to the nearest driver is given by $2\pi x I_{ahj} e^{-\pi I_{ahj} x^2}$, a Weibull distribution with parameters $k = 2$ and $\lambda = 1/\pi L$. We integrate the travel time over this density to obtain equation (26).

D.3 In-vehicle time adjustment

Our estimated congestion model predicts in-vehicle times very well for short trips, but it systematically overestimates times for long trips. This is likely because those long trips take highways, so travel times are shorter than the sum of edge-specific travel times that do not take highways.

To correct for this issue, we estimate a linear model of the form

$$\log\left(\frac{T_{mj}^{\text{vehicle}}}{\hat{T}_{mj}^{\text{vehicle}}}\right) = \alpha_j + \beta_j d_m + \epsilon_{mj},$$

where T_{mj}^{vehicle} is the Google Maps in-vehicle time for mode j in market m, $\hat{T}_{mj}^{\text{vehicle}}$ is the time predicted by our model, and d_m is the straight-line distance between the origin and destination for market m.

In our simulations, we use the estimates from this model to scale our predicted travel times by a factor $\exp(\hat{\alpha}_j + \hat{\beta}_j d_m)$.

D.4 Additional Parameters and Assumptions

Marginal costs: We take a marginal cost of \$0.396 per km for all car-based modes (including taxis and ride-hailing) from the AAA cost of driving. Taxis and ride-hailing also incur labor costs of \$10 per hour.

We combine several sources to obtain the marginal costs of public transit. For buses, we take the sum of several elements. First, we use capital costs of \$900,000 per bus that lasts 250,000 miles, which we take from diesel and electric bus purchases made by Chicago Transit Board. Second, we use fuel costs of \$3.26 per gallon with a fuel efficiency of 3.38 mpg, which we take from the National Transit Database (NTD)'s records for the CTA in 2020. Third, we set wages to \$33 per hour, which are also obtained from the NTD, assuming the average driving speed is 20 km/h, times a factor of 2 to account for benefits and the wages of supervisors, schedulers, etc. Finally, we use maintenance costs of \$2.76 per km reported in the NTD. These numbers add up to \$7.528 per km.

We also sum several costs to obtain the marginal costs for trains. First, we use capital costs of \$11M per train that lasts 2 million miles, based on the purchase price of the trains operated currently by the CTA and assuming each train has 10 rail cars. The CTA states that trains last approximately 43 years, make around 15 trips a day, and each trip is approximately 12.1 miles on average, which provides us with our estimate of lifetime mileage. Second, we set energy costs that are twice the fuel costs of a bus. Third, we set energy costs to \$9.06 per km, which we compute as the CTA's energy operating expenses divided by the total mileage of trains. Finally, we set maintenance costs of \$5.00 per km by dividing the total operational expenses by the total miles travelled by trains using CTA's 2020 budget. These numbers add up to \$18.68 per km.

As a sanity check, we compare these numbers relative to the quantities implied by the CTA's financial statements for 2019 values. One challenge is that those statements report operating expenses, some of which are not marginal costs (such as the wages of staff), and they do not account for the cost of capital. Nevertheless, we can obtain upper and lower bounds for marginal costs. These statements imply that the marginal cost of buses is between \$5.17 and \$12.51 per km, and the marginal costs of trains is between \$9.07 and \$40.38 per km. Reassuringly, both ranges include the values we use in our model.

Environmental externalities: For the social cost of carbon, we use the latest EPA proposal of \$190 per tonne as the baseline number.[41] For local pollutants, we obtain estimates based on Holland et al. (2016). We use their findings for Cook County to obtain a cost of 44.93 cents per gallon of gasoline and a cost of 41.32 cents per gallon of diesel fuel. They provide damages incurred due to use of different vehicle types in the United States based on pollutants emitted by vehicle type. We aggregate these values for Cook County, weighing vehicle types by the number of miles travelled. For gasoline-related damages, we restrict the sample to non-truck vehicles, and for diesel-related damages, we use the sample of diesel-only trucks.

[41] See EPA Issues Supplemental Proposal to Reduce Methane and Other Harmful Pollution from Oil and Natural Gas Operations.

Vehicle occupancy: We take the average occupancy of private cars to be 1.5 people, following the estimates for Chicago in Krile et al. (2019), and the average occupancy of ride-hailing and taxi trips to be 1.3 passengers, following Hou et al. (2020).

D.5 Equilibrium computation

Given prices and capacities $(\mathbf{p}, \mathbf{k})$, an equilibrium is a set $(\mathbf{q}, \mathbf{t})$ that satisfies $\mathbf{q} = q(\mathbf{p}, \mathbf{t})$ and $\mathbf{t} = T(\mathbf{q}, \mathbf{k})$, the demand and transportation technology equations. By plugging in the technology equation in the demand equation, the equilibrium condition can alternatively be written as $\mathbf{q} = q(\mathbf{p}, T(\mathbf{q}, \mathbf{k}))$. Thus, if we define the function $f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) = q(\mathbf{p}, T(\mathbf{q}, \mathbf{k}))$, an equilibrium is characterized by a vector of flows $\mathbf{q}^{\mathbf{p},\mathbf{k}}$ that is a fixed point of $f^{\mathbf{p},\mathbf{k}}$. After finding a fixed point, the equilibrium vector of travel times can then be computed as $\mathbf{t}^{\mathbf{p},\mathbf{k}} = T(\mathbf{q}^{\mathbf{p},\mathbf{k}}, \mathbf{k})$.

One naive way to search for an equilibrium is by fixed point iteration. However, this procedure typically diverges. We, instead, find a root of $f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) - \mathbf{q} = 0$ using a limited-memory version of Broyden's method. We use the actual vector of trips in the data as the initial point, and we use an identity matrix as the initial guess for the Jacobian. The full Broyden algorithm is:

