NBER WORKING PAPER SERIES

OPTIMAL URBAN TRANSPORTATION POLICY:
EVIDENCE FROM CHICAGO

Milena Almagro
Felipe Barbieri
Juan Camilo Castillo
Nathaniel G. Hickok
Tobias Salz

Working Paper 32185
http://www.nber.org/papers/w32185

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2024, Revised October 2025

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 insightful 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, Junbiao Chen, Diego Gentile Passaro, Shreya Mathur, Juan Andrés Russy, and Stefano Sperti 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Optimal Urban Transportation Policy: Evidence from Chicago
Milena Almagro, Felipe Barbieri, Juan Camilo Castillo, Nathaniel G. Hickok, and Tobias Salz
NBER Working Paper No. 32185
March 2024, Revised October 2025
JEL No. L0, L97, R0, R4

ABSTRACT

We characterize and quantify optimal urban transportation policies in the presence of congestion and environmental externalities. A municipal government sets public transit policies—fares and frequencies—and road prices to maximize welfare. 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 increased bus and train frequencies. Combining transit policies with road pricing slackens the budget constraint, allowing for even higher transit frequencies and lower prices, thereby increasing consumer surplus after rebates.

Milena Almagro
University of Chicago
Booth School of Business
and NBER
Milena.almagro@chicagobooth.edu

Felipe Barbieri
Dartmouth College
Tuck School of Business
Economics
felipe.barbieri@tuck.dartmouth.edu

Juan Camilo Castillo
University of Pennsylvania
Department of Economics
and NBER
jccast@upenn.edu

Nathaniel G. Hickok
Massachusetts Institute of Technology
nhickok@mit.edu

Tobias Salz
Massachusetts Institute of Technology
Department of Economics
and NBER
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 reduce congestion, environmental impact, and inequality have 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 on private road use. For example, London enacted a £15 cordon tax during the daytime and New York recently approved a cordon tax in Manhattan below 60th Street.[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 determining this optimal mix requires understanding how transportation modes interact with each other. In addition to mode substitution on the demand side and technological interactions through road congestion, we highlight a third channel: budget

[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.

constraints introduce important fiscal interactions across modes. Because building new transit infrastructure in the US is notoriously difficult, (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. Our analysis holds 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. These results emphasize the importance of coordinated policies across modes and provide an efficiency rationale for the London and New York plans to use road pricing to cross-subsidize public transit.

[7] For road pricing, Herzog (2024) finds that endogenizing sorting and traffic congestion attenuates welfare effects by around 20%, whereas Barwick et al. (2024) and Hierons (2024) find that it increases them by 18% and 10%, respectively.

Next, we apply this framework to Chicago, an ideal setting for our purposes. Both public and private transportation play an important role in this city. It has large economic disparities, so measuring the effects of transportation policies across different income levels is important. Furthermore, Chicago offers unusually rich data. We combine several data sources to construct a high-resolution dataset of travel flows, travel times, and prices for all relevant modes. We observe nearly all 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 is that there are no official car trip records. To overcome this problem, we estimate total trips from cellphone-location data, and then infer car trips as the residual after subtracting public-transit, ride-hailing, and taxi trips.

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 allow 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. Car operating costs and public transit prices are invariant to demand shocks (and we therefore directly use them as included instruments), but road travel times and ride-hailing prices are potentially endogeneous.

To instrument travel times, we use the straight-line distance between the origin and destination divided by mode-specific citywide average free-flow speeds. The resulting variation is due to the interaction between geography and mode technology, but independent of demand shocks and infrastructure. To address the endogeneity of ride-hailing prices, we exploit price variation from a surcharge on downtown ride-hailing trips during peak hours.[8] Our estimates reveal substantial heterogeneity in the value of time across travelers, ranging from $8 to $57 per hour for travelers in the bottom and the top income quintile, respectively.[9]

[8] Leccese (2022) studies the pass-through of this policy.

[9] For comparison, the average hourly wage in Chicago's metropolitan area in 2020 is $29.01. See Wage statistics from Bureau of Labor Statistics for the Chicago region.

We then estimate the road traffic congestion technology at a high resolution. 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 (2024). We find congestion elasticities between 0.09 and 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 scheduled frequency and the reliability of those schedules.

We simulate three main counterfactuals. First, we separately explore scenarios in which the planner 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 planner controls both.

When considering public transit policies alone, welfare gains strongly depend on whether the planner faces a budget constraint. Without the constraint, the planner would reduce fares and increase service frequencies, generating consumer surplus gains of $26.87 million per week and, after accounting for the fiscal burden, a net welfare gain of $5.2 million per week. However, once budget constraints are imposed, optimal frequencies increase more modestly while fares must rise to address the budget shortfall. Welfare rises by only $0.62 million per week, driven by consumer surplus gains of $0.95 million and partially offset by a $0.24 million increase in environmental externalities.

Road pricing, when used in isolation, allows for much larger welfare gains, but at the expense of travelers. The optimal road tax is 36 cents per kilometer. This leads to overall welfare gains of $3.80 million per week. The majority of these gains come from reducing the externalities generated by private cars, which account for 69.9% of trips. Without rebates, consumer surplus would decrease by $32.32 million per week, with middle-income consumers experiencing the greatest losses due to their reliance on cars. Even if the planner were to fully rebate resulting revenues, consumer surplus would fall by $0.39 million per week.

Combining road pricing and public transit policies unlocks even larger welfare gains. The planner collects substantial revenue from road pricing, which allows it to set transit policies optimally without budget considerations while still directly targeting the externalities from private cars. Public transit becomes virtually free, and bus and train frequencies increase by more than a third relative to the status quo. This combined approach increases overall welfare by \$6.97 million per week—nearly twice the sum of welfare gains from implementing transit policies and road pricing independently. Unlike under road pricing alone, consumer welfare rises if the planner rebates its surplus, benefiting both high- and low-income residents while leaving middle-income residents worse off.

These findings underscore the importance of coordinating transit policy with road pricing, as these policies are complementary. Road pricing can achieve substantial reductions in externalities, but travelers benefit only when toll revenues are used first to subsidize transit, with any remaining funds rebated to consumers.

We also investigate policies in which price and frequency vary by location and time. We find little extra gain from more granular pricing. However, we find large gains from granular route frequency adjustments that redirect available capacity to 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 based on 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. Building on these foundations, recent empirical work has studied the New York taxi market (Frechette et al., 2019; Buchholz, 2021), the dry bulk shipping industry (Brancaccio et al., 2020), and ride-hailing platforms (Castillo, 2025; Rosaia, 2025; Gaineddenova, 2022; Buchholz et al., 2025). Kreindler (2024) 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. (2024) 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 bus route network. While our policy simulations keep routes fixed and only varies their frequencies, 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. (2024) 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]

We also build on a classic theoretical literature in transportation economics. Early papers have focused on the interaction of 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. As in Mohring (1972), increasing utilization leads to lower per-passenger average cost of public transit provision.

[10] Several papers investigate alternative margins of adjustment in response to transportation policy (Tsivanidis, 2023; Fajgelbaum and Schaal, 2020; Severen, 2023; Herzog, 2024; Brinkman and Lin, 2022; Allen and Arkolakis, 2022; Bordeu, 2023). We depart from their work by allowing for rich demand substitution patterns across modes and heterogeneity, which is crucial for understanding the distributional effects of transit policies.

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, 2024) and different forms of road pricing (Hall, 2018; Cook and Li, 2025; 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 (CTA), is one of the largest in the nation. It includes a bus network of 127 routes, and a train rapid transit system—the "Chicago L"—that has eight routes and 145 stations. Full fares for buses and trains are \$2.25 and \$2.50, respectively.[11] The CTA has a history of budget shortfalls, making it important to account for 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.

[11] Reduced fares exist for students and seniors. There are also daily, 3-day, weekly, and monthly passes. See CTA Fares for additional details. The public transit prices that we use in the demand model are the average paid fare, which accounts for discounts.

[12] See, for instance, CTA avoids service cuts, fare hikes under proposed \$1.8 billion budget.

[13] See Chicago Taxi Fare Regulation.

2.2 Data description

We define each origin-destination-hour combination as a market. 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, resulting in 48 time periods. Our main dataset consists of travel flows, prices, and travel times for every mode in every market during January 2020.[14]

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.[15] 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).[16]

The second data source, published by the City of Chicago, contains the universe of de-identified taxi and ride-hailing trips.[17] It includes prices, pickup and dropoff locations as well as trip length and duration.

Third, we use 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.[18] We infer all motorized trips from the sequence of GPS coordinates for each individual device (details in Appendix B). 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. This restriction results in a sample of trips and travelers that is representative

[14] Our sample is drawn from Chicago's winter months. Although winter conditions persist for much of the year, travel behavior could shift in warmer weather. In section F.1 we conduct extensive robustness checks of the counterfactual results to changing parameter estimates, including the sensitivity to waiting and walking which is likely higher in the winter.

[15] 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)—are not included in our data because Metra is not managed by the CTA.

[16] We observe transit card identifiers, which allows us to identify chained trips. For such trips, our model accounts for the sum of walking and waiting from all segments. We code chained trips with train and bus segments as train trips.

[17] Source: Chicago Data Portal, Transportation Network Providers - Trips

[18] Source: Veraset: Location Data Provider

across many dimensions, as we show at the end of this section. We thus multiply the number of 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, bus, and train. While we observe the number of trips for buses, trains, taxis, and ride-hailing in the CTA data, we do not have official records of car trips. We recover them by subtracting public transit, taxi, and ride-hailing trips from the cellphone trips, which covers all motorized trips. Since we only see motorized trips in our data, we treat walking, biking, and not traveling as the outside option. We compute the potential market size by comparing the number of morning commuters to the number of residents (see Appendix B), which results in a number of potential commuters that is twice the number of trips we observe. Finally, we query Google Maps data to obtain travel times and routes by market for all modes, including those not chosen by travelers.

Our dataset has two advantages over survey data. First, survey data lack representativeness and their coverage diminishes at high spatial and temporal resolutions, leading to sparse data.[19] Our granular data allow us to estimate the relationship between vehicle flow and traffic speed throughout the city. Second, our data allow us to invert market shares (Berry, 1994; Berry et al., 1995) at a granular level and, thus, construct instrumental variable moment conditions that address the endogeneity of prices and travel times.

We add demographic information 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, and we then assign to each device the median income of that tract.[20] In our estimation, we divide

[19] In the CMAP 2019 Household Travel Survey, over 60% of origin-destination CA pairs have zero trips, and more than that when the data are broken down by time period.

[20] 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.

the population into income quintiles, whose income levels are the median income within that quintile. To compute the income distribution in each market, we count the frequency of every quintile and the corresponding assigned income.

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 B for more details.

2.3 Descriptive results

We present descriptive evidence in four parts. First, we explore the characteristics and usage of different modes. Next, we analyze how riders' income correlates with mode choices. We then document evidence of the low utilization of buses. Finally, we present empirical patterns of traffic congestion in the raw data.

Although Chicago has one of the most extensive public transit systems in the US, about 69.9% of trips are taken by car. Public transit accounts for 22.9% of trips, with buses taking slightly more than half of this share. Ride-hailing accounts for 6.3% of trips. Taxis represent just 0.9% of trips, so we omit them from the 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.

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 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

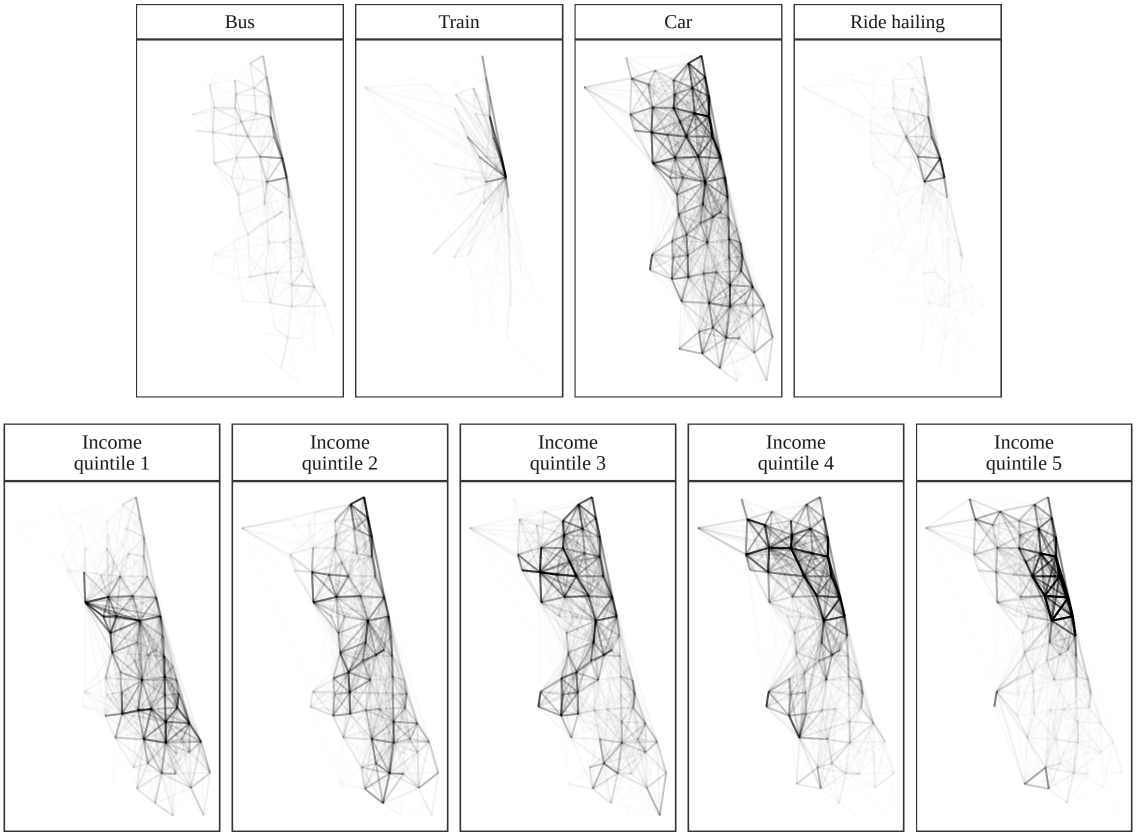


Figure 1: Trips by mode and by income

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.

average, are almost twice as fast as buses. The right panel shows that choosing a mode typically involves a tradeoff between prices and speed: faster modes tend to be more expensive, with the exception of cars.

Figure 3 shows car ownership across income levels. Ownership first increases with income but then flattens at the top of the income distribution. We account for car ownership in our demand estimation to avoid conflating references for non-car modes with travelers' ability to travel by car.

Figure 4 shows how mode choices vary by origin CA income. Lower income travelers are slightly more likely to use buses, while higher income people favor

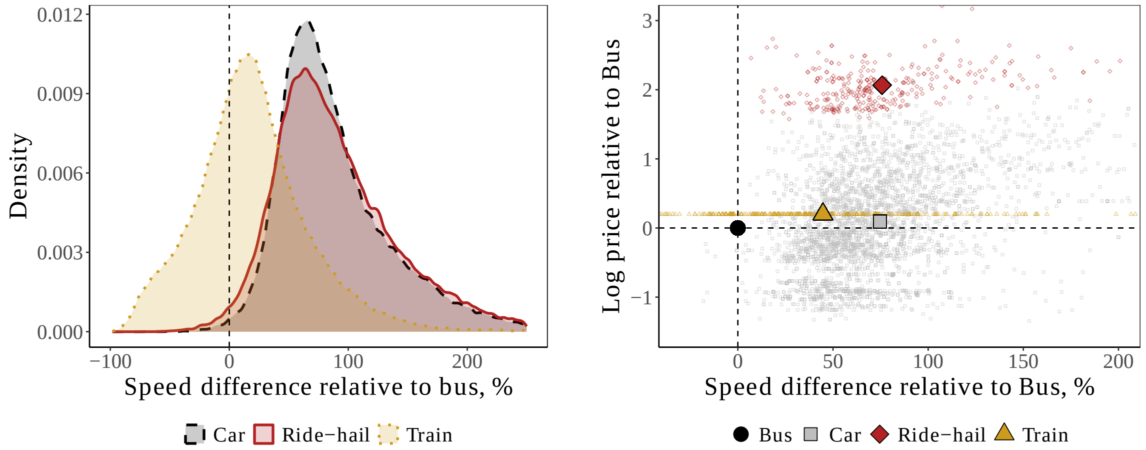


Figure 2: Speed and price differences across modes of transportation

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.