Algorithm 1 Equilibrium computation using Broyden's method

Set initial value of trips $\mathbf{q}$.
Compute initial times $\mathbf{t} = T(\mathbf{q}, \mathbf{k})$.
Compute deviation $\mathbf{d} = q(\mathbf{p}, \mathbf{t}) - \mathbf{q}$.
Set new vector of trips $\mathbf{q}' = \mathbf{q} + \gamma\mathbf{d}$ for a small step size $\gamma > 0$.
Compute new vector of times $\mathbf{t}' = T(\mathbf{q}', \mathbf{k})$.
Compute deviation $\mathbf{d}' = q(\mathbf{p}, \mathbf{t}') - \mathbf{q}'$.
Set initial approximation to inverse Jacobian $\mathbf{A} = \mathbb{1}$.
while $||\mathbf{d}'|| >$ *tolerance* **do**
    Define differences $\Delta\mathbf{q} = \mathbf{q}' - \mathbf{q}$ and $\Delta\mathbf{d} = \mathbf{d}' - \mathbf{d}$.
    Update vectors of trips $\mathbf{q} = \mathbf{q}'$ and deviation $\mathbf{d} = \mathbf{d}'$.
    Compute new approximation to inverse Jacobian $\mathbf{A} = \mathbf{A} + \frac{\Delta\mathbf{q} - \mathbf{A}\Delta\mathbf{d}}{\Delta\mathbf{q}^T\mathbf{A}\Delta\mathbf{d}}\Delta\mathbf{q}^T\mathbf{A}$.
    Compute new vector of trips $\mathbf{q}' = \mathbf{q} - \mathbf{A}\mathbf{d}$.
    Compute new vector of times $\mathbf{t}' = T(\mathbf{q}', \mathbf{k})$.
    Compute new deviation $\mathbf{d}' = q(\mathbf{p}, \mathbf{t}') - \mathbf{q}'$.
end

We make two adjustments to the above algorithm. First, we compute the approximation to the inverse Jacobian $\mathbf{A}$ with the limited-memory approach in Byrd et al. (1994). Second, when we compute the new vector $\mathbf{q}'$, we often obtain an infeasible vector of trips (the number of ride-hailing or taxi drivers is not enough to satisfy demand). Whenever that is the case, we iteratively update $\mathbf{q}' = \mathbf{q} + 1/2(\mathbf{q}' - \mathbf{q})$ until we get back to a feasible value.

D.6 Optimization

Having computed an equilibrium as described in Appendix D.5, we can compute welfare $W(\mathbf{p}, \mathbf{t})$ and the net revenue of the city $\Pi(\mathbf{p}, \mathbf{t})$. The unconstrained welfare maximization problem is

$$\max_{\mathbf{p},\mathbf{t}} W(\mathbf{p}, \mathbf{t}). \tag{27}$$

We solve this problem in two steps. First, we approximate the solution with a Nelder-Mead optimizer, starting from the true prices and capacities, and stopping after 100 iterations. Second, we run a quasi-Newton method starting from the Nelder-Mead optimum. This method differs from Newton's method in two ways, both of which greatly reduce the computational cost of our procedure. First, to avoid computing the Hessian of the objective function, we use the BFGS approximation (Nocedal and Wright, 2006), which only requires computing the gradient. Second, we approximate the gradient with central differences. Every time we compute a finite difference, instead of fully running Broyden's method until convergence to an equilibrium, we only take a few steps (typically three) starting from the central point, which allows us to obtain a good approximation to the gradient at a small fraction of the computational cost.

With a budget constraint, the welfare maximization problem is

$$\max_{\mathbf{p},\mathbf{t}} W(\mathbf{p}, \mathbf{t}) \qquad \text{s.t.} \qquad \Pi(\mathbf{p}, \mathbf{t}) = -B, \tag{28}$$

where B is the city's transportation budget. To solve this problem, we use the augmented Lagrangian method. We iteratively solve the following approximation

to the Lagrangian:

$$\max_{\mathbf{p},\mathbf{t}} W(\mathbf{p},\mathbf{t}) - \lambda_n \left(\Pi(\mathbf{p},\mathbf{t}) + B\right) + \mu_n \left(\Pi(\mathbf{p},\mathbf{t}) + B\right)^2. \tag{29}$$

We initialize this iterative procedure by setting $\mu_0 = 10^{-6}$ and $\lambda_0 = 0$. In every step n we use the method we described above to maximize the objective function, and we set $\mu_{n+1} = 2\mu_n$ and $\lambda_{n+1} = \lambda_n + \mu_n(\Pi^n + B)$, where Π^n is the net revenue at the n-th step optimum. In this algorithm, λ_n converges to the Lagrange multiplier that results in the budget constraint being satisfied with equality (Nocedal and Wright, 2006). This means that (29) converges to the true Lagrangian plus an extra penalty for deviations from the budget constraint—and thus, the sequence of solutions converge to the solution of (28).

E Model Fit

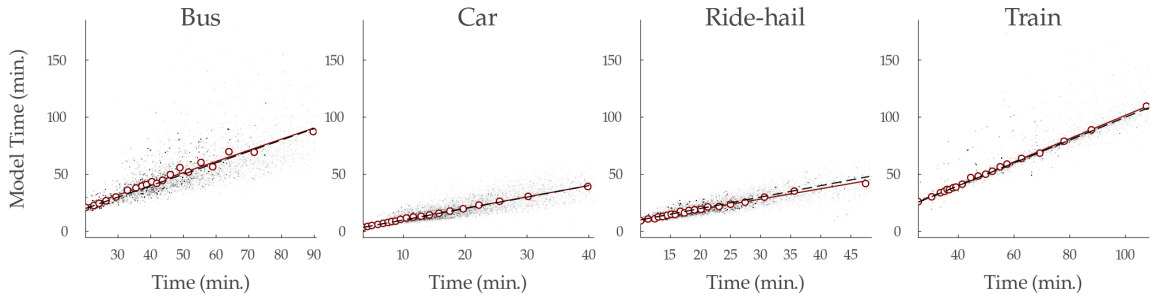
Figure A6 shows that the trip times and market shares by mode from our model fit the data well.

F Sensitivity Analysis

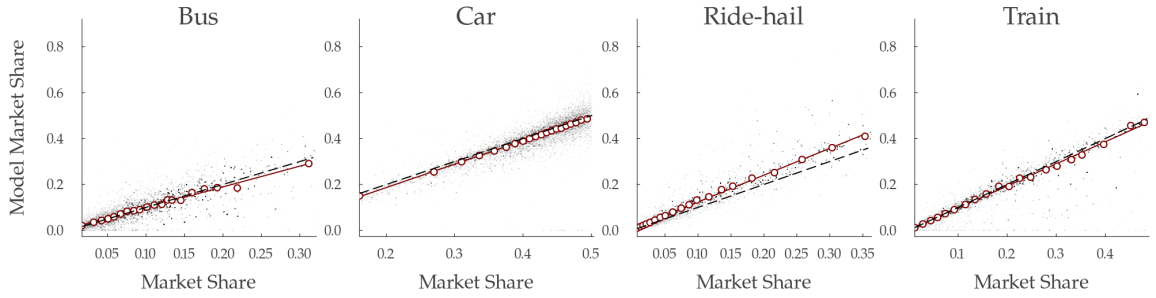
Figure A7 shows the extent to which our main results are sensitive to some of the key parameters of our model, focusing on the *Transit + Road Pricing* counterfactual. Each panel shows how a 10% increase in several parameters of the model affect the five choice variables of the city government. In the first two panels, we see that the result indicating public transit prices should be close to zero is very robust: optimal prices are always within 1.2 cents of our baseline results. On the other hand, our results about optimal wait time for public transit are more sensitive to parameters, as can be seen in the third and fourth panels. For five of the six parameters (the marginal cost of public transit, the price and time sensitivity of travelers, the relative disutility of walking and waiting, and the variability of bus

arrivals), a 10% change in the value of the parameter results in changes to the optimal bus wait time on the order of 0.6 minutes and in the optimal train wait time on the order of 0.2 minutes. These changes correspond to changes in frequencies of around 5%. Finally, the last panel shows that the road price is also quite robust: in every case, the optimal value is within 1.5 cents per km of the baseline value.

While estimates of the social cost of carbon continue to be revised Carleton and Greenstone (2022), Figure A7 shows that the only result on which it has an important impact is the optimal road price. If instead of the latest recommendation from the EPA ($190 per tonne of CO_2) we used the previous EPA figure of $51 per tonne, the optimal road price would drop by around 9 cents to $0.25 per km.



(a) Trip Times



(b) Market Shares

Figure A6: Model fit of trip times and market shares by mode

Notes: This figure compares observed trips times and market shares to model trip times and market shares separately for each mode. Each panel displays both a binscatter and a scatterplot for a sample of 25,000 markets, where markets are drawn randomly with replacement and sample weights are given by trip counts. The dashed line shows the 45 degree line.

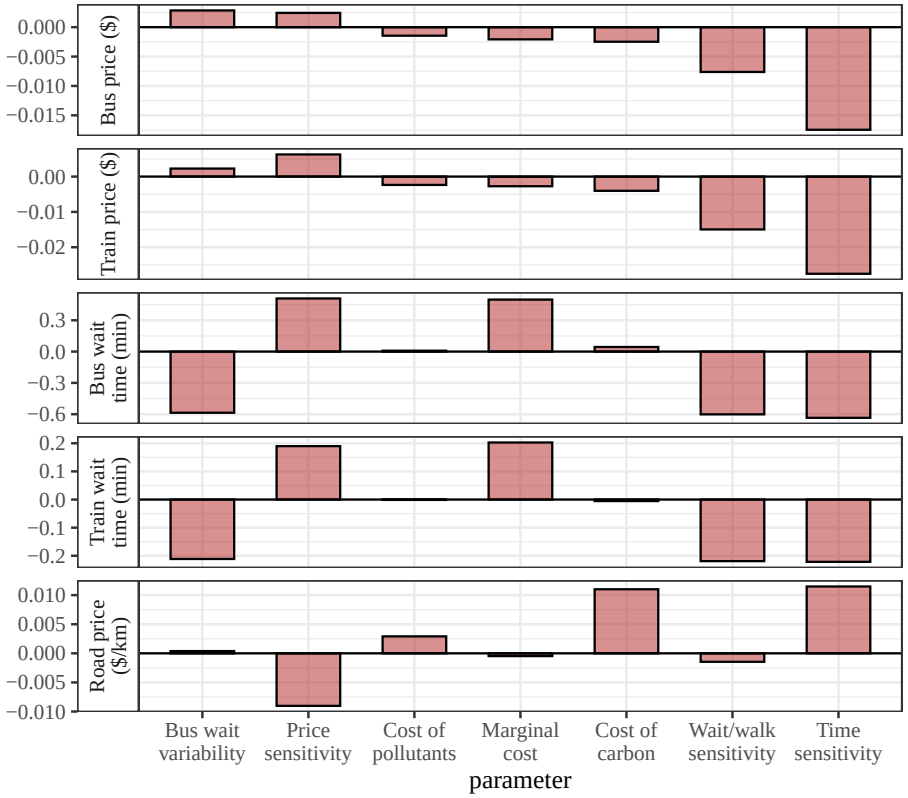


Figure A7: Robustness of counterfactual results

Notes: This figure presents how the choice variables of the social planner change in response to changes in some of the model parameters. We focus on the transit + road pricing counterfactual. In each panel, we show how a 10% increase in the model parameter specified in the x-axis affects the choice variable in the x-axis. Within each panel, we order model parameters from the one that causes the largest increase to the one that causes the largest decrease.

G Additional Results

G.1 Demand Robustness

To assess the robustness of our demand estimation results, we estimate a number of additional specifications. Results are shown in Table A2. We first relax the assumption that travelers have utility that is linear in time by including the square of time. The estimated coefficient on the square of time is $-.312$, implying that the disutility of travel time is increasing in the length of the trip. In particular, the marginal disutility of the first minute is about half as much as the marginal disu-

tility of the 60th minute.[42] Measured at the average trip length, the average VOT and time elasticity are both lower than in our main estimates. The average price elasticity is similar. It follows from equation 5 that counterfactuals using this alternative specification would therefore lead to larger frequency reductions and higher congestion prices (to a first order). Therefore, our results should be interpreted as conservative lower bounds on frequency reductions and road prices.

Next, we allow for travellers to not only care about average travel times, but also about reliability, in particular for public transit. To do so, we include the standard deviation of travel time for public transit modes (train and bus). We find that travelers are relatively insensitive to at least this measure of reliability, and our estimated coefficients imply a similar average VOT, price elasticity, and time elasticity as in our main specification.

In our third robustness specification, we additionally allow for heterogeneity in time sensitivity by income. We again adopt a Box-Cox functional form: $\alpha_T^i = \alpha_T + \frac{\alpha_{Ty}}{y_i^{1-\lambda_T}}$.[43] The estimated coefficients imply a similar average VOT, price sensitivity, and time sensitivity as in our main results. However, the dispersion in VOT is compressed because we estimate that low-income individuals who are more price elastic are also more time elastic. While this would mute the dispersion in distributional consequences that we estimate in our counterfactual results, the results would remain qualitatively unchanged since lower-income individuals still exhibit significantly lower VOT than higher-income individuals.

Finally, we allow for more flexible fixed effects by including a mode-destination fixed effect. This fixed-effect controls for additional unobserved factors that vary at the mode-destination level, including factors such as varying parking costs. We find that once again the estimated average VOT, price elasticity, and time elasticity are similar to our main specification, suggesting such factors are not biasing our estimation.

[42] Note that for estimation time is measured in hours.

[43] We also include an additional set of instruments that interacts free-flow times with indicators for each income quintile.