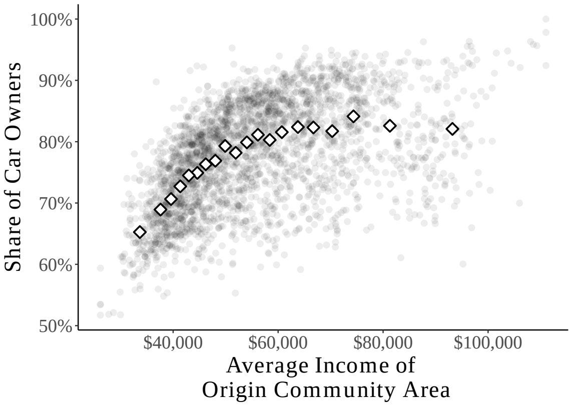


Figure 3: Car ownership by travelers’ income

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

trains, which are concentrated in affluent neighborhoods (see Figure 1). Car usage follows a subtle 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.

We now show that, although buses are the cheapest mode, they generally travel

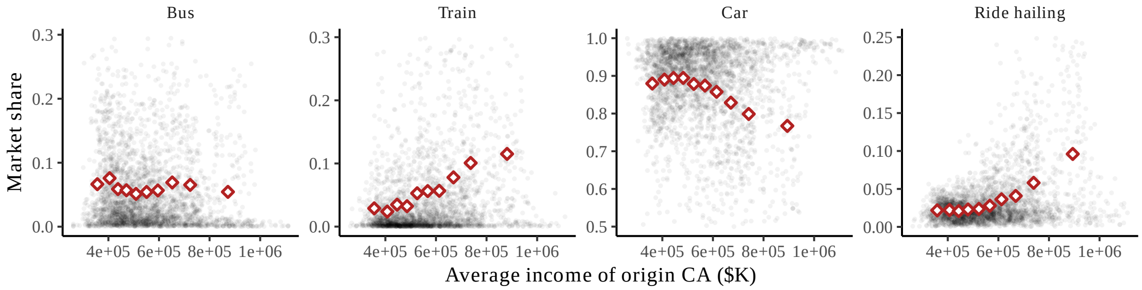


Figure 4: Mode market shares by travelers' income

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.

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 and a zero marginal cost of an additional passenger, an important observation to understand our counterfactual results.

Lastly, we present raw data patterns that show how traffic congestion impacts travel times. The left panel of Figure 6 plots travel times against vehicle flows between adjacent CAs, after residualizing on CA-pair fixed effects. The data exhibit a "hockey-stick" pattern: travel times are flat at lower vehicle flows but rise almost log-linearly beyond a certain point, suggesting that additional vehicles reduce travel speeds with an approximately constant elasticity. This motivates the empirical model of our congestion technology. The right panel shows that this pattern holds when we zoom into specific markets.

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

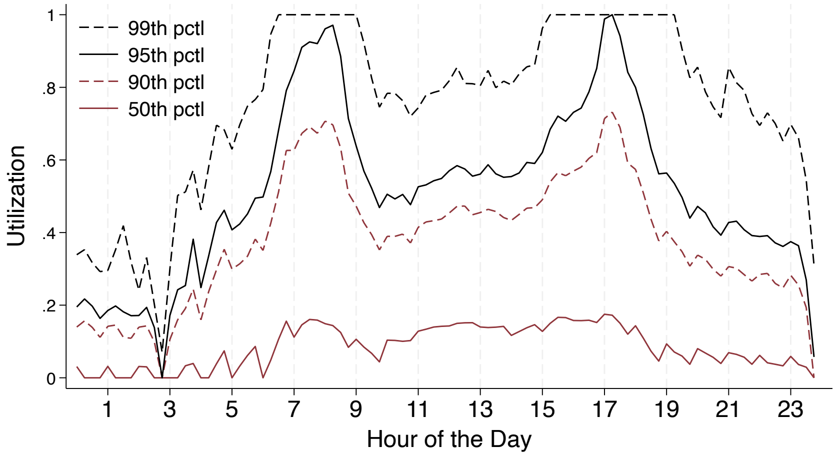


Figure 5: Bus utilization rates

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.

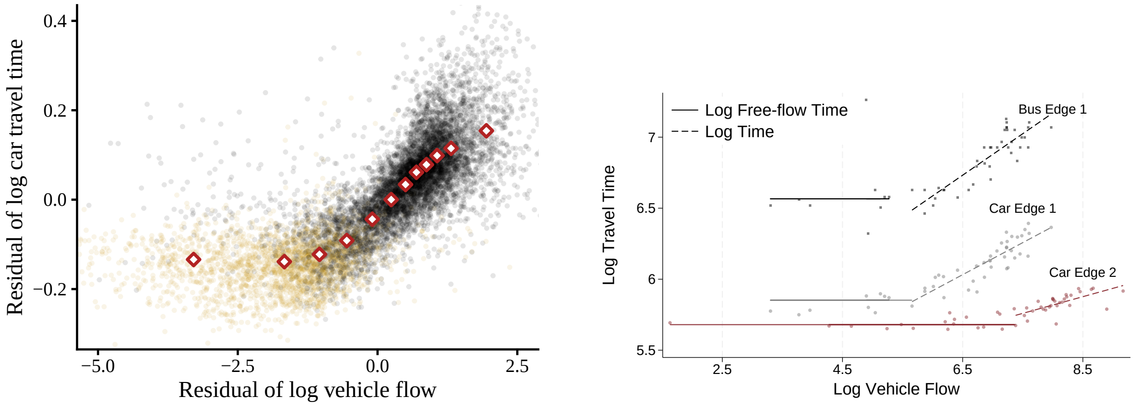


Figure 6: Relationship between flow of vehicles and travel times

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 (edges) 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. Edge is an origin-destination community area pair.

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

There is a mass of travelers with density $f(\cdot)$ who differ in their preferences for modes and whether they own a car, captured by type $\theta \in \mathbb{R}^n$.

A traveler 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. We normalize the utility of the outside option to zero. The traveler chooses the mode in her choice set that maximizes utility:

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

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. We assume that demand is invertible in $\mathbf{p}$, a property that is satisfied by standard discrete choice models, and we denote its inverse by $p(\mathbf{q}, \mathbf{t})$.

Gross consumer utility and consumer surplus are given by

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

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 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* $\mathbf{q}^*$ *and travel times* $\mathbf{t}^*$ *such that* (2) *and* (3) *hold.*

In this model, for any given fleet size and prices, travel times adjust to clear the market. In Appendix E.5, we show the existence of an equilibrium using Brouwer's fixed point theorem. We also show that the equilibrium is unique. This follows from the presence of congestion forces—stabilizing forces that spread out agents across space and travel modes—and the absence of agglomeration effects, which could generate coordination games with multiple equilibria.

3.2 The Social Planner's Problem

The city government's goal is to maximize welfare subject to a budget constraint. Its choice variables are the prices and fleet sizes of buses and trains as well as the road tax. Let $\mathcal{J}_G$ denote the set of modes the government controls—which varies across counterfactuals—with prices $\mathbf{p}_G$ and fleet sizes $\mathbf{k}_G$.

We now define welfare and the government's budget as functions of the allocation $(\mathbf{q}, \mathbf{k})$, exploiting the inverse demand function $p(\mathbf{q}, \mathbf{t})$. 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 the sum of consumer surplus, the government's revenue, and the profit of private mode operators minus externalities. After canceling out transfers, welfare can be expressed more succinctly as gross consumer utility minus the cost of transportation provision and externalities:[21]

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

The government's optimization problem is thus:

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

Note that $\mathbf{q}^*$ is an equilibrium quantity, which changes with $(\mathbf{p}, \mathbf{k})$. We omit its arguments for simplicity. The government maximizes welfare, subject to its budget constraint, by selecting the prices and fleet sizes of the modes it controls. However, it does not control all modes: it cannot set the price of ride-hailing and, in scenarios without road pricing, it cannot set the price of driving. As a result, the government is not able to implement all possible allocations $(\mathbf{q}, \mathbf{k})$. Because the budget enters

[21] This is the case because $W = CS + \Pi + \sum_{j \notin \mathcal{J}_G}(p_j q_j - C_j) - E = (U - \sum_j p_j q_j) + \sum_j (p_j q_j - C_j) - E$.

(4) as an inequality constraint, the planner is not forced to spend money inefficiently, which means that each dollar spent on public transit will generate at least one dollar of welfare.

To derive optimality conditions, we introduce superscript notation for derivatives of costs, externalities, travel times, and utilities with respect to some quantity x. For example, C_j^q denotes the derivative of the cost with respect to the number of rides q of mode j and T_{kj}^q denotes the change in travel time of mode k with respect to additional trip of mode j. Also, let Ω_{lj} represent elements of the inverse Jacobian of $\mathbf{q}(\mathbf{p},\mathbf{t})$ with respect to p. 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 in equilibrium—that is, holding all fleet sizes constant but allowing travel times to adjust. We can now obtain an expression for optimal prices:

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

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

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:*

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

$$\tilde{M}_j^q \equiv \sum_{k \neq j} D_{kj} \left(\mathbf{1}_{k \in \mathcal{J}_G} \cdot (C_k^q - p_k) - \sum_l \tilde{u}_l^T \cdot T_{lk}^q + \sum_{l \in \mathcal{J}_G} q_l \cdot \Omega_{lk} \right). \tag{7}$$

Proof. See Appendix D.2. □

This proposition characterizes optimal prices implicitly, as the solution to a system of $|\mathcal{J}_G|$ equations. In Appendix D.1, we derive a similar expression that decomposes optimal travel times into analogous terms.

We now explain Equation 5 in detail.[22] Imagine first an unconstrained social planner with $\lambda = 0$. Optimal prices take a Pigouvian form: they are equal to marginal costs plus corrections for marginal externalities, congestion effects, and an additional term that we call *diversion*, which we explain below. Importantly, the relevant marginal costs and externalities are those of an additional trip and not of an additional vehicle. For public transit, for instance, most of the costs (labor, vehicle depreciation, fuel, and energy) are related to the number of vehicles, and the marginal cost of a passenger is negligible holding the number of vehicles fixed.

Congestion 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.

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 weighted sum over modes of deviations of prices from a standard Pigouvian solution, $\left(C_k^q + E_k^q - \sum_l u_l^T \cdot T_{lk}^q - p_k\right)$ where weights are diversion ratios, D_{kj}. This term is zero whenever all other modes are already priced at the Pigouvian solution.

The last term arises due to budget constraints. The need to raise revenue to meet the budget makes the planner behave like a monopolist and introduces a market power markup in Equation 5. The social planner now also under-weights environmental externalities and there is 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. Fi-

[22] For readers seeking straightforward intuition, Appendix A derives similar results based on a simple illustrative model.

nally, 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 decomposing 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 D.3.

3.3 Empirical Model

We now move to the empirical version 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 a trip from CA a to CA a' at hour h.[23] 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:

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

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, ξ_{mj} are other unobservable components of demand, and

[23] 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.

ϵ^i_{mj} is an idiosyncratic taste shock. The value of time (VOT) is α_T/α^i_p. A higher value of time means that passengers assign greater disutility to time spent traveling, which affects the optimal policy. This leads the planner to provide more frequent public transit service and increase road taxes to reduce travel times.

Motivated by disparities in mode choice across the income distribution (Section 2.3), we allow α^i_p to vary across income levels.[24] This leads to heterogeneity in the time-money tradeoff, which is important to quantify distributional effects.

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 $\mathcal{J}^i_m$ with a probability equal to the empirical fraction of car owners among consumers of type i in market m.

The taste shock ϵ^i_{mj} 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 $\mathcal{J}^i_m$. 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 rather than to the outside option, implying that diversion terms play a larger role in the optimal prices from Proposition 1. Concretely, the taste shock takes the form $\epsilon^i_{mj} = \varsigma^i_{mg(j)} + (1-\rho)\eta^i_{mj}$, where $g(j)$ is the nest good j belongs to and 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 $j \neq 0$ in market m is therefore given by:

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

[24] Appendix S1.1 details how we assign individual cellphones to income groups.

Integrating over i—accounting for the distribution of price coefficients α_p^i and the choice sets $\mathcal{J}_m^i$—mode shares and trips for mode j in market m are:

$$\mathbb{P}_{mj} = \int \mathbb{P}^i_{mj}\, di \qquad \text{and} \qquad q_{mj} = N_m \cdot \mathbb{P}_{mj}. \tag{10}$$

Consumer surplus, in dollars, is given by:

$$\sum_m \int \frac{1}{\alpha_p^i} \mathbb{E}\left[\max_{j \in \mathcal{J}_m^i} \left\{ u^i_{mj} \right\} \right] di, \tag{11}$$

where the expectation integrates over the distribution of errors ϵ^i_{mj}.

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.[25]

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'.[26] 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'. We fix paths to those suggested by Google Maps.[27]

[25] 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.

[26] Because the city is a directed graph, the edge $e = (a, a')$ is different from edge $e' = (a', a)$.

[27] Although travelers could reoptimize paths in counterfactuals, a robustness exercise shows that this changes average travel times in our main counterfactuals by less than 0.1%.

Total vehicle flow during hour h on edge e is defined as:

$$F_{eh} = \sum_j w_j \cdot f_{ehj}, \tag{12}$$

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 have different effects on congestion. For cars, the number of vehicles is a function of trips $f_{ehj} \equiv \sum_{m \in \mathcal{M}_{hj}^e} q_{mj}$, where $\mathcal{M}_{hj}^e$ is the set of all markets in which travelers take a route that goes through edge e.[28] For buses, the number of vehicles is a function of frequencies $f_{ehj} \equiv \sum_{r \in \mathcal{R}_j^e} k_{rj}$, where $\mathcal{R}_{hj}^e$ is the set of bus routes that go through e.

For road-based modes, travel time over edge e at time h for mode j is:[29]

$$T_{ehj}^{\text{vehicle}} = \max\{T_{ej}^0, A_{ehj} \cdot F_{eh}^{\beta_j}\}. \tag{13}$$

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_{ej}^0 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. The coefficient A_{ehj} is a scale factor that captures spatial patterns like geography and road infrastructure, as well as time-varying shifters like weather patterns.

We define the in-vehicle time for mode j in market m as

$$T_{mj}^{\text{vehicle}} = \psi_{mj} \sum_{e \in P_{mj}} T_{ehj}^{\text{vehicle}}. \tag{14}$$

The sum simply adds the travel times over all edges in the path P_{mj}. We correct

[28] 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 E.4.

[29] For trains, we assume in-vehicle times are constant. We take Google Maps expected times.

it by a mode-specific distance-based factor ψ_{mj} that accounts for higher speeds on long trips due to highway usage. See Appendix E.3 for details.

Next, we explain how public transit wait times are determined. While we assume that passengers do not plan their arrival at the public transit stop, we do account for the uncertainty due to schedule violations. 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.[30] Route frequency is determined by its fleet size k_{r_mj}, with a mean time between vehicles of $1/k_{r_mj}$. From a traveler's perspective, however, the expected wait time also depends on reliability: irregular service with schedule violations lengthens mean wait times. With perfectly regular service, passengers arrive on average halfway between vehicles, so expected wait time is $1/(2k_{r_mj})$. Under random (Poisson) arrivals, it is $1/k_{r_mj}$.

We estimate a model that nests both extremes (for details see Appendix E.1). The expected wait time for passengers is given by

$$T^{wait}_{mj} = \frac{1+\omega^2}{2k_{r_mj}},$$

where ω is the coefficient of variation of the time between vehicles.[31] We estimate ω using schedule deviations: trains have $\omega = 0$, while for buses $\hat{\omega}^2 = 0.194$, implying substantially more variability than for trains.[32]

For ride-hailing, wait time T^{wait}_{mj} depends on three main factors: (1) they are lower in periods during which many drivers are working, since there are more idle drivers; (2) they are higher when demand for ride-hailing trips is high, depleting idle drivers; and (3) they are lower 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 E.2 presents the details of this driver movement model.

[30] We assume that the set of routes is fixed and equal to the routes running in Chicago in our data.

[31] If Google Maps suggests a route with transfers, the wait time is the sum of individual wait times.

[32] To estimate this number, we compute realized times between buses and divide them by their average at the hour by route level. The variance of this ratio is our estimate $\hat{\omega}^2 = 0.194$.

3.4 Costs and environmental externalities

We assume costs and environmental externalities are proportional to vehicle-miles driven. For cars and ride-hailing, the number of miles depends on how many passengers choose these modes. For buses and trains, the number of miles driven depends on their frequency; hence, the marginal cost of an additional passenger is effectively zero. This is a good approximation as long as vehicles are not operating at capacity. Figure 5 shows that this 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 the 2022 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).[33] Appendix E.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.