Table A2: Demand Estimation Robustness

	(1)	(2)	(3)	(4)
α_T	-0.574	-1.702	-1.286	-1.929
	(0.048)	(0.025)	(0.028)	(0.018)
α_p	-2.169	-3.080	-1.173	-1.000
	(0.115)	(0.158)	(0.032)	(0.041)
α_{py}	-0.508	-0.643	-0.089	-0.022
	(0.026)	(0.026)	(0.014)	(0.022)
ρ	0.359	0.314	0.335	0.191
	(0.012)	(0.015)	(0.009)	(0.010)
α_{T^2}	-0.312	·	·	·
	(0.014)			
$\alpha_{std(T)}$	·	-0.114	·	·
		(0.046)		
α_{Ty}	·	·	-25.611	·
			(2.060)	
λ_T	·	·	-1.457	·
			(0.073)	
Mode FEs	✓	✓	✓	
Mode-Destination FEs				✓
Market FEs	✓	✓	✓	✓
Transfer & Multimodal Controls	✓	✓	✓	✓
Policy Moment	✓	✓	✓	✓
Car Ownership	✓	✓	✓	✓
Nest	✓	✓	✓	✓
Avg. VOT	8.30	13.00	10.82	12.78
VOT (Bot. Quintile)	2.01	2.68	8.98	5.15
VOT (Top Quintile)	17.82	28.86	15.89	22.59
Avg. Price Elast.	-0.64	-0.64	-0.70	-0.63
Avg. Time Elast.	-0.76	-1.14	-1.25	-1.21
M	91,561	74,512	91,561	91,561
N	280,185	222,142	280,185	280,185

Notes: This table presents a number of robustness checks for our main specification in section 4.1. The average VOT is computed by first computing the within market average VOT as the weighted average of α_T/α_p^i and then averaging across markets, with weights given by market size. Similarly, the average elasticities are computed as the weighted average of own-price and own-time elasticities across all mode-market observations, with weights given by market size. In specification (2), markets for which we cannot compute the standard deviation of time are dropped.

G.2 Decomposition of train prices and waiting times

Figure follows the expressions in Section C.3 to decompose the optimal prices and wait times for trains, similar to Figure 9.

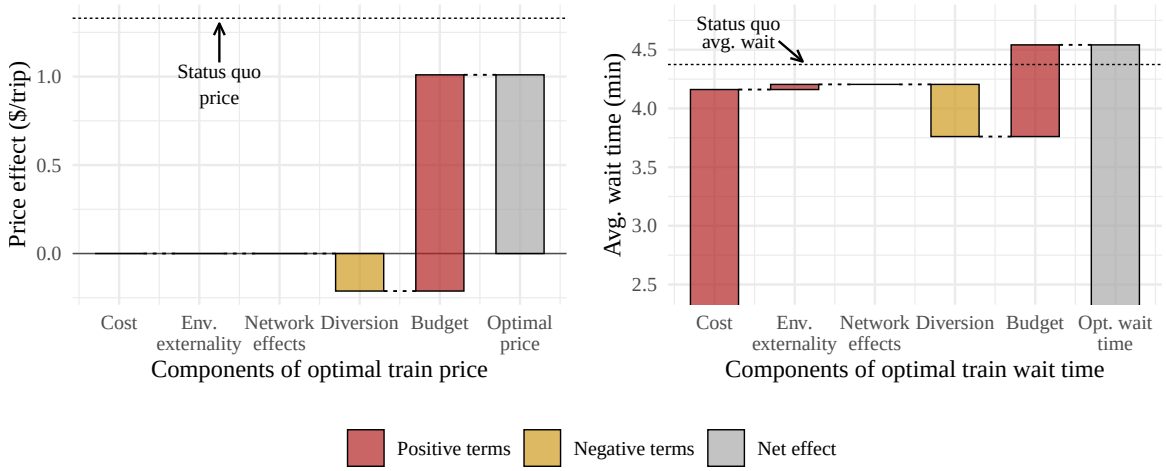


Figure A8: Decomposition of optimal price and waiting times for trains

Notes: This graph shows the a decomposition of the optimal prices and travel times for buses corresponding to our theoretical decomposition in Section 4. Red bars indicate terms that lead prices and travel times to be higher and yellow bars indicate terms that lead prices to be lower.

Additional References

Byrd, R.H., Nocedal, J. and Schnabel, R.B. (1994). Representations of quasi-newton matrices and their use in limited memory methods. *Mathematical Programming* 63(1):129–156.

Carleton, T. and Greenstone, M. (2022). A guide to updating the us government's social cost of carbon. *Review of Environmental Economics and Policy* 16(2):196–218.

Castillo, J.C., Knoepfle, D. and Weyl, G. (2024). Surge pricing solves the wild goose chase. Working paper.

Hou, Y., Garikapati, V., Weigl, D., Henao, A., Moniot, M. and Sperling, J. (2020). Factors influencing willingness to share in ride-hailing trips. Tech. rep., National Renewable Energy Lab (NREL).

Krile, R., Landgraf, A. and Slone, E. (2019). Developing vehicle occupancy factors and percent of non-single occupancy vehicle travel. Tech. Rep. FHWA-PL-18-020, Federal Highway Administration (FHWA).

Supplementary Appendix

S1 Data Construction

S1.1 Cellphone location records

This subsection details how we construct our sample of trips based on the raw cellphone data. The raw data is composed of a sequence of pings. Each ping contains a timestamp, latitude, longitude, and a device identifier. The final output from this process is a dataset with a fraction of the universe of trips that took place in Chicago. A sequence of filtering steps leaves us with 5% of devices. We verify that the owners of these devices are representative and then scale up the number of trips by a factor such that the aggregate number of car trips is consistent with what is reported by the Chicago Metropolitan Agency for Planning (CMAP) 2019 Household Travel Survey.[44]

Data filtering We start by subsetting cellphone pings to a rectangle around the city of Chicago (i.e., latitude between 41.11512 and 42.494693, longitude between -88.706994 and -87.527174) for the month of January 2020.

Next, using the cellphone device identifier, the timestamp and geolocation of each ping, we calculate the time between two consecutive pings as well as the geodesic distance. These distances allow us to obtain the speed between consecutive pings. We then filter out "noisy" pings by using distance, time, and speed variables. In particular, we remove pings that are moving at an excessive speed since these pings are likely to be GPS "jumps" resulting from noise in the measurement of the GPS coordinates of the device.[45] We also drop "isolated" pings since they are not helpful for identifying whether people are moving. Additionally, we only keep pings belonging to a "stream" of pings.[46] We define a stream

[44] Source: My Daily Travel survey (website)
[45] 40 meters per second, i.e. about 145 kilometers per hour
[46] In particular, we only keep pings that satisfy the following two conditions: (i) no more than ten

of pings as a sequence of pings for the same cellphone identifier such that a ping always has another ping within the next 15 minutes and within 1,000 meters. We drop streams with less than 3 pings. Finally, we aggregate pings to the minute of the day by taking the average location and timestamp across pings within each minute for a given cellphone identifier. In what follows, we focus on the remaining filtered pings aggregated at the minute level.