4 Estimation and Computation

4.1 Demand model

In this section, we explain how we estimate price and time coefficients α_T and α_p^i as well as the nesting parameter ρ. Recall that the utility of traveler i for taking mode j in market m is given by:

$$U_{mj}^i = \xi_{mj} + \alpha_T \cdot T_{mj} + \alpha_p^i \cdot p_{mj} + \epsilon_{mj}^i. \tag{15}$$

We assume that the price coefficient takes the form $\alpha_p^i = \alpha_p / y_i^{1-\alpha_{py}}$ (as in Miravete et al., 2023), so that α_{py} captures how the price coefficient varies with income y_i.

We assume that the unobserved shock takes the form $\xi_{mj} = \lambda_{od(m)} + \lambda_{t(m)} + \lambda_j +$

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

$\tilde{\xi}_{mj}$: it is the sum of fixed effects for origin-destination, hour, and transportation mode as well as a remaining term. The fixed effects thus capture common shocks along each dimension, such as origin-specific demand, time-specific shocks, or average mode quality.

To estimate our demand parameters, 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 a vector of moment conditions that we detail below.

Our estimation needs to address two endogeneity concerns. The first relates to the estimation of α_p^i: prices could be correlated with unobserved demand shocks $\tilde{\xi}_{mj}$. We first exploit the fact that the prices of cars and public transit are fixed and therefore not affected by time-varying demand shocks. In this respect, our strategy mirrors the use of coarse retail prices, which are not affected by local demand variation (DellaVigna and Gentzkow, 2019). Imposing this orthogonality between prices and demand shocks allows us to construct the following moment condition:

$$\mathbb{E}[p_{mj} \cdot \tilde{\xi}_{mj} \cdot \mathbb{1}\{j \neq \text{ride-hailing}\}] = \mathbf{0}.$$

Since our model includes origin-destination, hour, and mode fixed effects, this moment exploits how prices vary differentially across markets for different modes. For instance, car prices increase with distance, whereas public transit prices do not. We also include moments in which we interact prices with income quintiles π_m^y to identify heterogeneity in the sensitivity to prices.

While this variation is sufficient to identify the price coefficient, it does not exploit variation in ride-hailing prices. We incorporate such variation using an alternative strategy that addresses the fact that ride-hailing prices may respond to unobserved demand shocks (e.g., via surge pricing). Based on a city surcharge on ride-hailing trips beginning or ending downtown between 6 a.m. and 10 p.m., we run a differences-in-differences specification to estimate an own-price elasticity of $\hat{\eta} = -1.42$ (see Appendix C). We then add to our GMM estimator an indirect inference moment that matches the model-predicted elasticity to this estimated elasticity:

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

where $\tilde{\eta}_{mj}$ is the model-implied own-price elasticity and $\mathcal{M}_{\text{surcharge}}$ denotes the set of markets affected by the surcharge. The term $\hat{\eta}$ is a local elasticity for travelers who use ride-hailing in certain markets. We thus match it to the elasticity predicted by our model for that selected group, rather than matching it to the average elasticity across all travelers.

The second endogeneity issue concerns travel times, as these are an equilibrium object: positive demand shocks $\tilde{\xi}_{jm}$ for road-based modes lead to more travel, inducing congestion and longer travel times. This endogeneity biases the travel time coefficient upward, just as demand shocks bias the price coefficient in standard demand-supply models. To address this concern, we exploit the fact that travel times vary with distance in a mode-specific way, reflecting differences in their speeds. For example, buses take longer than private cars to cover the same distance because they make frequent stops and are harder to maneuver.

Following this idea, we construct an instrument by dividing the straight-line distance by the average mode-specific speed:

$$Z^1_{mj} = \frac{D_m}{\frac{1}{M} \cdot \sum_m S^0_{mj}},$$

where D_m is the straight-line distance between the origin and destination of market m and $\frac{1}{M} \cdot \sum_m S^0_{mj}$ is the city-wide free-flow speed for mode j. This instrument satisfies the exclusion restriction because it is independent of unobserved demand shocks influencing travel mode choices: the straight-line distance is purely geographic and therefore unaffected by infrastructure or any other local unobserved factors, and the city-wide free-flow speed by construction does not vary with any local factors. Our demand model includes origin-destination fixed effects, which control for distance and mitigate concerns that the remaining demand errors may reflect commuter sorting based on commuting preferences (e.g., higher-VOT workers living closer to work).

In addition to the main coefficients, we also need to identify the nest parameter ρ, which governs the strength of the correlation of shocks among modes within the nest. Berry (1994) notes that estimating ρ requires instruments that shift the probability of choosing mode j conditional on selecting one of the goods inside the nest (i.e., one of the inside options). We use three such instruments Z^2_{mj}, Z^3_{mj}, and Z^4_{mj} that are common in the literature: (i) the number of modes in the nest; (ii) the difference between product characteristics—free-flow travel times, in this case—of mode j and the average of the other modes in the nest; and (iii) the square of this difference, to increase power (Gandhi and Houde, 2019). The latter two use free-flow travel times, which are exogenous and not an equilibrium object, unlike observed travel times.

We collect all the instruments in vector $\mathbf{Z}_{mj} = (Z^1_{mj}, Z^2_{mj}, Z^3_{mj}, Z^4_{mj})$ and construct the following additional moment:

$$\mathbb{E}[\mathbf{Z}_{mj}\tilde{\xi}_{mj}] = \mathbf{0}.$$

Table 1 shows estimates for several specifications of our model, gradually building up to the main specification that we outline above. The first column presents OLS estimates of a logit model without a nested error structure or heterogeneity across consumers. Specification (2) shifts to a GMM estimator, including all our moments and allowing for heterogeneity in price sensitivity, but does not account yet for car ownership or for the nested structure of taste shocks. Compared to specification (1), accounting for endogeneity increases the sensitivity to travel times and prices as well as the value of time, as expected. Specification (3) adds car ownership by allowing for random choice set variation across travelers based on car ownership by home census tract. Specification (4) also introduces the nested structure of taste shocks. We find an average city-wide VOT of \$20.70, that ranges from \$8.22 for the bottom income quintile to \$36.80 for the top income quintile. Model (5) is our main specification. It is the same model as specification

Table 1: Demand estimation results

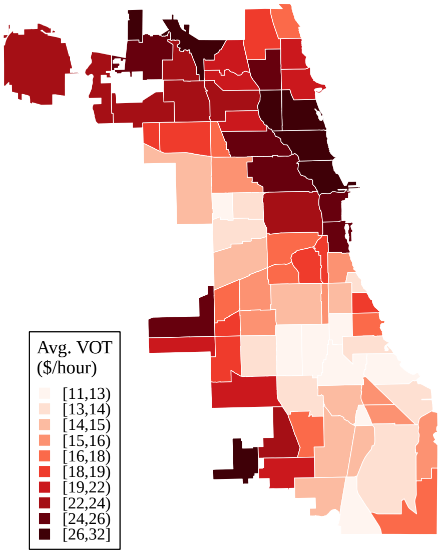
	Pooled				Peak/Off-Peak	
	(1)	(2)	(3)	(4)	(5)	
					Peak	Off-Peak
Time (α_T)	-0.913	-2.460	-2.542	-1.888	-1.674	-2.581
	(0.012)	(0.030)	(0.031)	(0.024)	(0.031)	(0.042)
Price (α_p)	-0.056	-1.234	-0.877	-0.622	-0.775	-0.934
	(0.001)	(0.054)	(0.055)	(0.045)	(0.141)	(0.106)
Income (α_{py})		-0.242	-0.050	-0.037	-0.362	0.005
		(0.026)	(0.034)	(0.038)	(0.094)	(0.053)
Nest (ρ)				0.442	0.514	0.381
				(0.001)	(0.012)	(0.019)
Estimator	OLS	GMM	GMM	GMM	GMM	
Policy Moment		✓	✓	✓	✓	
Car Ownership			✓	✓	✓	
Nest				✓	✓	
Avg. VOT ($/h)	16.20	20.42	20.29	20.70	27.83	17.58
VOT (Bot. Quintile)	16.20	6.58	7.95	8.22	8.00	7.19
VOT (Top Quintile)	16.20	39.63	36.29	36.80	57.36	30.30
Avg. Price Elast.	-0.19	-0.46	-0.46	-0.53	-0.45	-0.70
Avg. Time Elast.	-0.50	-1.34	-1.37	-1.59	-1.63	-1.89
M	92,326	91,941	91,595	91,595	42,995	48,600
N	285,331	284,562	283,704	283,704	138,066	145,638

Notes: This table presents demand estimation results from the specifications outlined in section 4.1. All specifications include origin-by-destination, hour, mode fixed effects, and they include control for dummies for multi-modal trips as well as transfers. 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. Standard errors are computed using a sandwich formula.

(4), 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 $8.00 to $57.36 per hour during peak hours (for the lowest and highest income quintiles, respectively), and from $7.19 to $30.30 per hour during off-peak hours. The city-wide average VOT is $23.30—around 80% of the mean hourly wage in Chicago, within the range of estimated values in the literature between 50% and 100% (Small, 2012).[34] After including income

[34] The ratio of VOT to hourly wages increases from 0.585 in the bottom to 0.807 in the top income quintile. Our annual income levels are [$26,154, $41,076, $53,750, $70,154.5, $111,024.5]. We as-

Figure 7: Value of time across space and diversion ratios



(a) Overall diversion ratios

From \ To	**Bus**	**Car**	**RideH**	**Train**	**Outside**
Bus	·	0.28	0.11	0.14	0.47
Car	0.12	·	0.06	0.09	0.73
RideH	0.22	0.26	·	0.13	0.40
Train	0.22	0.24	0.12	·	0.42

(b) Div. ratios for top income quintile

From \ To	**Bus**	**Car**	**RideH**	**Train**	**Outside**
Bus	·	0.45	0.14	0.09	0.32
Car	0.11	·	0.12	0.08	0.69
RideH	0.14	0.44	·	0.10	0.33
Train	0.13	0.39	0.16	·	0.31

(c) Div. ratios for bottom income quintile

From \ To	**Bus**	**Car**	**RideH**	**Train**	**Outside**
Bus	·	0.32	0.05	0.14	0.49
Car	0.12	·	0.02	0.10	0.77
RideH	0.23	0.28	·	0.12	0.37
Train	0.24	0.29	0.04	·	0.43

Notes: The map on the left shows the average VOT implied by our main demand specification (column (6) of Table 1) for trips originating in each CA. Income heterogeneity is driven by differences in the price coefficient. The tables on the right present the average diversion ratios implied by our main demand specification for all consumers as well as for the highest- and lowest-income consumers. Individual diversion ratios are averaged across markets, weighted by market size.

heterogeneity in column (3), the average VOT is quite stable across different specifications, and robustness checks show similar values between $19.39 and $25.90 per hour (Appendix F.1).

We now present the main spatial patterns implied by our estimates. The left panel of Figure 7 shows the VOT by origin CA. It tends to be higher in the North Side, which is characterized by higher incomes. Most South Side CAs display low VOTs. Exceptions include Midway Airport and the neighborhoods of Beverly, Mount Greenwood, and Morgan Park—white-flight destinations in the 1950s–60s. These estimates correlate closely with the patterns in the bottom panel of Figure 1.

The right panel of Figure 7 presents substitution patterns in the form of diversion ratios, both for the whole population and separately for the top and bottom

sume 2000 worked hours during a year.

income quintiles. 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:

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

where $F_{eh} = \sum_j w_j f_{ehj}$. We set $w_{car} = w_{ride-hail} = 1$ and $w_{bus} = 2$, implying that buses congest twice as much as cars.[35]

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_{ej}^0 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 because of unusual early-morning schedules.

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

$$\log T_{ehj}^{\text{vehicle}} = a_{ej} + \beta_j \log F_{eh} + \varepsilon_{ehj}. \tag{16}$$

The coefficient β_j is the congestion elasticity outside of free-flow times and measures how responsive vehicle flows are to reductions in traffic. A higher value for β_j will lead to lower optimal road congestion charges, as has been foreshadowed by Equation 5. The edge-mode fixed effect a_{ej} captures any edge-specific differences in geography or infrastructure that determine travel times. The remaining

[35] These values follow London's Traffic Modeling Guidelines.

error ε_{ehj} captures unobservable shocks that vary across hours of the week h within edge e.

Table 2: Traffic congestion estimation results

Dependent Variable: Log travel time in traffic

	Bus			Car		
	(1)	(2)	(3)	(4)	(5)	(6)
Log Flow	0.083***	0.053***	0.089***	0.129***	0.109***	0.174***
	(0.006)	(0.006)	(0.009)	(0.004)	(0.004)	(0.004)
Edge FE	✓	✓	✓	✓	✓	✓
Weather controls		✓	✓		✓	✓
IV			✓			✓
within R^2	0.074	0.117	0.107	0.399	0.522	0.440
First-stage F			2795.094			4200.774
Observations	8367	8367	8367	11724	11724	11724

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 2 presents the estimates of Equation 16 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).[36]

A remaining concern is that travelers may re-optimize their choices in response to expected but unobservable local traffic shocks, such as planned construction during certain hours of the day. To address those concerns, specifications (3) and

[36] Adding controls X_{ehj} amounts to adjusting the model so that $\log A_{ehj} = a_{ej} + \gamma X_{ehj} + \varepsilon_{ehj}$.

(6) instrument traffic flows with the city-wide number of travelers by hour, following Kreindler (2024). This strategy is valid as long as the city-wide demand for travel is driven by daily patterns—commuting to work in the morning, leisure activities in the evening, etc.—and not local shocks.

We estimate congestion elasticities of 0.08-0.09 for buses and 0.13-0.17 for cars, matching previous studies (Akbar and Duranton, 2017; Couture et al., 2018). The main elasticities that we use for our model are those in columns (3) and (6).

4.3 Solving for Equilibrium and the Planner's Problem

Before we move on to describe our counterfactuals, we restate our equilibrium Definition 1 and social welfare function in the context of our empirical 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 model:* $T_{mj}^{vehicle} = \psi_m \sum_{e \in r_{mj}} \max\{T_{ej}^0, A_{ehj} \cdot F_{eh}^{\beta_j}\}$*. 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 frequency of buses and trains. Wait times for cars are zero.*

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

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.[37]

Appendix E.5 describes the algorithm we use to find an equilibrium (a limited-memory version of Broyden's method). 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.[38] Appendix E.6 shows that our model fits the data well.

Our empirical social welfare function is based on the estimated parameters and model structure detailed in Section 4. It includes several components. The first component is consumer surplus, following equation (11).[39] The second component are the costs of operating the transportation system and environmental externalities, using the values specified in Appendix E.4. The third component is the government's revenue, accounting for public transit prices as well as road taxes. The last component is the profit of private transportation providers (i.e., ride-hailing companies). To solve the social planner's problem, we follow the augmented Lagrangian method (Nocedal and Wright, 2006), where we iteratively maximize problems that approximate the Lagrangian of the main problem until convergence. Every evaluation of the Lagrangian requires solving for the transportation equilibrium. Further details are provided in Appendix E.7.

5 Optimal Policy Design

In what follows, we explore counterfactual policy designs based on the optimality conditions we derive in Proposition 1. We start by analyzing coarse policies that

[37] 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 $\mathbf{q} = q(\mathbf{p}, T(\mathbf{q}, \mathbf{k})) = f^{\mathbf{p},\mathbf{k}}(\mathbf{q})$.

[38] 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.

[39] The heterogeneity of our empirical model arises from price coefficients α_p^i, car ownership status, and idiosyncratic shocks $\epsilon^{\mathbf{i}} = (\epsilon_{ij})_j$. To compute consumer welfare, we integrate over those dimensions.

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 then 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 3. Each column represents one counterfactual policy, reporting results relative to the *Status Quo* (column 1). We also refer to Figures 8 and 9, which decompose the forces that give rise to the optimal policies for buses and cars, as in our theoretical results from Section 3.2.[40] In these graphs, 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.

We start with *Transit*, where the planner would want to set somewhat negative prices for buses and trains. Because the social cost of an additional passenger—the sum of marginal costs, environmental externalities, and congestion—is zero, the only non-zero component of this optimal price is the diversion term, which is negative because the planner wants to divert travelers away from socially underpriced cars (see Figure 8). The optimal wait times are 5.34 minutes for buses and 3.14 minutes for trains. Marginal costs, environmental externalities, and congestion all work in favor of fewer buses and thus longer waits. The only countervailing force is diversion—encouraging people to shift away from private cars. Accounting for all these forces, the optimal wait times are lower than those in the status quo.