Defining movements, stays, and trips We identify two consecutive (aggregated) pings as a "movement" for a given cellphone identifier if their distance is at least 50 meters or if their implied speed is at least 3 meters per second (6.7 miles per hour or 10.8 kilometers per hour). We then define a "stay" as a sequence of two or more successive pings with no movement.

Finally, we take all streams of pings and define trips as being a stream (i) with movement, (ii) that starts with a stay, and (iii) that ends with a stay. We remove all trips with a total geodesic trip distance between the starting and ending point below 0.25 miles (about 400 meters).

Estimation of home locations and traveler's income This subsection details how we assign a home location and an income level to each individual cellphone identifier.

We start by assigning all cellphone pings to census blocks for the subset of pings within Chicago during our sample period.[47] Next, we focus on pings during night hours, defined as between 10pm and 8am, when individuals are more likely to be at home.

Using this subset of pings, we attribute a score system for each hour between 10pm and 8am. Specifically, regardless of the number of pings, scores are assigned as follows:

- A value of 10 to all census blocks that were pinged between 1 am and 5 am.

minutes to either the next or the previous ping, (ii) no more than 5,000 meters to either the next or the previous ping.

[47] See Appendix S1.1 for the sample restrictions.

- A value of 5 to all census blocks that were pinged between 11 pm and 1am or between 5 am and 7 am.

- A value of 2 to all census blocks that were pinged between 10pm and 11pm, or between 7am and 8am.

The basic idea is to assign a higher score to blocks where the cellphone owner is more likely to be at home. Finally, we sum the scores across all census blocks for each cellphone ID - month combination and keep the census block with highest score. If this highest-score census block appears on at least 3 or more separate nights during the month, we assign it as the cellphone's home census block for that month. Otherwise, we consider the cellphone as having an unknown home location, which we believe captures occasional Chicago visitors such as tourists. Throughout the text, we refer to these devices as *visitors*. Figure S9 plots the share of visitors by origin locations. We see that, for trips done by visitors, the most common origin locations are the city center (center right), both airports (top left and center left), as well as Hyde Park the neighborhood home to the University of Chicago (right, south of the center).

For all cellphones with an assigned home location, we impute their income by using the census tract median household income.[48]

Next, for each market, we estimate traveler's income distribution.[49] First, we take median income by tracts and divide tracts according to quintiles.[50] Next, we assign an income quintile to each device according to their home location. Since we can follow how devices travel across space and over time, for each market, we can measure the quintile from each traveler departing from its destination. Finally, for each market, we construct shares of traveler's income quintile. For markets with less than 5 trips, we impute market-level income shares using the underlying distribution of census tract-level income for the origin CA of that market.

[48] We compute the census-tract median income percentile using the 2010 Census data.
[49] Recall, a market is defined as an (origin CA, destination CA, hour of the week)-tuple.
[50] For 2010, income quintiles are defined using the following cut-offs: $34,875$, $46,261$, $60,590$ and $85,762$ U.S. Dollars.

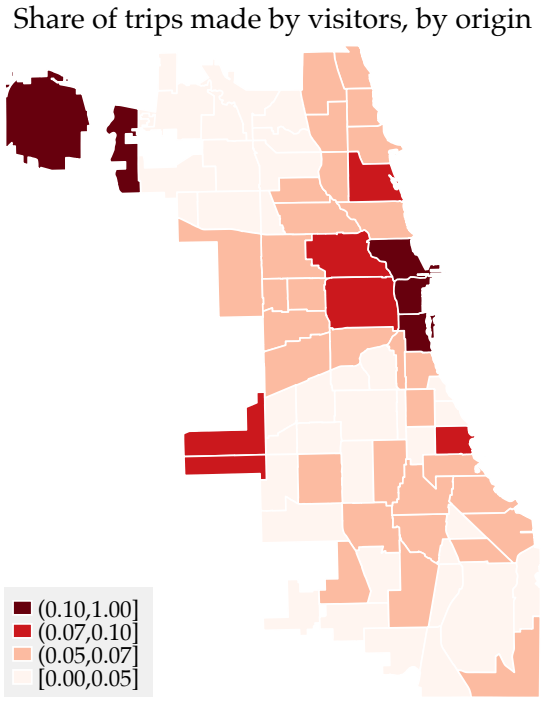


Figure S9: Share of visitors by origin location

Notes: This figure shows the share of trips at the origin CA level made by visitors. In our cellphone trips data, each market (origin-destination-hour triple) has a share of trips made by visitors. To construct the shares displayed in the figure, we take the weighted average of the share of trips made by visitors across destinations and hours of the week, for each origin CA, using inside market size (number of cellphone trips per market) as weight.

S1.1.1 Survey Data Sparsity

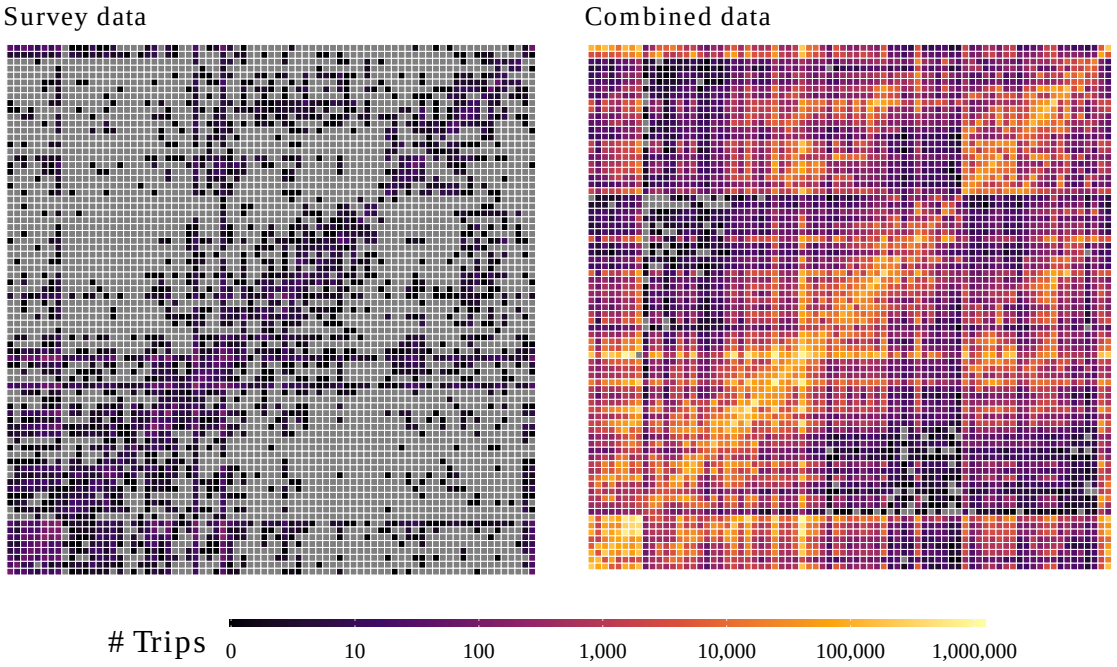


Figure S10: Combined vs. Survey Data: Flows Across Community Areas

Notes: These figures show the number of trips from every origin CA to every destination CA in our combined data (right panel) and in the survey data (left panel). Each row represents an origin CA and each column represents a destination CA. Grey points represent empty cells.

S1.2 Travel times, routes, and schedules

Travel times and routes Similar to Akbar et al. (2023), we query and geocode trips using Google Maps. For each mode of transportation, we query 30,796,848 counterfactual trips and obtain their distance, duration, and route.[51] Importantly, we can measure trip duration for the same origin-destination tuple over the time of the weekday (or weekend) and how this varies with traffic conditions. Moreover, using the detailed "steps" of the public transit Google Maps queries, we obtain

[51] One trip for each (origin census tract, destination census tract, hour of day, weekend dummy) combination. We use all the 801 Chicago census tracts boundaries for the year 2010 from the Chicago Data City portal website.

walk times from the origin latitude/longitude to the "best" train or bus station.[52]

We also obtain Google Maps data on train trip times by querying Google Maps three times for each pair of train stations in Chicago. These times represented three broad time categories: weekday peak, weekday non-peak, and weekend. In particular, the first query requested a trip time of 8am on Wednesday July 6th, 2022, the second query requested a trip time of 11am on Wednesday July 6th, 2022, and the third query requested a trip time of 11am on Saturday July 9th 2022.

Public transit schedules We obtain historical GTFS data from Open Mobility Data. These data contain bus and train schedules for December 2019 through February 2020.

S1.3 Constructing Mode-Specific Trips

Mode-specific trips are constructed using five main sources: (1) Taxi and TNP trips data from the City of Chicago, (2) Google Maps data, (3) cellphone trips data, (4) historical GTFS data containing public transit route schedules, and (5) Chicago public transit data from the MIT Transit Lab and the CTA.

Taxi and Transportation Network Provider (TNP) data We obtain trip times, distances, and origin-destination census tracts for both Taxi and Transportation Network Provider (TNP) trips from the City of Chicago's Data Portal.[53]

Cellphone trips data We construct cellphone trips from cellphone pings using the procedure detailed in Appendix S1.1. This procedure results in a trip-level dataset. Since our cellphone data only captures a portion of the total trips, we adjust for this by assigning an inflation factor to each trip. To account for varying rates of unobserved trips across different city areas, we allow inflation factors to

[52] The "best" bus or train station is not necessarily the closest one, depending on the destination and/or the time of the day.

[53] For privacy reasons, during periods of the day and for locations with very few trips, only the origin and/or destination CA of a trip is reported. See this page for a discussion of the approach to privacy in this data set.

vary by the neighborhood of the trip's origin.[54] Specifically, we calibrate these factors to ensure that the number of car trips beginning in each neighborhood in our dataset matches the corresponding number in the Chicago Metropolitan Agency for Planning (CMAP) Household Travel Survey.[55]

Public transit data We obtain individual public transit trips for the city of Chicago via a partnership between the MIT Transit Lab and the CTA. Each observation corresponds to a a passenger swiping in to access the bus or the train station. For buses, we observe the specific bus stop, bus line, and boarding time. For trains, we observe the station and swiping time. Drop-off locations are given to us and imputed following Zhao et al. (2007).

This data notably excludes trips taken via the Metra, which is a suburban rail system operating in and around Chicago. Metra is managed by a different agency, the Regional Transportation Authority. An additional limitation is that we do not observe trips paid for via cash or trips whose destination could not be imputed. To account for these sources of missing trips, we assign each observed trip an inflation factor. This inflation factor is computed at the day-mode level such that

$$infl_{dm} T_{dm} = R_{dm},$$

where dm indexes the day-mode, T is the total number of observed trips, and R is the observed aggregate daily ridership for the CTA, which we obtain from the City of Chicago's Data Portal. The average such inflation factor is 2.0.

We also do not observe travel times for train trips, and so we are forced to impute these travel times. To do so, we first match each train trip to the historical GTFS schedule data. To compute the match for a given train trip, we first find all scheduled trips between the origin and destination stops of that trip. We then take the match to be the scheduled trip whose boarding time is closest to the observed

[54] Each neighborhood is a group of about 8-9 CAs. The exact make-up of neighborhoods can be found on Wikipedia.
[55] Source: My Daily Travel survey (website)

boarding time. We then take the scheduled travel time as the travel time. This matching process enables us to compute travel times for close to 90% of train trips.

For trips that have no matches in the schedule data, we impute travel times using Google Maps data.[56] In particular, we first assign each trip one of three time categorizations: weekend (if Saturday or Sunday), peak weekday (if between 5-9:59am or 2-6:59pm on a weekday), or non-peak weekday (otherwise). We then take the time to be the travel time of the matching train trip from the Google Maps data.

We also compute travel distances for each trip. We use the Haversine formula to compute distances, with radius equal to 6371.0088, which is the mean radius of Earth in km. For bus trips, we compute the travel distances as the Manhattan distance between the boarding and alighting coordinates, while for train trips we compute the travel distances as the Euclidean distance between the boarding and alighting coordinates.

S1.4 Market Share Calculations

We first append together the transit, TNP, taxi, and cellphone trips data. We incorporate walk times to bus/train stations from the Google Maps data. We drop any trips that have a negative trip time, trip time exceeding 6 hours, negative prices, or missing values for origin, destination, distance, duration, mode, trip time, or price. Since our trip data is at the vehicle level, we account for unobserved vehicle occupancy by scaling trip numbers and prices using the average vehicle occupancy for that mode, which we report in Appendix D.4.