These price and wait time changes in the *Transit* scenario increase welfare by $3.89 million per week relative to the status quo. Consumer surplus increases by

[40] The expressions that we use for this higher-dimensional problem, which include spillovers across markets, are derived in Appendix D.3. Appendix F.2 presents figures for trains, which are almost identical to those for buses.

Table 3: Counterfactual results

		Status Quo (1)	Transit (2)	Transit, Budget (3)	Road Pricing (4)	Transit + Road Pricing (5)
Panel A: Prices						
Avg. Price ($)	Bus	1.09 [1.09, 1.09]	-0.61 [-0.89, -0.46]	1.46 [0.60, 2.84]	1.09 [1.09, 1.09]	-0.01 [-0.12, 0.11]
	Train	1.33 [1.33, 1.33]	-0.81 [-1.20, -0.60]	1.90 [0.73, 3.88]	1.33 [1.33, 1.33]	-0.05 [-0.22, 0.10]
Road Tax ($/km)		0	0	0	0.36 [0.30, 0.46]	0.34 [0.26, 0.45]
Panel B: Wait Times and Frequencies						
Avg. Wait (min)	Bus	7.07 [7.06, 7.09]	5.34 [3.68, 7.43]	6.45 [4.42, 8.97]	7.07 [7.07, 7.08]	5.28 [3.67, 7.25]
	Train	4.44 [4.44, 4.45]	3.14 [2.22, 4.08]	3.67 [2.64, 4.65]	4.44 [4.44, 4.45]	3.13 [2.22, 4.05]
Δ Frequency	Bus	0%	32.5% [-4.8%, 92.2%]	9.6% [-21.2%, 60.1%]	0%	33.9% [-2.5%, 92.9%]
	Train	0%	42.8% [10.6%, 101.5%]	22.4% [-2.1%, 69.0%]	0%	43.1% [11.4%, 101.5%]
Panel C: Trips						
Number of Trips (M/week)	Bus	3.63 [3.62, 3.66]	5.41 [4.97, 6.05]	3.63 [3.42, 3.94]	4.04 [3.94, 4.17]	5.51 [5.15, 6.14]
	Train	2.76 [2.73, 2.78]	3.76 [3.72, 3.93]	2.75 [2.61, 3.04]	3.01 [2.97, 3.06]	3.76 [3.73, 3.96]
	Ride-hailing	2.94 [2.77, 3.21]	2.74 [2.56, 2.96]	2.93 [2.79, 3.10]	3.08 [2.95, 3.24]	2.91 [2.73, 3.12]
	Car	21.33 [21.11, 21.47]	20.12 [19.87, 20.22]	21.33 [21.17, 21.36]	18.62 [17.94, 19.15]	17.88 [17.41, 18.15]
	Total	30.66 [30.64, 30.69]	32.03 [31.90, 32.35]	30.65 [30.62, 30.75]	28.74 [28.39, 29.02]	30.06 [29.70, 30.52]
Panel D: Welfare						
Δ Welfare ($M/week)		0	3.89 [2.15, 17.94]	0.62 [0.10, 11.22]	3.80 [3.49, 4.28]	6.97 [5.03, 21.34]
Δ CS ($M/week)		0	26.87 [14.81, 58.37]	0.95 [-0.07, 12.89]	-32.32 [-41.56, -26.20]	-7.84 [-12.30, 12.18]
Δ City Surplus ($M/week)		0	-21.67 [-37.84, -12.42]	0	31.93 [25.32, 41.81]	12.18 [8.26, 13.61]
Δ Transit Surplus ($M/week)		0	-21.67 [-37.84, -12.42]	0	0.77 [0.61, 1.01]	-15.82 [-29.39, -8.81]
Road Taxes ($M/week)		0	0	0	31.15 [24.37, 41.20]	28.00 [20.98, 37.73]
Δ Externalities ($M/week)		0	-0.17 [-0.90, 1.04]	0.24 [-0.50, 1.37]	-2.72 [-3.26, -2.29]	-2.44 [-3.53, -0.96]

Notes: This table compares prices, frequencies, trips, and welfare relative to the *Status Quo* (column 1) across 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*. The main values represent counterfactuals based on point estimates of the model parameters. Square brackets below each value represent bootstrap-based 95% confidence intervals.

$26.87 million per week, or $9.95 per resident, due to reductions in both prices and wait times. This represents a substantial gain; for comparison, the city-wide cost of transportation is $29 per passenger per week. However, the planner's deficit also grows by $21.67 million per week. Environmental externalities fall slightly by $0.17 million per week: while lower prices shift travelers away from environmentally costly cars, the effect is almost entirely offset by more buses and trains running.

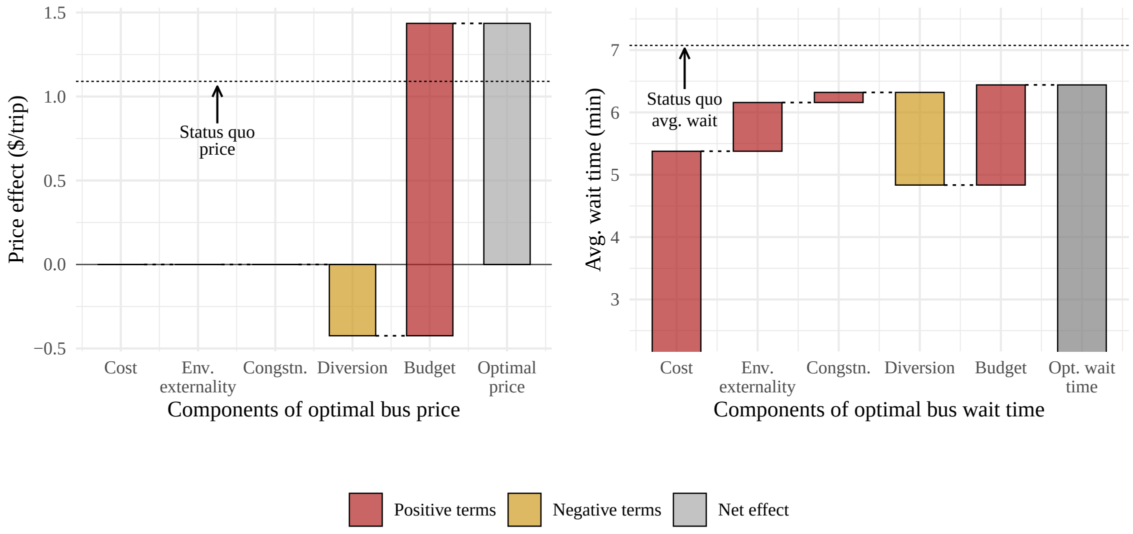


Figure 8: Optimal bus price and wait time decomposition for *Transit, Budget*

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 lead to very different results relative to the above unconstrained policy. Welfare increases only by $0.62 million per week relative to the status quo. The main difference is that traveler surplus increases much less, since a binding budget constraint leads the government to make two adjustments that hurt travelers: rising fares and reducing the frequency of buses and trains by around 20%, as can be seen by the budget terms in Figure 8.[41] These effects can be further decomposed into the markup

[41] At the optimal constrained policy, the marginal value of public funds (Hendren and Sprung-

term and the Spence distortion. 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 higher wait times for public transit. Despite the reduction in frequencies, the number of public transit trips remains almost unchanged 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 in Chicago deviate from the optimum. In *Transit, Budget*, both bus and train frequencies are higher, and their prices must be increased to balance the budget. One possible reason the CTA deviates from this optimum is that it faces pressure to keep fares affordable to low income travelers: as we show in Section 5.2, the price and frequency adjustments in *Transit, Budget* are regressive.

We now turn to *Road Pricing*.[42] When it is the only lever available to the government, the optimal per-km tax is 36 cents, or $15.2 per day for the average car commuter.[43] Figure 9 shows that this almost doubles the status quo price of driving a car (the marginal cost). About one third of the tax is due to environmental externalities and the remaining two thirds are due to congestion externalities. The diversion term is nearly zero, since two opposing forces cancel out: it is optimal to divert passengers towards public transit but away from ride-hailing. The budget term is zero because road tax revenue generates a fiscal surplus, so the budget constraint becomes nonbinding.

The overall welfare gains from *Road Pricing*, $3.80 million per week, are much larger than what budget-constrained transit policies alone can achieve. However, these gains predominantly result from a reduction in environmental externalities, while travelers are worse off. In the absence of rebates (the numbers shown in Table 3), consumer surplus decreases by $32.32 million per week, or $12 per resi-

Keyser, 2020), which we derive from the Lagrange multiplier of the budget constraint, is 1.37: every dollar that the government spends translates into a $1.37 increase in welfare.

[42] In our *Road Pricing* counterfactuals, ride-hailing trips do not pay road taxes. In a different counterfactual, we find that the status quo price of ride-hailing is 1% lower than the optimal price: the markup charged by ride-hailing companies is almost identical to the optimal Pigouvian tax.

[43] If, instead of a per-km tax, the government sets a cordon price for cars entering downtown Chicago, the optimal level is $8.75, resulting in welfare gains of $1.24 million per week.

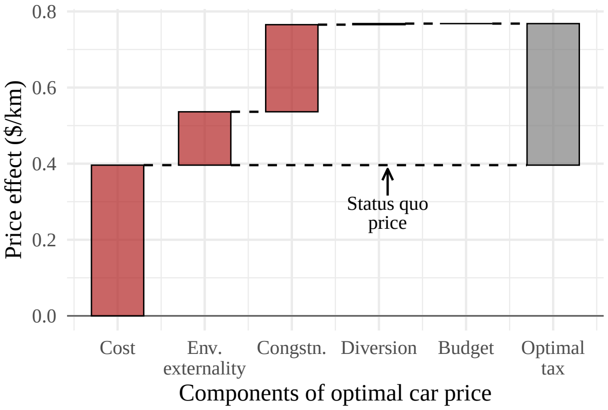


Figure 9: Optimal car price decomposition in the *Road Pricing* scenario

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.

dent per week. Even if the government fully rebated the revenue it collected from road taxes, consumers would lose $0.39 million in weekly surplus.[44] The number of public transit trips increases by 0.66 million per week, while car trips go down by 3.45 million per week.

Finally, when transit polices are combined with road pricing (*Transit + Road Pricing*) the planner can achieve welfare gains of $6.97 million per week—almost twice the sum of its parts, columns (3) and (4). Transit policies and road pricing are complementary due to cross-subsidization. Setting road taxes at $0.34 per km generates a government surplus, so the budget constraint no longer binds and resulting budget distortions are no longer present. The government is then able to spend part of the extra revenue to set 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 and it increases the frequencies of both buses and trains.

After optimally setting public transit frequencies and prices, some tax surplus remains, which the government can either keep or rebate back to consumers. As with *Road Pricing*, consumer surplus decreases without rebates—by $7.84 million

[44] *Road Pricing* also impacts commuters who enter and exit the City of Chicago. Assuming these travelers are perfectly inelastic, $27.45 million in tax revenue would be collected from them.

per week. On the other hand, if the government rebates its surplus, consumer surplus can increase by as much as $4.34 million per week. Combining efficient road pricing and cross-subsidizing public transit thus ends up benefiting travelers.

5.1 Sensitivity Analysis

In Figure 10 we explore the extent to which the optimal policies from Table 3 are sensitive to changes in key model parameters. We find that the optimal prices are very robust: 10% changes in parameter values change public transit prices by less than three cents and road taxes by less than two cents per km. By contrast, optimal wait times are more sensitive: for five of six parameters (public transit costs, price and time sensitivity, walking and waiting disutility, and bus variability), a 10% change shifts bus wait times by about 0.6 minutes and train wait times by about 0.2 minutes, corresponding to frequency changes of roughly 5%.

5.2 Distributional Effects

Figure 11 shows the effects of our counterfactuals on consumers across income quintiles. The left panel measures changes in consumer surplus per trip. Without rebates, most policies are regressive because higher public transit frequencies and lower congestion disproportionately benefit higher-income travelers, given their higher value of time. Under policies that involve road pricing, almost all income groups are worse off without rebates (the sole exception are the highest income travelers, who are roughly indifferent in *Transit + Road Pricing*). Losses are somewhat U-shaped because middle income consumers are the most reliant on cars, as 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. Thus, 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 to residents as a flat refund (dashed lines), which leads

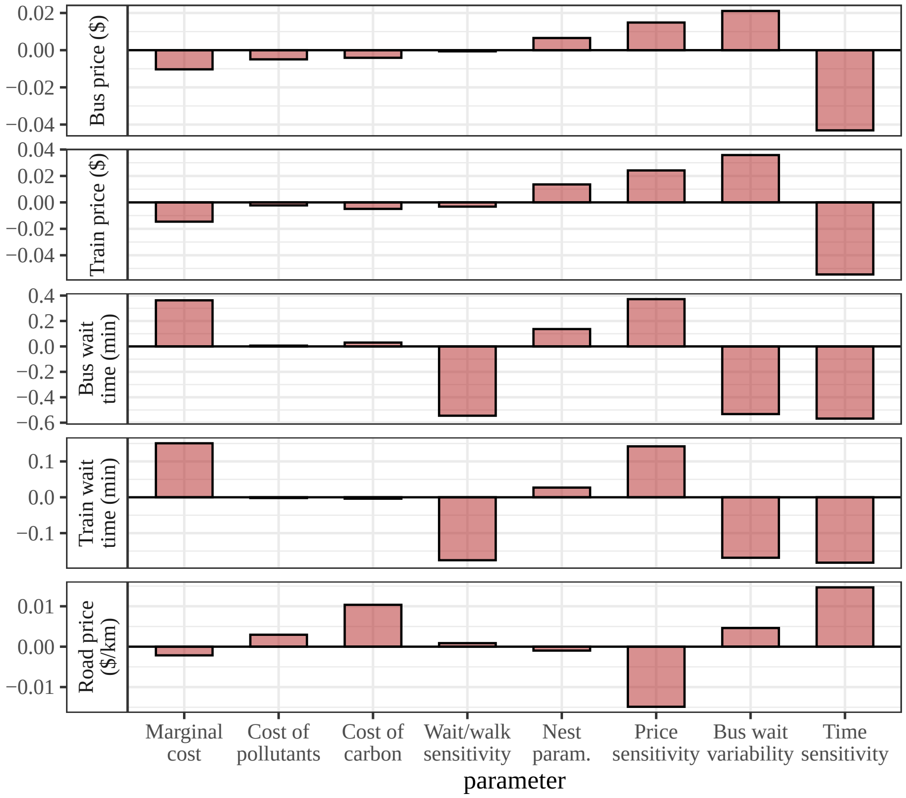


Figure 10: Robustness of counterfactual results

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.

the lowest income consumers to be better off.

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 4 presents results from several granular policies that we consider.

Setting different road taxes for the city center (CBD), 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

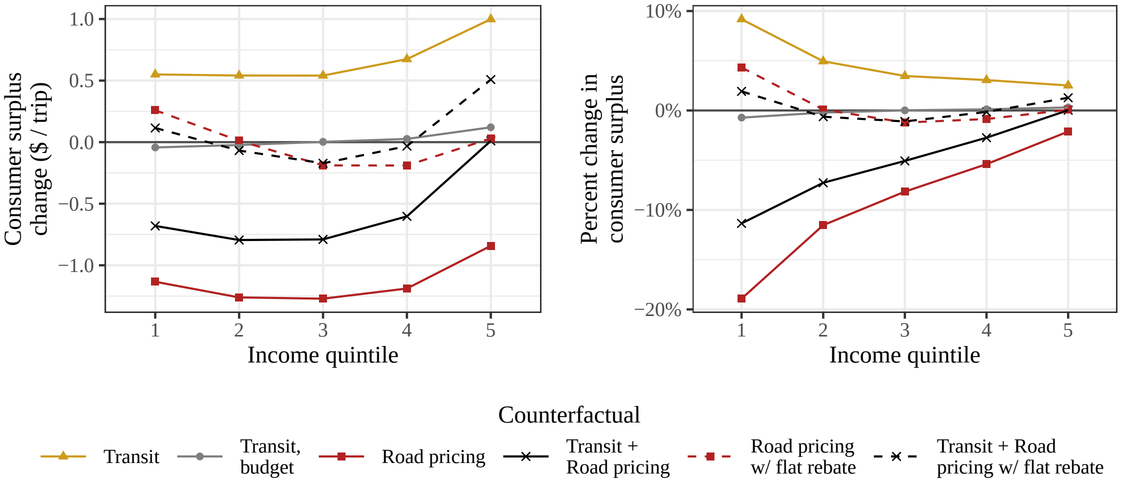


Figure 11: Change in consumer surplus across income quintiles

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 3. Dashed lines represent scenarios in which road pricing revenue is rebated back to residents as a flat refund.

of day. Furthermore, differences in the remaining two components of the optimal tax—congestion 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 almost double the frequency of high utilization routes while decreasing the frequency of low utilization routes by around 20%.[45] These adjustments result in welfare gains that are over five times larger than those from uniform frequency adjustments. However, these improvements only achieve around half of the welfare gains from combined road pricing and transit policies.