We calculate market shares at the (origin CA, destination CA, hour-of-the-week) level using a two-step process. First, we aggregate trips at the (origin CA, destination CA, hour-of-the-week, date) level. We then let the number of car trips be the residual after subtracting public transit, taxi, TNP, and shared trips from the cell-

[56] Manual inspection suggests these trips typically involve an unobserved transfer between two lines.

phone trips.[57] Car prices are computed as 0.6374 U.S. Dollars per trip mile, which is AAA's estimate of per mile driving costs for an average 2020 model.[58] Finally, we obtain trip counts at the (origin CA, destination CA, hour-of-the-week, date) level by averaging across dates.

S1.5 Market Size

To compute market shares, we need to take a stance on the size of the market, which captures how many people could be traveling at a given moment in time. For simplicity, we assume that market sizes are proportional to the total number of observed trips. To determine the factor of proportionality, we compare the population of each CA to the total number of trips originating from that CA in the morning hours (5-9:59am) on weekdays. The median ratio across CAs is 2.61. Implicitly, this factor assumes that the number of potential travelers in each CA in these morning hours is given by the total population, which is likely an upper bound. We also compute a more conservative factor by assuming the set of potential travelers is made up of commuters and school-age children, which gives a median factor of 1.48. Corresponding to roughly the midpoint of these two factors, we set our proportionality factor to 2.

We restrict ourselves to markets where we observe car trips so that cars are always an available mode. These markets capture 96% of observed trips.

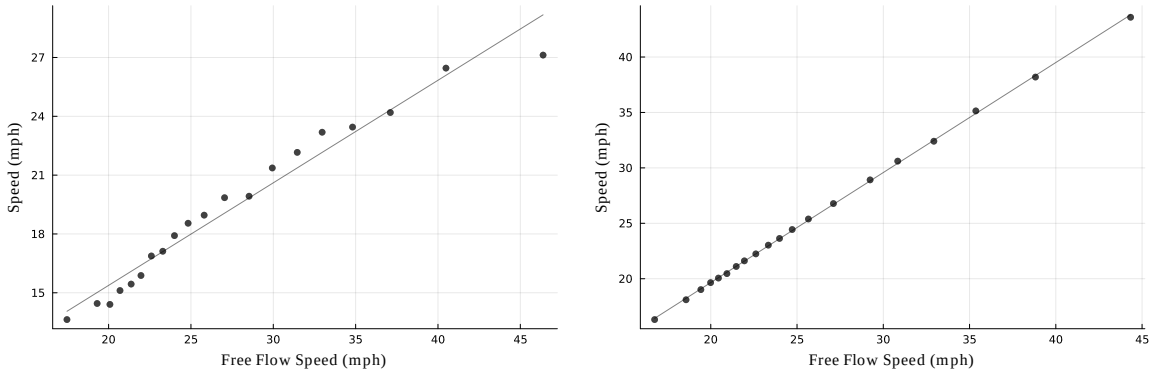
S2 Additional Empirical Results

S2.1 First-Stage Coefficients

Figure S11 shows a visualization of the first-stage for our free-flow instrument. The left panel shows free-flow speeds relative to actual travel speeds at 8am for cars. While actual travel speeds are typically lower than free-flow speeds due to conges-

[57] If the residual is negative we assume that there are no car trips.
[58] Source: AAA brochure "Your driving costs".

tion, there is still a large and positive correlation between the two. That is, markets with high free-flow speeds still have relatively higher actual travel speeds even when there is congestion. The right panel additionally shows free-flow speeds relative to actual travel speeds at 12am for cars. Once again, there is a strong and positive correlation between the two. Moreover, the fact that these two speeds are very close provides a measure of validation for our free-flow speeds, as we would expect there to be very little congestion at this time in most markets. Finally, table S3 shows the first-stage coefficients for the rest of our instruments.



(a) 8am (b) 12am

Notes: This figure shows a binscatter of car free-flow speeds vs. travel speeds across markets at 8am (left panel) and 12am (right panel).

Figure S11: Free-Flow Speeds vs. Actual Travel Speeds

Table S3: First-Stage Coefficients

	Time	Price
	(1)	(2)
Free-Flow Time	0.970***	-9.228***
	(0.007)	(0.090)
Non-TNP Price	0.004***	-0.632***
	(0.001)	(0.018)
Frac. Transfers	0.280***	-0.306***
	(0.003)	(0.020)
Frac. Multimodal	0.166***	0.597***
	(0.004)	(0.029)
Local Diff. x TNP Indic.	-0.014***	-1.445***
	(0.001)	(0.016)
Quad. Diff. x TNP Indic.	0.013***	1.946***
	(0.004)	(0.081)
Local Diff.	-0.024***	-0.336***
	(0.002)	(0.013)
Quad. Diff.	0.055***	3.063***
	(0.006)	(0.072)
π^1 x Non-TNP Price	-0.020***	0.691***
	(0.001)	(0.021)
π^2 x Non-TNP Price	-0.011***	0.308***
	(0.001)	(0.025)
π^3 x Non-TNP Price	-0.013***	0.234***
	(0.001)	(0.024)
π^4 x Non-TNP Price	-0.006***	0.076**
	(0.001)	(0.028)
Mode Fixed Effects	Yes	Yes
Market Fixed Effects	Yes	Yes
F	9,074.944	5,190.973
N	273,833	273,833

Notes: This table presents the first-stage coefficients for the instruments used to estimate demand in section 4.1. In particular, we regress times and prices on the full vector of instruments as well as mode and market fixed effects. Singleton observations (markets with only a single mode) are dropped.

S2.2 Bus Utilization

While our model does not consider capacity constraints for buses when solving for the optimal policy, we can consider *ex-post* the extent to which this constraint might bind. Our results imply frequency reductions for buses that are typically less than

30%. We consider whether these frequency reductions would result in binding capacity constraints, holding ridership levels fixed, by computing the fraction of buses that exceed 70% and 80% utilization across hours of the day. Figure S12 shows that this constraint is unlikely to make a first-order impact on our results as only 10% of buses reach even 70% utilization, and only during the morning and afternoon rush hours.

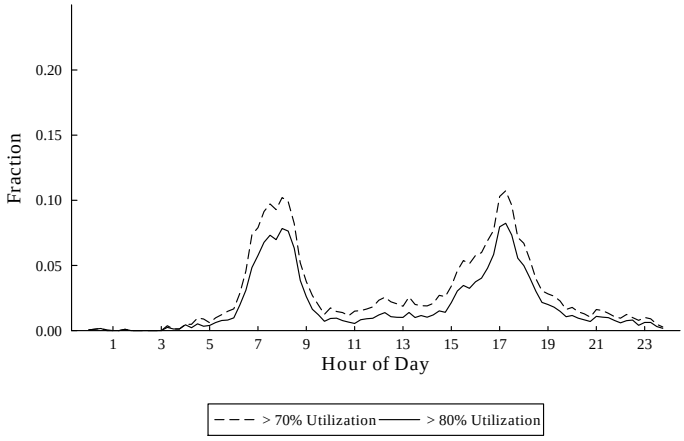


Figure S12: Bus Capacity

Notes: This figure shows the fraction of buses that exceed 80% (solid) and 70% (dashed) utilization over the course of the day.