[45] There could be extra costs from having to re-allocate public transit capacity across hours or locations. While the CTA already adjusts frequency up and down to accommodate fluctuations in demand across times and locations, our results ignore any extra re-allocation costs beyond those incurred in the status quo.

Table 4: Granular counterfactual results

Panel A: Road Pricing	Uniform	Time Heterogeneity	Spatial Heterogeneity	Time + Spatial Heterogeneity
	(1)	(2)	(3)	(4)
Base ($/km)	0.36	0.34	0.34	0.32
Rush Hour ($/km)	·	0.39	·	0.35
CBD ($/km)	·	·	0.56	0.50
Rush Hour × CBD ($/km)	·	·	·	0.63
∆ Welfare ($M/week)	3.80	3.82	3.83	3.85

Panel B: Transit, Budget		Uniform	Time Heterogeneity	Utilization Heterogeneity
		(1)	(2)	(3)
∆ Bus Frequency (%)	Base	9.61	-5.90	-21.70
	Rush Hour	·	24.20	·
	High Utilization	·	·	96.73
∆ Train Frequency (%)	Base	22.37	-5.92	22.90
	Rush Hour	·	56.27	·
Bus Price ($)		1.46	1.39	1.33
Train Price ($)		1.90	1.81	1.72
∆ Welfare ($M/week)		0.62	1.50	3.39

Notes: This table presents levels and changes in prices, changes in frequencies, and changes in 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 (2024) 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 out-bound and return trips jointly, we only model individual trips. This choice arises from a data limitation: we are only able to link a small fraction of consecutive trips made by the same rider. Once again, the main challenge this brings to our model is that we cannot 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. However, prior research finds that long-run adjustments to transportation policy are limited. Herzog (2024) finds that sorting attenuates the welfare effects of time savings due to road pricing by around 20%. Barwick et al. (2024) show that residential sorting increases the overall welfare effects of road pricing by 18%, and Hierons (2024) finds that sorting only accounts for 10% of total welfare gains of cordon pricing in New York City.

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 a combined policy approach generates complementarities through cross-subsidization, yielding welfare gains larger than the sum of its parts.

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.

Barry, J.J., Newhouser, R., Rahbee, A. and Sayeda, S. (2002). Origin and destination estimation in new york city with automated fare system data. *Transportation Research Record* 1817(1):183–187.

Barwick, P.J., Li, S., Waxman, A., Wu, J. and Xia, T. (2024). Efficiency and equity impacts of urban transportation policies with equilibrium sorting. *American Economic Review* 114(10):3161–3205.

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., Matejka, F. and Salz, T. (2025). Personalized pricing and the value of time: Evidence from auctioned cab rides. *Econometrica* 93(3):929–958.

Castillo, J.C. (2025). Who benefits from surge pricing? *Econometrica* 93(5):1811–1854.

Cook, C. and Li, P.Z. (2025). 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.

Forkenbrock, D.J. (1999). External costs of intercity truck freight transportation. *Journal of Transportation and Statistics* 2(1):1–14.

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.

Hendren, N. and Sprung-Keyser, B. (2020). A unified welfare analysis of government policies. *The Quarterly journal of economics* 135(3):1209–1318.

Herzog, I. (2024). The city-wide effects of tolling downtown drivers: Evidence from london's congestion charge. *Journal of Urban Economics* 144:103714.

Hierons, T. (2024). Spreading the jam: Optimal congestion pricing in general equilibrium.

Holland, S.P., Mansur, E.T., Muller, N.Z. and Yates, A.J. (2016). Are there environmental benefits from driving electric vehicles? the importance of local factors. *American Economic Review* 106(12):3700–3729.

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).

Kreindler, G. (2024). Peak-hour road congestion pricing: Experimental evidence and equilibrium implications. *Econometrica* 92(4):1233–1268.

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.

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).

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, pp. 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.

Mohring, H. (1972). Optimization and scale economies in urban bus transportation. *American Economc Reveiew* 62(4):591–604.

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

OECD and ECMT (2003). *Reforming Transport Taxes*. Paris: Organisation for Economic Co-operation and Development.

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. (2025). Competing platforms and transport equilibrium. 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. and Verhoef, E.T. (2007). *The Economics of Urban Transportation*. New York: Routledge.

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 Stylized Model

This section presents the simplest possible model that we believe can illustrate our main theoretical findings from Section 3.2. For simplicity, we assume that travel times are fixed and, thus, omit the dependence on travel times throughout.

Consider a market in which travelers choose between traveling by car, by bus, and an outside option (such as walking or biking). The price for a bus trip is the fare p_b, and the price for a car trip p_c equals the cost of driving plus, possibly, a road tax. Demand for bus and car trips is given by $q_b(p_b, p_c)$ and $q_c(p_b, p_c)$. Demand for each mode is decreasing in its own price and increasing in the other price.

The cost of a bus trip, which is borne by the local government, is c_b. The cost of a car trip, which is borne by the traveler, is c_c. Additionally, each trip causes environmental externalities given by e_b and e_c.

The city government wants to maximize welfare. In the first best scenario, the government is able to set bus fares and road taxes, and does not face a budget constraint. In that case, the government's problem is

$$\max_{(p_b, p_c)} CS(p_b, p_c) + (p_b - c_b)q_b + (p_c - c_c)q_c - e_b q_b - e_c q_c.$$

The optimal prices are given by:

$$p_b = c_b + e_b \qquad \text{and} \qquad p_c = c_c + e_c. \tag{17}$$

This is the standard result in optimal taxation: the optimal price is equal to the marginal cost plus the marginal environmental externality.[46]

[46] To derive equation 17, it's easier to write the problem in terms of the quantities q_b and q_c, noting that consumer surplus can be written as $U(q_b, q_c) - p_c q_c - p_b q_b$, where U represents gross surplus:

$$\max_{(q_b, q_c)} U(q_b, q_c) - (c_b + e_b) \cdot q_b - (c_c + e_c) \cdot q_c.$$

Now suppose that the government faces a budget constraint B, and that it cannot set a road tax (so that $p_c = c_c$).The government's problem is then

$$\max_{p_b} U(q_b(p_b), q_c(p_b)) - (c_b + e_b) \cdot q_b(p_b) - (c_c + e_c) \cdot q_c(p_b) \quad \text{s.t.} \qquad (c_b - p_b)q_b(p_b) \leq B.$$

The first order condition is

$$\frac{\partial q_b}{\partial p_b}\left(\frac{\partial U}{\partial q_b} - (c_b + e_b) + \lambda\left(p_b + q_b\frac{\partial p_b}{\partial q_b} - c_b\right)\right) + \frac{\partial q_c}{\partial p_b}\left(\frac{\partial U}{\partial q_c} - (c_c + e_c)\right) = 0,$$

where λ is the Lagrange multiplier for the budget constraint. Substituting in $\frac{\partial U}{\partial q_b} = p_b$ and $\frac{\partial U}{\partial q_c} = p_c$, noting that $p_c = c_c$, and rearranging gives the following expression for the optimal bus price:

$$p_b = c_b + e_b - \frac{1}{1+\lambda}\underbrace{D_{bc} \cdot e_c}_{\text{Diversion}} + \frac{\lambda}{1+\lambda}\underbrace{(\mu_b - e_b)}_{\text{Budget}}, \tag{18}$$

where $D_{bc} = -\frac{\partial q_c}{\partial p_b} / \frac{\partial q_b}{\partial p_b}$ is the diversion ratio from buses to cars, and $\mu_b = -q_b \frac{\partial p_b}{\partial q_b}$ is the standard monopolist markup.

Equation 18 showcases the two new forces in Proposition 1 beyond Pigouvian taxes. First, there is a diversion term. The price of buses should thus be lower to the extent that (i) cars are under-priced since drivers do not pay for their environmental externality, and (ii) lowering the price of buses diverts travelers away from cars. This term is more complicated in Proposition 1 because, in our full model, the extent to which buses are underpriced also depends on traffic congestion externalities. The second force is that, to stay on budget, the government behaves somewhat like a monopolist—in this case, it does not give full weight to externalities, and it sets a market power markup.

The first order conditions are $\partial U/\partial q_b - c_b - e_b = 0$ and $\partial U/\partial q_c - c_c - e_c = 0$.

To derive the final expression, note that $\partial U/\partial q_b = p_b$ and $\partial U/\partial q_c = p_c$. The usual definition of gross utility for demand of one good is $U(q) = \int_0^q p(x)dx$, where $p(x)$ is inverse demand. It is thus clear that $\partial U/\partial q = p$. In the two-good case, gross utility $U(\mathbf{q}) = \int_{\mathbf{0}}^{\mathbf{q}} \mathbf{p}(\mathbf{r}) \cdot d\mathbf{r}$ is only well-defined when the demand function is integrable, in which case the gradient theorem gives $\partial U/\partial q_j = p_j$.

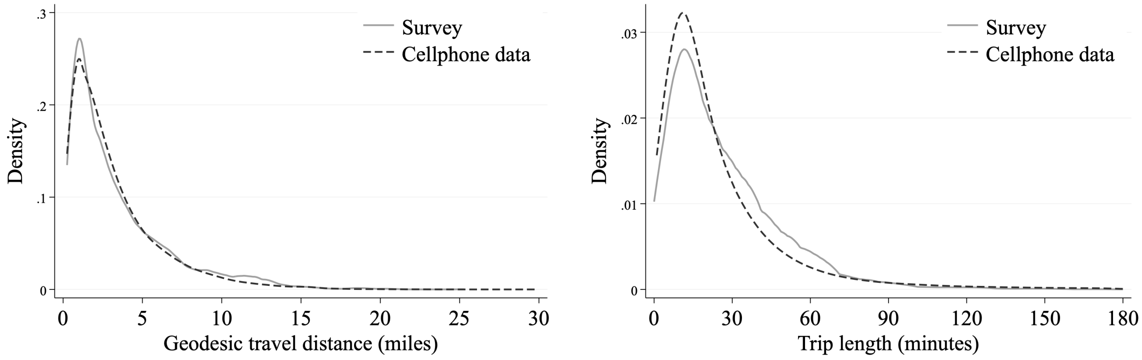
B Data Construction and Validation

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 and to January 2020.[47] 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 0.4km (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 and the probability of owning a car equal to tract's car ownership rates.

We validate our data in two ways. First, Figure A1 shows that survey and cellphone data travel time and travel distance distributions are very similar, showing that our cellphone data accurately represents travel patterns. Second, Figure A2 shows that the share of the tract population covered by the cellphone data is fairly constant and around 5% for all percentiles of the income distribution. This suggests that our cellphone location records cover a representative sample of the population in terms of income.

[47] Specifically, our subsample of pings is restricted to those with latitudes between 41.11512 and 42.494693, and longitudes between -88.706994 and -87.527174. This includes 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.

C Downtown Surcharge

On January 6, 2020, Chicago introduced a surcharge on ride-hailing trips starting or ending in a Downtown Zone (Figure A3) during peak hours (weekdays, 6 am–10 pm).[48] Single rides are taxed at $1.25 outside the zone and $3.00 inside it. Before the implementation of this surcharge, all trips faced a uniform surcharge of $0.72, regardless of location or time.[49]

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

$$y_{odt} = \mu_{od} + \alpha_t + \beta_t \cdot treat_{od} + \epsilon_{odt},$$

where y_{odt} is either log price or log trips, od refers to origin/destination CA. Time t is measured in 15-min intervals. $treat_{od}$ refers to all trips between areas subject to

[48] See City of Chicago website.

[49] See ABC7 Chicago.

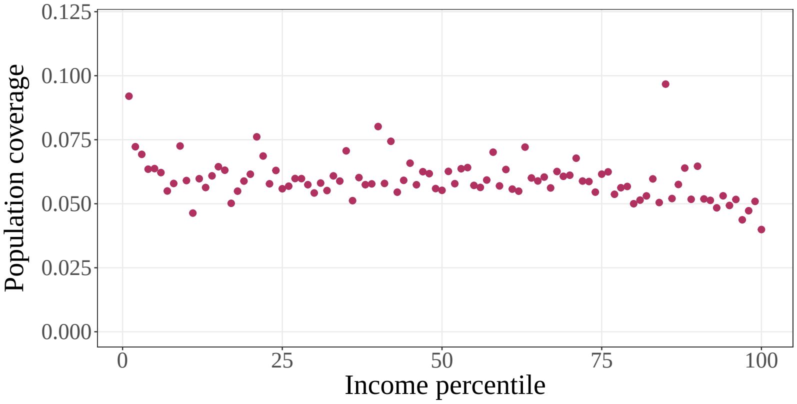


Figure A2: Representativeness across income groups

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.

the 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.

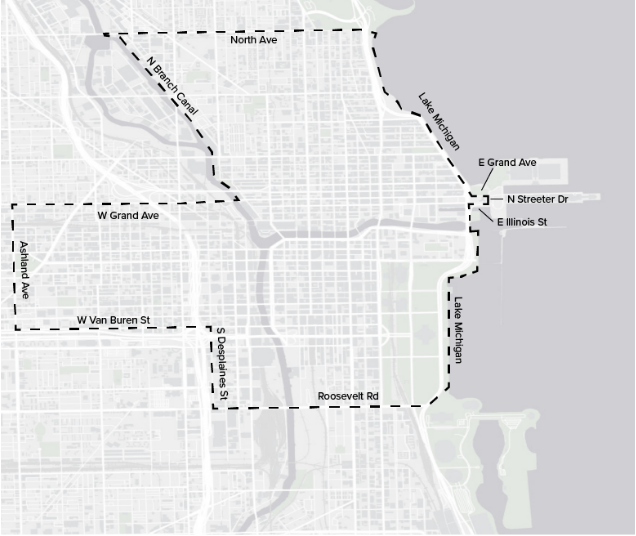
D Proofs and Additional Theoretical Results

D.1 Optimality condition for fleet size

We first introduce some notation. We decompose derivatives of travel times with respect to fleet sizes as $T^k_{jk} = \check{T}^k_{jk} + \tilde{T}^k_{jk}$, where $\check{T}^k_{jk}$ accounts for effects on waiting times (it is zero when $k \neq j$), and $\tilde{T}^k_{jk}$ accounts for effects due to travel times.

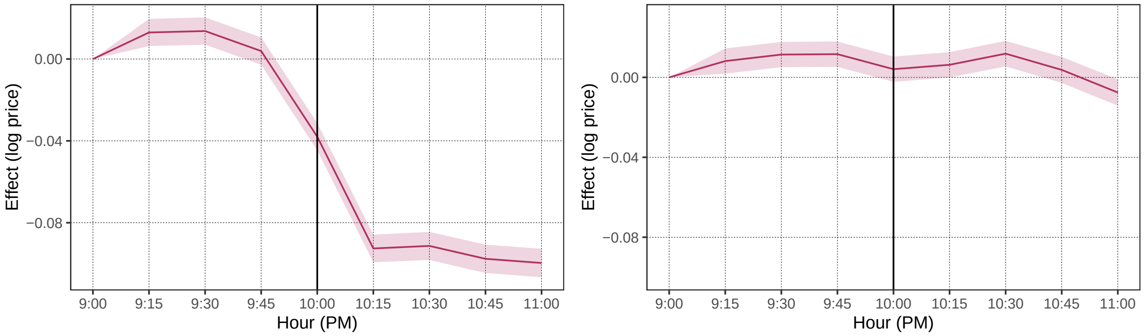
Proposition 2. *The first order conditions for the social planner's problem* (4) *with respect*

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.

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

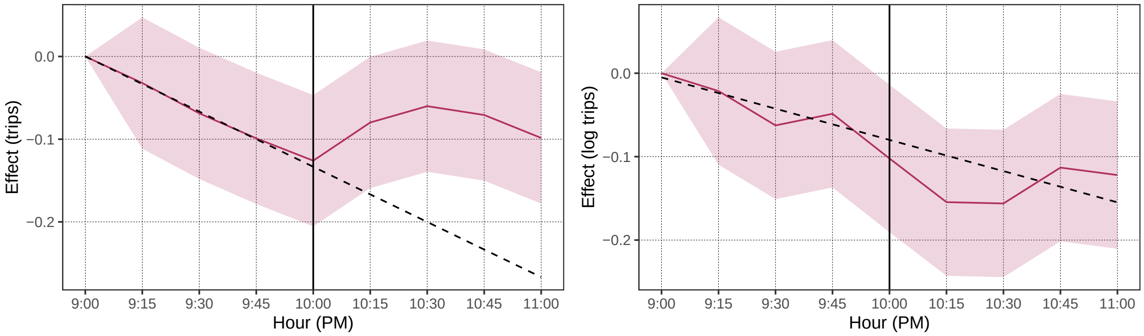


Notes: The left panel shows ride-hailing prices of areas affected by the surcharge of $3 relative to unaffected adjacent areas around 10 pm, after which the surcharge no longer applies. The right panel shows the same Figure in 2019, when the surcharge policy was not in place yet.

to fleet sizes can be written as

$$\overbrace{u_j^T \check{T}_{jj}^k}^{\substack{\textit{Direct benefit}\\ \textit{of fleet size}}} = \overbrace{C_j^k}^{\substack{\textit{Mg. cost}\\ \textit{of fleet size}}} + \overbrace{E_j^k}^{\substack{\textit{Mg. env. externality}\\ \textit{of fleet size}}} \quad \overbrace{- \sum_l u_l^T \cdot \tilde{T}_{lj}^k}^{\substack{\textit{Congestion}\\ \textit{effects}}} + \overbrace{M_j^k}^{\textit{Diversion}} + \\ \frac{\lambda}{1+\lambda}\Bigg(E_j^k + \underbrace{\sum_k (\tilde{u}_k^T - u_k^T) \cdot T_{lj}^k}_{\substack{\textit{Spence}\\ \textit{distortion}}} + \underbrace{\tilde{M}_j^k - M_j^k}_{\substack{\textit{Diversion}\\ \textit{distortion}}} \Bigg), \qquad (19)$$

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



Notes: The left 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 right panel shows the same Figure in 2019, when the surcharge policy was not in place yet and we see that the downwards trend continues.

where M_j^k and $\tilde{M}_j^k$ are defined as:

$$M_j^k = \sum_l \frac{\partial q_l}{\partial k_j} \left(C_l^q + E_l^q - \sum_m u_m^T \cdot T_{ml}^q - p_l \right)$$

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

Proof. See Appendix D.2 □

This result takes a very similar form to equation (5). Instead of the price, the left hand side is the direct benefit of an increase in the fleet size—on those riders taking that mode—which can be thought of as the direct benefit of an additional trip. The marginal cost, marginal externality, and congestion effects terms are almost identical, except that they are derivatives with respect to fleet sizes.

The diversion term follow a similar intuition to those for equation (5): they are 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.

D.2 Proof of Propositions 1 and 2

We start by proving two lemmas in which we obtain expressions that are used in the main proof. We first define $U(\mathbf{q},\mathbf{t}) = U(p(\mathbf{q},\mathbf{t}),\mathbf{t})$, gross utility as a function of trips and travel times. The first lemma characterizes its derivatives with respect to trips. As is standard when choices originate from utility maximization, these derivatives are equal to prices.

Lemma 1. *The marginal gross utility of an additional mode-j trip is given by the price of mode j:*

$$\frac{\partial U}{\partial q_j} = p_j.$$

Proof. Taking the derivative of gross utility with respect to q_j (using the Leibniz integral rule), we get that $\frac{\partial}{\partial q_j}U(\mathbf{q},\mathbf{t}) = \sum_k \int_{\partial\Theta_k(\mathbf{q},\mathbf{t})} u_k(t_k,\theta)e_k^j(\theta)f(\theta)\,d\theta + \sum_{k,l\neq k}\int_{\partial\Theta_{kl}(\mathbf{q},\mathbf{t})} u_k(t_k,\theta)e_k^j(\theta)f(\theta)\,d\theta$, where $\partial\Theta_j(p,t)$ is the boundary between $\Theta_j(p,t)$ and $\Theta_0(p,t)$, $\partial\Theta_{jk}(p,t)$ is the boundary between $\Theta_j(p,t)$ and $\Theta_k(p,t)$, and $e_k^j(\theta)$ denotes by how much $\Theta_k(\mathbf{q},\mathbf{t})$ expands at θ as q_j increases. There is no term corresponding to the interior because t is fixed (and so is $u_k(t_k,\theta)$).

The second sum involves two terms for every pair. We can collect them together by noting that $e_k^j(\theta) = -e_l^j(\theta)$ at the boundary between $\Theta_k(\mathbf{q},\mathbf{t})$ and $\Theta_l(\mathbf{q},\mathbf{t})$. The above expression is thus $\sum_k \int_{\partial\Theta_k(\mathbf{q},\mathbf{t})} u_k(t_k,\theta)e_k^j(\theta)f(\theta)\,d\theta + \sum_{k,l>k}\int_{\partial\Theta_{kl}(\mathbf{q},\mathbf{t})}(u_k(t_k,\theta) - u_l(t_l,\theta))e_l^j(\theta)f(\theta)\,d\theta$. To avoid counting every border twice, this expression only consider pairs (k,l) such that $k > l$ based on any arbitrary ordering of modes.

Agents at the boundaries are indifferent, so $u_k(t_k,\theta) = p_k$ for the first sum and $u_k(t_k,\theta) - u_l(t_l,\theta) = p_k - p_l$ for the second sum. Substituting and rearranging terms, we get $\sum_k \int_{\partial\Theta_k(\mathbf{q},\mathbf{t})} p_k e_k^j(\theta)f(\theta)\,d\theta + \sum_{k,l\neq k}\int_{\partial\Theta_{kl}(\mathbf{q},\mathbf{t})} p_k e_k^j(\theta)f(\theta)\,d\theta = \sum_k p_k \left(\int_{\partial\Theta_k(\mathbf{q},\mathbf{t})} e_k^j(\theta)f(\theta)\,d\theta + \sum_{l\neq k}\int_{\partial\Theta_{kl}(\mathbf{q},\mathbf{t})} e_k^j(\theta)f(\theta)\,d\theta\right)$. The term in parentheses is how much $\Theta_k(p,t)$ expands in total into all other regions, so it is equal to $\partial q_k/\partial q_j$, which is thus equal to 1 for $k = j$ and 0 for $k \neq j$. Hence, $\partial U(\mathbf{q},\mathbf{t})/\partial q_j = p_j$. □

Our second lemma concerns partial derivatives $\partial p_j/\partial t_k$ of the inverse demand function $p(\mathbf{q}, \mathbf{t})$. Our analysis follows Weyl and White (2010).

We first define some notation. Let $\mathbb{E}_j[u_j^t]$ represent the mean of $|\partial u_j(t,\theta)/\partial t|$—the marginal disutility of travel time—among agents that are indifferent between mode j and the outside option. Also let $\mathbb{E}_{kj}[u_j^t]$ represent the mean of $|\partial u_j(t,\theta)/\partial t|$ among agents that are marginal between modes j and k. In other words, these two quantities represent the mean disutility of time for the set of agents that are marginal between two options.

The next lemma shows that diagonal terms $\partial p_j/\partial t_j$ take the form of a weighted mean of terms $\mathbb{E}_j[u_j^t]$ and $\mathbb{E}_{jk}[u_j^t]$, while off-diagonal terms $\partial p_j/\partial t_k$ for $j \neq k$ take the form of differences between terms $\mathbb{E}_j[u_j^t]$ and $\mathbb{E}_{jk}[u_j^t]$. This means that $\sum_{l\in\mathcal{J}_G} q_l \cdot \partial p_l/\partial t_j$—how much more revenue the government can extract after a change in t_j, holding trips fixed—is equal to a weighted sum of marginal disutilities of travel time ($\mathbb{E}_j[u_j^t]$ and $\mathbb{E}_{kj}[u_j^t]$) with sum of weights equal to the number of travelers using government-controlled modes.

Lemma 2. *$\partial p_l/\partial t_j$ is equal to a weighted sum $w_j^{lj}\mathbb{E}_j[u_j^t] + \sum_{k\neq j} w_{kj}^{lj}\mathbb{E}_{kj}[u_j^t]$, where the sum of weights $w_j^{lj} + \sum_{k\neq j} w_{kj}^{lj}$ is equal to one for $l = j$ and equal to zero for $l \neq j$.*

$\tilde{u}_j^T \equiv \sum_{l\in\mathcal{J}_G} q_l \cdot \partial p_l/\partial t_j$ takes the form of a weighted sum $\tilde{w}_j^j\mathbb{E}_j[u_j^t] + \sum_{k\neq j}\tilde{w}_{kj}^j\mathbb{E}_{kj}[u_j^t]$, where the sum of weights $\tilde{w}_j^j + \sum_{k\neq j}\tilde{w}_{kj}^j$ is equal to $\sum_{l\in\mathcal{J}_G} q_l$.

Proof. The function $p(\mathbf{q}, \mathbf{t})$ is defined implicitly by $\mathbf{q} = q(\mathbf{p}, \mathbf{t})$. By the implicit function theorem, the matrix of its partial derivatives $\partial p_j/\partial t_k$ is $J_t^p = -\left[J_p^q\right]^{-1} J_t^q$, where J_p^q and J_t^q represent the Jacobians of $q(p,t)$ with respect to p and t, respectively.

We now obtain expressions for those Jacobians. Consider first an increase in p_j. This induces N_j customers to switch from mode j to the outside option, and it induces N_{kj} customers to switch from mode j to mode k. Thus, $\frac{\partial q_j}{\partial p_j} = -N_j - \sum_{k\neq j} N_{kj}$ and $\frac{\partial q_k}{\partial p_j} = N_{kj}$ for $k \neq j$, where N_j and N_{jk} are given by integrals over the boundaries $\partial\Theta_j(p,t)$ and $\partial\Theta_{jk}(p,t)$, $N_j = \int_{\partial\Theta_j(p,t)} n_j(\theta) f(\theta)\, d\theta$ and $N_{kj} = \int_{\partial\Theta_{kj}(p,t)} n_{kj}(\theta) f(\theta)\, d\theta$. The integrands represent the number of travelers at θ that are willing to switch modes in response to a unit increase in utility (i.e., a decrease in price). They are equal to the product of $f(\theta)$, the density of agents, times the volume (in θ space) of types that are willing to switch modes, which is given by $n_j(\theta)$ or $n_{jk}(\theta)$. That volume is given by the directional derivative of the inverse of $\partial u_j/\partial\theta$ or $\partial(u_k - u_j)/\partial\theta$ in the direction that is normal to the boundary.

Now consider an increase in t_j. This induces M_j customers to switch from mode j to the outside option, and it induces M_{kj} customers to switch from mode j to mode k, where $M_j = \int_{\partial\Theta_j(p,t)} \frac{\partial u_j}{\partial t_j} n_j(\theta) f(\theta)\, d\theta$ and $M_{kj} = \int_{\partial\Theta_{kj}(p,t)} \frac{\partial u_j}{\partial t_j} n_{kj}(\theta) f(\theta)\, d\theta$. These are similar integrals as before, but the integrand now accounts for the fact that an increase in times no longer induces a unit decrease in utility, but an increase of $\partial u_j(\theta, t_j)/\partial t_j$. They are thus weighted sums of $\frac{\partial u_j}{\partial t_j}$, which can also be written as $M_j = \mathbb{E}_j[u_j^t]N_J$ and $M_{kj} = \mathbb{E}_{kj}[u_j^t]N_{kj}$. We can thus write $\frac{\partial q_j}{\partial t_j} = -\mathbb{E}_j[u_j^t]N_j - \sum_{k\neq j} \mathbb{E}_{kj}[u_j^t]N_{kj}$ and $\frac{\partial q_k}{\partial t_j} = \mathbb{E}_{kj}[u_k^t]N_{kj}$ for $k \neq j$.

We now use Cramer's rule to compute the elements of J_t^p. Element (l, j) is $-\det(\tilde{J}_{lj})/\det(J_p^q)$, where $\tilde{J}_{lj}$ is the matrix J_p^q with the l-th column replaced by the j-th column of J_t^q. By using the Laplace expansion for determinants, $(J_t^p)_{lj} = -\frac{\sum_k (J_t^q)_{kj}(-1)^{k+l}\det(\check{J}_p^q(k,l))}{\det(J_p^q)} = w_j^{lj}\mathbb{E}_j[u_j^t] + \sum_{k\neq j} w_{kj}^{lj}\mathbb{E}_{kj}[u_j^t]$, where $\check{J}_p^q(k, j)$ is the submatrix of J_p^q without row k and column j. This is a weighted sum of terms $\mathbb{E}_j[u_j^t]$ and $\mathbb{E}_{kj}[u_j^t]$, where the weights take the form $w_j^{lj} = N_j(-1)^{j+l}\det(\check{J}_p^q(j, l))/\det(J_p^q)$ and $w_{kj}^{lj} = -N_{kj}[(-1)^{j+l}\det(\check{J}_p^q(j, l)) - (-1)^{k+j}\det(\check{J}_p^q(k, l))]/\det(J_p^q)$. Hence, the weights are simply functions of N_j and N_{jk}.

If the marginal value of all marginal agents is equal to the same value $\bar{u}$, then $J_t^q = -\bar{u}J_p^q$, which means that $J_t^p = -(J_p^q)^{-1}J_t^q$ is equal to $\bar{u}$ times the identity matrix. Thus, we can conclude that the sum of the weights corresponding to term $(J_t^p)_{lj}$ is equal to minus one for diagonal elements ($l = j$) and is zero otherwise.

We then have that $\sum_{l\in\mathcal{J}_G} q_l \frac{\partial p_l}{\partial t_j} = \sum_{l\in\mathcal{J}_G} q_l \big(w_j^{lj}\mathbb{E}_j[u_j^t] + \sum_{k\neq j} w_{kj}^{lj}\mathbb{E}_{kj}[u_j^t]\big) = \big(\sum_{l\in\mathcal{J}_G} q_l w_j^{lj}\big)\mathbb{E}_j[u_j^t] + \sum_{k\neq j}\big(\sum_{l\in\mathcal{J}_G} q_l w_{kj}^{lj}\big)\mathbb{E}_{kj}[u_j^t] = \tilde{w}_j^j\mathbb{E}_j[u_j^t] + \sum_{k\neq j}\tilde{w}_{kj}^j\mathbb{E}_{kj}[u_j^t]$, where $\tilde{w}_j^j + \sum_{k\neq j}\tilde{w}_{kj}^j = \sum_{l\in\mathcal{J}_G} q_l$. □

We now present the main proof of Propositions 1 and 2.

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

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

The first order condition for p_j is

$$\sum_l \frac{\partial q_l^*}{\partial p_j} \left[\frac{\partial U}{\partial q_l} + \sum_m u_m^T T_{ml}^q - C_l^q - E_l^q + \lambda \left(\mathbf{1}_{l \in \mathcal{J}_G} \cdot (p_l - C_l^q) + \sum_{m \in \mathcal{J}_G} q_m \frac{dp_m}{dq_l} \right) \right] = 0.$$

The first order condition for k_j is:

$$\sum_m u_m^T T_{ml}^k - C_l^k - E_l^k + \lambda \left(\sum_{m \in \mathcal{J}_G} q_m \frac{dp_m}{dk_j} - \mathbf{1}_{l \in \mathcal{J}_G} \cdot C_l^k \right) +$$
$$\sum_l \frac{\partial q_l^*}{\partial k_j} \left[\frac{\partial U}{\partial q_l} + \sum_m u_m^T T_{lk}^q - C_l^q - E_l^q + \lambda \left(\mathbf{1}_{l \in \mathcal{J}_G} \cdot (p_l - C_l^q) + \sum_{m \in \mathcal{J}_G} q_m \frac{dp_m}{dq_l} \right) \right] = 0.$$

By taking its total derivative, the term $\sum_{m \in \mathcal{J}_G} q_m dp_m / dq_l$ can be written as $\sum_{m \in \mathcal{J}_G} (q_m \partial p_m / \partial q_l + \sum_n q_m \cdot \partial p_m / \partial t_n \cdot \partial T_n / \partial q_l) = \sum_{m \in \mathcal{J}_G} (q_m \Omega_{ml} + \sum_n q_m \cdot \partial p_m / \partial t_n \cdot \partial T_n / \partial q_l)$. Similarly, $\sum_{m \in \mathcal{J}_G} q_m dp_m / dk_l = \sum_{m \in \mathcal{J}_G} \sum_n q_m \cdot \partial p_m / \partial t_n \cdot \partial T_n / \partial k_l$. Substituting Lemma 2 into these two expressions, we obtain $\sum_{m \in \mathcal{J}_G} q_m \frac{dp_m}{dq_l} = \sum_{m \in \mathcal{J}_G} q_m \Omega_{ml} + \sum_m \tilde{u}_m^T T_{ml}^q$ and $\sum_{m \in \mathcal{J}_G} q_m \frac{dp_m}{dk_l} = \sum_m \tilde{u}_m^T T_{ml}^k$.