S3 Additional Counterfactual Results

S3.1 Decomposition of welfare effects

In this section, we decompose the change in consumer surplus and in environmental externalities attributed to different channels. The change in consumer surplus is a product of two forces: the direct change in prices and the indirect effect in time due to changes in mode choices. The change in the environmental externalities is also due to two channels: the change in frequencies and the change in travelers' mode choices (substitution). Table S4 shows how each of these channels contribute to the overall aggregate effects across different scenarios.

Focusing on the counterfactual where the planner only sets public transit prices and frequencies subject to a budget constraint, column 3, we see that consumers face two opposing effects. On the one hand, lower prices means an increase in consumer surplus of $3.8M per week. On the other hand, lower frequencies increase the overall travel times and, in turn, decreases consumer surplus by $3.4M per week. In terms of externalities, most of the reduction accrues through the reduction in frequencies and fewer vehicles running throughout the city.

When the planner only set road pricing, we see a large reduction in consumer surplus of $25.5M per week. The reason is that consumers face an increase of prices for the most common mode of transportation, namely private cars. Because, due to this increase in prices, consumers stop traveling by car, traveling speed goes up, which translates into lower overall travel times and an increase in consumer surplus of $2M per week.

Simultaneously setting public transit prices and frequencies as well as road pricing can be viewed as the combination of the previous two cases. However, in this case we have two opposing effects for both prices and travel times that net each other out in the aggregate overall results.

Finally, when the planner sets all prices and reduces ride-hailing prices by 45%, we see some interactions of these policies accruing through two channels. First, consumer surplus increases by $14.5M per week relative to the previous scenario. However, as travelers start substituting toward ride-hail, travel speed decreases and overall travel time increases, which partially undoes the effect of congestion of car taxes.

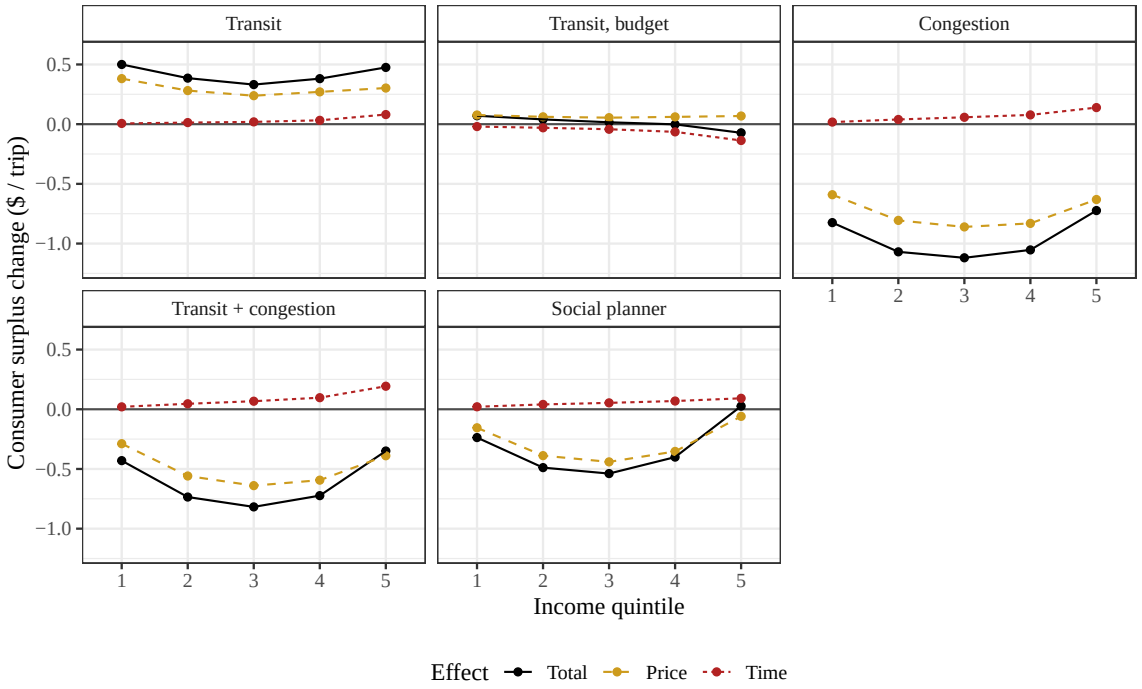
Next, we zoom in on how consumer surplus changes across the income distribution. The results, in percentage terms, can be seen in Figure S13. Observe that the absolute effect of prices for the public transit counterfactuals is larger for lower income consumers, as they are the ones who are more likely to use those modes of transportation. Conversely, the effect of time is more pronounced for higher income consumers, as they are the ones with the highest VOT.

Table S4: Decomposition of Consumer Surplus and Environmental Externalities

		Status quo	Transit	Transit, budget	Road pricing	Transit + Road pricing	Social planner
Δ CS ($M/week)	Total	0	12.647	0.025	-29.113	-18.536	-9.472
	Price	0	11.375	2.476	-31.843	-22.062	-11.093
	Time	0	1.272	-2.451	2.730	3.526	1.621
	Capacity	0	0.436	-2.480	0	0.429	0.799
	Substitution	0	0.837	0.028	2.730	3.097	0.822
Δ Externality ($M/week)	Total	0	-0.616	-0.347	-3.585	-3.692	-3.086
	Capacity	0	-0.038	-0.220	0	-0.032	-0.010
	Substitution	0	-0.578	-0.127	-3.585	-3.660	-3.076
Δ Avg. Speed (km/h)		0.00%	0.61%	0.09%	2.93%	3.14%	2.53%

Notes: This table represent the change in consumer surplus and environmental externalities attributed to different channels. Changes in consumer surplus (first row) are divided into changes in prices (second row) and times (third row). Changes in times are a product in changes in fleet size (fourth row) and substitution of consumers across modes (fifth row). Total changes in externalities (sixth row) are decomposed into changes in fleet size (seventh row) and substitution across consumer (eighth row).

Figure S13: Decomposition of consumer surplus through different channels



Notes: These graphs presents changes in consumer surplus across income quintiles for four different counterfactual scenarios scenarios. Each of the lines represent the change in consumer surplus from each of the channels that affect traveler's utility.