Substituting these two expressions as well as Lemma 1 into the first order conditions and rearranging terms yields the following two expressions:

$$\sum_l \frac{\partial q_l}{\partial p_j} \Bigg[(1+\lambda) \Big(p_l - C_l^q + \sum_m u_m^T T_{ml}^q - E_l^q \Big)$$
$$+ \lambda \Big(E_l^q - \mathbf{1}_{l \notin \mathcal{J}_G} \cdot (p_l - C_l^q) + \sum_{m \in \mathcal{J}_G} q_m \Omega_{ml} + \sum_m (\tilde{u}_m^T - u_m^T) T_{ml}^q \Big) \Bigg] = 0$$

$$\begin{aligned}&\sum_m (1+\lambda)\big(u_m^T T_{ml}^k - C_l^k - E_l^k\big) + \lambda\Bigg(E_l^k + \sum_m (\tilde{u}_m^T - u_m^T) T_{mj}^k\Bigg)\\ &\quad + \sum_l \frac{\partial q_l}{\partial k_j}\Bigg[(1+\lambda)\Bigg(p_l - C_l^q + \sum_m u_m^T T_{ml}^q - E_l^q\Bigg)\\ &\qquad\qquad + \lambda\Bigg(E_l^q - \mathbf{1}_{l \notin \mathcal{J}_G} \cdot (p_l - C_l^q) + \sum_{m \in \mathcal{J}_G} q_m \Omega_{ml} + \sum_m (\tilde{u}_m^T - u_m^T) T_{ml}^q\Bigg)\Bigg] = 0.\end{aligned}$$

Isolating p_j and $u_j^T \check{T}_{jj}^k$ and rearranging yields expressions (5) and (19). □

D.3 Generalizing Propositions 1 and 2 to multiple markets

Consider a city government that faces many markets m. The transportation system 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 is not 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, whose Lagrangian is

$$\begin{aligned}\max_{\sigma} U(\mathbf{q}(\sigma), T(\mathbf{q}(\sigma), \mathbf{k}(\sigma))) - C(\mathbf{q}(\sigma), \mathbf{k}(\sigma)) - E(\mathbf{q}(\sigma), \mathbf{k}(\sigma)) + \\ \lambda\left[\sum_{m, j \in \mathcal{J}_G} p_{mj}(\sigma) q_{mj}(\sigma) - C(\mathbf{q}(\sigma), \mathbf{k}(\sigma))\right], \qquad (20)\end{aligned}$$

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

The first-order condition for this Lagrangian is

$$0 = \sum_{mj} p_{mj} \frac{dq_{mj}}{d\sigma} + \sum_{nkmj} \frac{\partial U}{\partial t_{nk}} \frac{\partial t_{nk}}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} + \sum_{nkr} \frac{\partial U}{\partial t_{nk}} \frac{\partial t_{nk}}{\partial k_r} \frac{dk_r}{d\sigma} - \sum_{mj} \frac{\partial C}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} -$$
$$\sum_{mj} \frac{\partial E}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} - \sum_{r} \frac{\partial C}{\partial k_r} \frac{dk_r}{d\sigma} - \sum_{r} \frac{\partial E}{\partial k_r} \frac{dk_r}{d\sigma}$$
$$+\lambda \Bigg\{ \sum_{mjk,n \in \mathcal{J}_G} q_{nk} \frac{\partial p_{nk}}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} + \sum_{mjkol,n \in \mathcal{J}_G} q_{nk} \frac{\partial p_{nk}}{\partial t_{ol}} \frac{\partial t_{ol}}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} +$$
$$\sum_{j,m \in \mathcal{J}_G} p_{mj} \frac{dq_{mj}}{d\sigma} - \sum_{mj} \frac{\partial C}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} - \sum_{r} \frac{\partial C}{\partial k_r} \frac{dk_r}{d\sigma} \Bigg\}. \tag{21}$$

Suppose that σ is a price instrument, in which case $\frac{dk_r}{d\sigma}$ is equal to zero for all r. Then, following the ideas from Proposition 1 and after some algebra, this first order condition can be written as

$$p_j^\sigma = C_j^\sigma + E_j^\sigma - U_j^{A,\sigma} + M_j^\sigma + \frac{\lambda}{1+\lambda} \left\{ \mu_j^\sigma - E_j^\sigma - \Delta U_j^\sigma + \Delta M_j^\sigma \right\}, \tag{22}$$

where $w_{mj}^\sigma = \frac{\frac{dq_{mj}}{d\sigma}}{\sum_n \frac{dq_{nj}}{d\sigma}}$, $D_{kj}^\sigma = \frac{\sum_m \frac{dq_{mk}}{d\sigma}}{\sum_m \frac{dq_{mj}}{d\sigma}}$, $p_j^\sigma = \sum_m p_{mj} w_{mj}^\sigma$ $C_j^\sigma = \sum_m \frac{\partial C}{\partial q_{mj}} w_{mj}^\sigma$, $E_j^\sigma = \sum_m \frac{\partial E}{\partial q_{mj}} w_{mj}^\sigma$, $U_j^\sigma = \sum_{nkm} \frac{\partial U}{\partial t_{nk}} \frac{\partial T_{nk}}{\partial q_{mj}} w_{mj}^\sigma$, $\tilde{U}_j^\sigma = \sum_{mnkol,k \in \mathcal{J}_G} q_{nk} \frac{\partial p_{nk}}{\partial t_{ol}} \frac{\partial T_{ol}}{\partial q_{mj}} w_{mj}^\sigma$, $\mu_j^\sigma = \sum_{mn,k \in \mathcal{J}_G} q_{nk} \Omega_{nkmj}^\sigma w_{mj}^\sigma$, $M_j^\sigma = \sum_{k \neq j} D_{kj}^\sigma (C_k^\sigma + E_k^\sigma - U_k^\sigma - p_k^\sigma)$, $\tilde{M}_j^\sigma = \sum_{k \notin j} D_{kj}^\sigma (1_{k \in \mathcal{J}_G} (C_k^\sigma - p_k^\sigma) + \mu_k^\sigma - \tilde{U}_k^\sigma)$, $\Delta U_j^\sigma = \tilde{U}_j^\sigma - U_j^\sigma$, and $\Delta M_j^\sigma = \tilde{M}_j^\sigma - M_j^\sigma$.

This equation resembles very closely equation (5). To make this generalization, the key insight is that the relevant price, marginal cost, marginal externality, congestion effects, and diversion ratios are weighted averages of individual-market quantities across markets. The weight given to market m is $w_{mj}^\sigma = \frac{\frac{dq_{mj}}{d\sigma}}{\sum_n \frac{dq_{nj}}{d\sigma}}$: the extent to which a change in σ affects the number of trips in that market. Thus, the planner should put more weight on markets that are more affected by the policy instrument σ.

Based on this expression, one can find an explicit expression for a per-km road 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:

$$\tau = \frac{1}{r_j^\sigma} \left(E_j^\sigma - U_j^{A,\sigma} + M_j^{W,\sigma} \right), \tag{23}$$

where $r_j^\sigma = \sum_m r_{mj} w_{mj}^\sigma$ is the average distance per trip. This is a standard Pigouvian expression: the optimal price is equal to the average per-km externality plus the average per-km congestion effects and a diversion term. There is no 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

$$-\check{U}^{k,\sigma} = C^{k,\sigma} + E^{k,\sigma} - \hat{U}^{k,\sigma} + M^{k,\sigma} + \frac{\lambda}{1+\lambda} \left\{ -E^{k,\sigma} - \Delta U^{k,\sigma} + \Delta M^{k,\sigma} \right\}, \tag{24}$$

where $C^{k,\sigma} = \sum_r \frac{\partial C}{\partial k_r} \frac{dk_r}{d\sigma}$, $E^{k,\sigma} = \sum_r \frac{\partial E}{\partial k_r} \frac{dk_r}{d\sigma}$, $U^{k,\sigma} = \sum_{nkr} \frac{\partial U}{\partial t_{nk}} \frac{\partial t_{nk}}{\partial k_r} \frac{dk_r}{d\sigma}$, $\Delta q_k^\sigma = \sum_m \frac{dq_{mk}}{d\sigma}$, $\tilde{U}^{k,J,\sigma} = \sum_{nolr, k \in \mathcal{J}_G} q_{nk} \frac{\partial p_{nk}}{\partial t_{ol}} \frac{\partial T_{ol}}{\partial k_r} \frac{dk_r}{d\sigma}$, $M^\sigma = \sum_{k,m} \Delta q_k^\sigma (C_k^\sigma + E_k^\sigma - U_k^\sigma - p_k^\sigma)$, $\tilde{M}^\sigma = \sum_{km} \Delta q_k^\sigma (1_{k \in \mathcal{J}_G} (C_k^\sigma - p_k^\sigma) + \mu_k^\sigma - \tilde{U}_k^\sigma)$, $\Delta U^{k,\sigma} = \tilde{U}_t^{k,\sigma} - U_j^{k,\sigma}$, $\Delta M^{k,\sigma} = \tilde{M}_j^{k,\sigma} - M_j^{k,\sigma}$, and all other terms are defined as before. We decompose $U^{k,\sigma} = \check{U}^{k,J,\sigma} + \hat{U}^{k,\sigma}$ into the direct effect on travelers taking routes affected by σ due to waiting $\check{U}^{k,J,\sigma}$ as well as remaining effects $\hat{U}^{k,\sigma}$.

Once again, this equation resembles equation (19) very closely. Quantities are also aggregated across markets through a weighted average in which the weight given to market m is $w_{mj}^\sigma = \frac{\frac{dq_{mj}}{d\sigma}}{\sum_n \frac{dq_{nj}}{d\sigma}}$.

E Model Details

E.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 because the longer the gap between vehicles, the more riders arrive between them (k_{mj} is simply a normalization factor so the density integrates to one) . 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 $T_{mj}^{wait} = \int \frac{1}{2} t \cdot (t \cdot k_{mj} \cdot \phi(t))\, dt = \frac{1+\omega^2}{2k_{mj}}$.

E.2 Model of Waiting Times for Ride-Hailing and Taxis

Consider mode j (taxi or ride-hailing). Let q_{ahj} denote the number of trips originating in a during hour h, and I_{ahj} the number of idle drivers there. We assume a matching technology where the expected rider wait time is $T_{ahj}^{W} = A_{aj}^{W} I_{ahj}^{-\phi_j}$, where A_{aj}^{W} captures matching inefficiency in location a and ϕ_j is an elasticity governing how waiting times fall as idle drivers increase.[50]

Driver availability follows a parsimonious spatial model. Let L_{hj} be the total number of drivers in hour h, with busy drivers $B_{hj} = \sum_{od} T_{odh}^{\text{vehicle}} q_{odhj}$, where T_{odh}^{vehicle} 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 $I_{hj} = L_{hj} - B_{hj}$. The probability that an idle driver is in location a during hour h is given by

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

where $F_{ha} = \sum_b (q_{bahj} - q_{abhj})$ is the net inflow of mode-j trips into 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 depends on two terms. First, μ_a, which are fixed effects that capture driver's preferred areas. 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,

[50] This flexible formulation nests simple taxi and ride-hailing models. E.g., $\phi_j = 1$ in the taxi model of Lagos (2003); $\phi_j = 1/n$ in the n-dimensional ride-hailing model of Castillo et al. (2024).

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 and the equation governing T^W_{ahj} determine waiting times.

Estimation We first estimate the parameters A^W_{aj} and ϕ_j that map idle drivers into waiting times. For CA a, assume idle drivers I_{ahj} are uniformly distributed and pickup time conditional on distance is $t(x) = M_{aj} x^{c_j}$. The implied expected pickup time is[51]

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

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 estimate M_{aj} and c_j by regressing log travel time on log travel distance for all Google Maps car trips within the same CA, including CA fixed effects. We obtain $c_j = 0.614$ (s.e. = 0.0025), which implies $\phi_j = c_j/2 = 0.307$. A^W_{ahj} follows from equation (26) and from the fixed effects estimates.

We next estimate the driver location parameters (μ_a, λ, ρ). Since drivers are unobserved, we use Uber data on average waiting times T^W_{ahj} at the CA–hour level. Inverting equation (26) gives I_{ahj}, which we use to estimate (μ_a, λ, ρ) by maximum likelihood from equation (E.2). Because μ_a has 77 elements, we solve the problem with an inner loop that computes the optimal μ_a given λ and ρ using a contraction mapping (Berry et al., 1995), and an outer loop that maximizes over λ and ρ. We obtain estimates $\hat{\lambda} = 0.048$ (s.e. = 0.0003) and $\hat{\rho} = 0.799$ (s.e. = 0.028).

[51] With driver density I_{ahj}, the nearest-driver distance follows a Weibull distribution with pdf $2\pi x I_{ahj} e^{-\pi I_{ahj} x^2}$. Integrating $t(x)$ over this density yields equation (26).

E.3 In-vehicle time adjustment

Our congestion model predicts in-vehicle times very well for short trips but slightly overestimates long ones, likely because they tend to use more highways. To correct this, we regress the log ratio of Google Maps times T_{mj}^{vehicle} to the sum of travel times over edges, $\sum_{e \in P_{mj}} T_{ehj}^{\text{vehicle}}$, on straight-line distance d_m:

$$\log \left(\frac{T_{mj}^{\text{vehicle}}}{\sum_{e \in P_{mj}} T_{ehj}^{\text{vehicle}}} \right) = \alpha_j + \beta_j d_m + \epsilon_{mj}$$

We then scale the sum of travel times by $\psi_{mj} = \exp(\hat{\alpha}_j + \hat{\beta}_j d_m)$ in simulations.

E.4 Additional Parameters and Assumptions

Marginal costs. For car-based modes (taxis, ride-hailing, private cars), we use \$0.396 per km from the AAA cost of driving, plus labor costs of \$10 per hour for taxis and ride-hailing.

For buses, we combine four components. Capital costs are \$900,000 per bus lasting 250,000 miles, based on diesel and electric bus purchases by the Chicago Transit Board. Second, fuel costs are \$3.26 per gallon with a fuel efficiency of 3.38 mpg, which we take from National Transit Database (NTD) data for the CTA in 2020. Third, labor costs are \$33 per hour (NTD), assuming 20 km/h average speed and doubling to account for benefits and support staff. Finally, we use maintenance costs of \$2.76 per km (NTD). These numbers add up to \$7.528 per km.[52]

For trains, capital costs are \$11M per train lasting 2 million miles, based on CTA train purchases and assuming 10 cars per train. 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. Energy costs are \$0.07 per kWh (from a CTA report) with consumption of 5.88 kWh per mile (NTD). Labor costs are \$9.06 per km (operator CTA's expenses divided by

[52] We exclude road wear-and-tear externalities, which are negligible at \$0.0006–\$0.001 per km (Forkenbrock, 1999; OECD and ECMT, 2003; Small and Verhoef, 2007).

mileage), and maintenance is \$5.00 per km from the CTA's 2020 budget. Total costs equal \$17.73 per km.

As a sanity check, we compare our estimates with the CTA's 2019 financial statements. While these report all operating expenses rather than marginal costs and exclude capital, they provide bounds: \$5.17–\$12.51 per km for buses and \$9.07–\$40.38 per km for trains. Both ranges encompass our values.

Environmental externalities. For the social cost of carbon, we use \$190 per tonne.[53] For local pollutants, we follow Holland et al. (2016), using their Cook County estimates: 44.93¢ per gallon of gasoline (non-truck vehicles) and 41.32¢ per gallon of diesel (diesel trucks), aggregated by vehicle miles traveled. They report damages by vehicle type, which we aggregate for Cook County weighting by vehicle miles traveled. 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. For trains, we use an electricity consumption of \$5.88 kWh per mile (NTD) and Holland et al. (2016)'s environmental cost of \$0.111 per kWh for the Chicago electricity grid.

Vehicle occupancy. We assume average occupancies of 1.5 for private cars (Krile et al., 2019) and 1.3 for ride-hailing and taxis (Hou et al., 2020).

E.5 Equilibrium

Let $f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) \equiv q(\mathbf{p}, T(\mathbf{q}, \mathbf{k}))$ be a function that maps the feasible set of trip vectors $\mathcal{Q}$ into itself. An equilibrium, for $(\mathbf{p}, \mathbf{k})$, is a fixed point $\mathbf{q}^* \in \mathcal{Q}$ of this map:

$$\mathbf{q}^* = f^{\mathbf{p},\mathbf{k}}(\mathbf{q}^*).$$

After obtaining $\mathbf{q}^*$, equilibrium travel times can be computed as $\mathbf{t}^* = T(\mathbf{q}^*, \mathbf{k})$.

Proposition 3 (Existence)**.** *A fixed point $\mathbf{q}^* \in \mathcal{Q}$ of $f^{\mathbf{p},\mathbf{k}}(\cdot)$ exists.*

Proof. $f^{\mathbf{p},\mathbf{k}}(\cdot)$ is continuous because it is a composition of continuous functions. $\mathcal{Q}$

[53] See EPA Supplemental Proposal.

is a simplex, so it is a convex, compact set. Hence, $f^{\mathbf{p},\mathbf{k}}(\cdot)$ has a fixed point by Brouwer's fixed point theorem. □

To prove that there is a unique equilibrium, we follow Castillo (2025). We rely on the following assumption:

Assumption 1. $(f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) - f^{\mathbf{p},\mathbf{k}}(\mathbf{q}')) \cdot (\mathbf{q} - \mathbf{q}') < 0$ *for every* $\mathbf{q}, \mathbf{q}' \in \mathcal{Q}$.

Assumption 1 captures congestion forces (traffic congestion and ride-hailing externalities). When trip volumes rise in a mode–market ($\Delta\mathbf{q} = \mathbf{q} - \mathbf{q}' > 0$), travel times increase ($f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) - f^{\mathbf{p},\mathbf{k}}(\mathbf{q}') < 0$), discouraging further demand. Thus, higher flows raise costs, generating the negative correlation that Assumption 1 formalizes.

Although we cannot prove this assumption—some forms of congestion violate it—we show that, in our empirical model, thousands of random pairs $\mathbf{q}, \mathbf{q}' \in \mathcal{Q}$ always satisfied $(f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) - f^{\mathbf{p},\mathbf{k}}(\mathbf{q}')) \cdot (\mathbf{q} - \mathbf{q}') < 0$.[54]

Proposition 4 (Uniqueness)**.** *Under Assumption 1, $f^{\mathbf{p},\mathbf{k}}(\cdot)$ has a unique fixed point.*

Proof. Let $g_\gamma : \mathcal{Q} \to \mathcal{Q}$ be defined by $g_\gamma(\mathbf{q}) = (1-\gamma)\mathbf{q} + \gamma f^{\mathbf{p},\mathbf{k}}(\mathbf{q})$, where $\gamma \in (0,1)$. The set of fixed points of $f^{\mathbf{p},\mathbf{k}}$ and g_γ is the same. We now show that there exists some γ such that g_γ is a contraction mapping. This implies, by the contraction mapping theorem, that g_γ has a unique fixed point—and, hence, so does $f^{\mathbf{p},\mathbf{k}}$.

$f^{\mathbf{p},\mathbf{k}}$ is a continuous function with compact domain $\mathcal{Q}$ (a simplex), so it is uniformly continuous by the Heine-Cantor theorem. Thus, there exists $\beta < \infty$ such that $\frac{||f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) - f^{\mathbf{p},\mathbf{k}}(\mathbf{q}')||}{||\mathbf{q}-\mathbf{q}'||} < \beta$ for all $\mathbf{q}, \mathbf{q}' \in \mathcal{Q}$. This, in turn, means that, for all $\mathbf{q}, \mathbf{q}' \in \mathcal{Q}$, $||g_\gamma(\mathbf{q}) - g_\gamma(\mathbf{q}')||^2 = (1-\gamma)^2||\mathbf{q} - \mathbf{q}'||^2 + 2\gamma(1-\gamma)(f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) - f^{\mathbf{p},\mathbf{k}}(\mathbf{q}')) \cdot (\mathbf{q} - \mathbf{q}') + \gamma^2||f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) - f^{\mathbf{p},\mathbf{k}}(\mathbf{q}')||^2 < [(1-\gamma)^2 + \gamma^2\beta^2]\,||\mathbf{q} - \mathbf{q}'||^2$, where the inequality holds because of Assumption 1. Since β is bounded, the term in brackets is $1 - 2\gamma + O(\gamma^2)$, which is less than one for small-enough $\gamma > 0$. Thus, there exists some $\gamma > 0$ and some $\delta \in (0,1)$ such that $||g_\gamma(\mathbf{q}) - g_\gamma(\mathbf{q}')|| \leq \delta||\mathbf{q} - \mathbf{q}'||$ for all $\mathbf{q}, \mathbf{q}' \in \mathcal{Q}$. □

[54] The assumption is violated when two modes impose larger cross-externalities on each other than on themselves, which is unlikely in practice.

E.5.1 Equilibrium computation

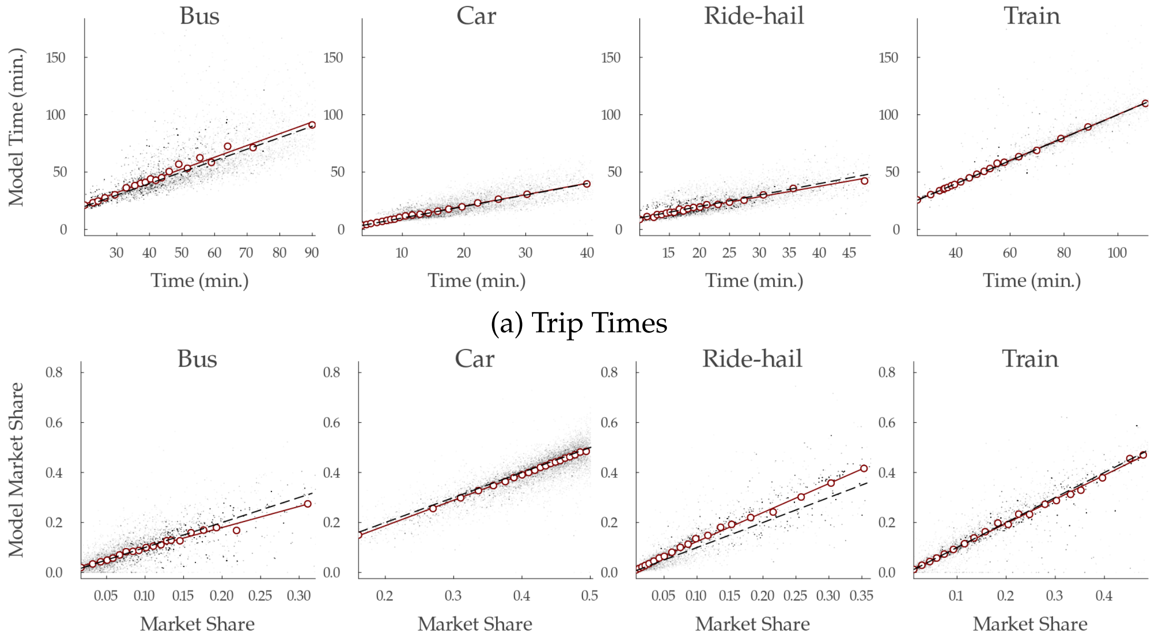
A naive approach is fixed-point iteration, but it typically diverges. Dampened fixed point iteration is guaranteed to converge—it is the same as applying fixed point iteration on g_γ, a contraction with the same fixed points as $f^{\mathbf{p},\mathbf{k}}$—but requires small γ and thus many iterations. Instead, we solve $f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) - \mathbf{q} = 0$ using a limited-memory Broyden's method. The full algorithm is:

```
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 previous algorithm. First, we approximate the inverse Jacobian $\mathbf{A}$ using the limited-memory method of Byrd et al. (1994). Second, if the updated vector $\mathbf{q}'$ is infeasible (due to insufficient ride-hailing or taxi drivers), we iteratively replace $\mathbf{q}'$ with $\mathbf{q} + \frac{1}{2}(\mathbf{q}' - \mathbf{q})$ until feasibility is restored.

E.6 Model Fit

Figure A6 shows that the trip times and market shares from our model fit the data well. The differences arise from our model of traffic congestion (Section 4.2), which predicts edge-level times perfectly but produces small discrepancies when times are aggregated to the path between origin and destination. The factor explained in Section E.3 corrects overall mismatches, but market-specific gaps persist.



(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.

E.7 Optimization

Having computed an equilibrium (Appendix E.5), we evaluate welfare $W(\mathbf{p}, \mathbf{k}) = W(\mathbf{q}^*(\mathbf{p}, \mathbf{k}), \mathbf{k})$ and city revenue $\Pi(\mathbf{p}, \mathbf{k}) = \Pi(\mathbf{q}^*(\mathbf{p}, \mathbf{k}), \mathbf{k})$. The unconstrained welfare maximization problem is

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

We solve it in two steps: (i) approximate the solution with Nelder–Mead (100 iterations, starting from observed prices and capacities), and (ii) refine with a quasi-Newton method. To reduce the computational burden, we deviate from Newton's method in two ways. First, we use the BFGS approximation (Nocedal and Wright, 2006) to the inverse Hessian. Second, we approximate gradients with central differences, taking only a few Broyden steps (typically three) at each evaluation.

With a budget constraint, the welfare maximization problem is

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

where B is the city's transport budget. We use an augmented iterative Lagrangian approach:

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

We initialize $\mu_0 = 10^{-6}$ and $\lambda_0 = 0$. At each step n, we maximize the objective with the method described above, update $\mu_{n+1} = 2\mu_n$, and set $\lambda_{n+1} = \lambda_n + \mu_n(\Pi^n + B)$, where Π^n is revenue at the step-n optimum. In this algorithm, λ_n converges to the value such that the budget constraint is satisfied with equality (Nocedal and Wright, 2006). This means that (29) converges to the true Lagrangian with a penalty term for deviations from the constraint, and the solutions converge to (28). To verify global optimality, we ran the optimization from hundreds of random initial points; in every case, the algorithm converged to the same solution.

Externality due to crowding As an additional check, we run counterfactuals incorporating bus and train crowding externalities. Following Hörcher et al. (2017), we scale in-vehicle time by $1 + 0.265s + 0.119d$, where s is the share of standing passengers and d their density per square meter. Table A1 reproduces Table 3 with this adjustment. Results remain qualitatively unchanged, though the planner chooses somewhat higher frequencies and slightly higher prices.

F Additional Results

F.1 Demand Robustness

To assess robustness, we estimate several alternatives (Table A2), always including dummies for multimodal trips and transfers. Column (1) adds market fixed effects: average VOT rises and the low–high income gap widens. Because of higher VOTs,

Table A1: Counterfactual results accounting for crowding externality

		Status Quo	Transit	Transit, Budget	Road Pricing	Transit + Road Pricing
		(1)	(2)	(3)	(4)	(5)
Panel A: Prices						
Avg. Price ($)	Bus	1.09	-0.21	1.77	1.09	0.41
	Train	1.33	-0.54	2.21	1.33	0.21
Road Tax ($/km)		0	0	0	0.33	0.33
Panel B: Wait Time and Frequencies						
Avg. Wait (min)	Bus	7.33	5.32	6.44	7.40	5.27
	Train	4.95	2.80	3.33	5.02	2.79
Δ Frequency	Bus	0%	37.96%	12.76%	0%	39.79%
	Train	0%	70.27%	43.25%	0%	70.87%
Panel C: Welfare						
Δ Welfare ($M/week)		0	5.28	2.15	3.20	8.24
Δ CS ($M/week)		0	27.74	2.62	-30.02	-6.44
Δ City Surplus ($M/week)		0	-20.93	0	29.41	12.38
Δ Transit Surplus ($M/week)		0	-20.93	0	0.66	-15.22
Road Taxes ($M/week)		0	0	0	28.75	27.60
Δ Externalities ($M/week)		0	0.02	0.34	-2.47	-2.20

Notes: This table presents a version of Table 3 where public transit passengers cause an externality due to vehicles getting crowded.

our optimal policy counterfactuals would also predict higher frequencies. Column (2) adds mode–destination fixed effects to net out factors such as destination-specific parking or access costs; the average VOT rises only slightly, at most, modest effects on optimal frequencies. Column (3) introduces an inner nest for public transit, which shows stronger substitution within transit relative to private modes; average VOT and the own-price/time elasticities remain essentially unchanged. Importantly, stronger bus–rail substitution implies that changing only bus prices or frequencies would mostly reallocate riders to rail (and vice versa). Because our counterfactuals move both bus and rail in parallel (or neither), these offsets would largely cancel, so optimal policies would not suffer large differences. Column (4) incorporates reliability by including the standard deviation of travel time for bus

Table A2: Demand Estimation Robustness

	(1)	(2)	(3)	(4)	(5)
Time (α_T)	-3.079	-2.774	-1.658	-1.954	-1.649
	(0.032)	(0.031)	(0.023)	(0.031)	(0.078)
Price (α_p)	-2.777	-1.488	-0.636	-1.030	-0.630
	(0.184)	(0.098)	(0.112)	(0.074)	(0.078)
Income (α_{py})	-0.646	-0.330	-0.108	-0.235	-0.086
	(0.032)	(0.033)	(0.089)	(0.037)	(0.063)
Nest (ρ)	0.359	0.196	0.231	0.463	0.399
	(0.015)	(0.013)	(0.014)	(0.011)	(0.011)
Time Std. Dev. ($\alpha_{std(T)}$)				0.488	
				(0.049)	
Time Het. (α_{Ty})					-0.132
					(0.103)
Time Income (λ_T)					0.065
					(1.719)
Inner Nest (ρ_{inner})			0.516		
			(0.016)		
Estimator	GMM	GMM	GMM	GMM	GMM
Market FE	✓	✓			
Mode-Dest. FE		✓			
Avg. VOT ($/h)	25.90	22.79	20.44	19.39	19.94
VOT (Bot. Quintile)	5.40	6.70	7.56	6.22	7.68
VOT (Top Quintile)	58.31	45.78	37.49	37.11	36.04
Avg. Price Elast.	-0.63	-0.54	-0.43	-0.59	-0.47
Avg. Time Elast.	-2.31	-1.77	-1.36	-1.61	-1.34
M	91,595	91,595	91,595	73,828	91,595
N	283,704	283,704	283,704	223,048	283,704

Notes: This table presents 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.

and rail. The estimated sensitivity to this variability is limited relative to sensitivity to mean travel time; implied average VOT and elasticities are close to baseline, yielding similar counterfactual prescriptions. Column (5) incorporates income-heterogeneous time preferences via a Box–Cox form: $\alpha_T^i = \alpha_T + \frac{\alpha_{Ty}}{y_i^{1-\lambda_T}}$, producing a slightly lower average VOT and a comparable inter-quintile range; price and time elasticities remain similar, suggesting that this additional heterogeneity does not generate substantively different substitution patterns.[55]

Taken together, average VOT remains stable across specifications—ranging from $19.39 to $25.90—and price/time elasticities are also similar. Given the small

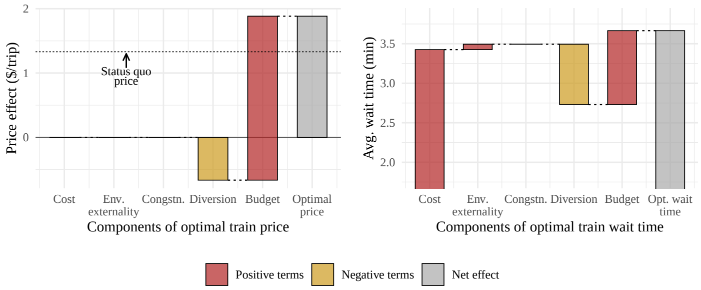
[55] We also instrument travel times interacted with income-quintile indicators.

differences, we expect counterfactual results to be similar. Thus, we retain our baseline specification for parsimony and transparent identification, while noting that these other specifications deliver qualitatively equivalent conclusions.

F.2 Decomposition of train prices and waiting times

Figure A7 is as Figure 8 but for trains.

Figure A7: 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).

Hörcher, D., Graham, D.J. and Anderson, R.J. (2017). Crowding cost estimation with large scale smart card and vehicle location data. *Transportation Research Part B: Methodological* 95:105–125.

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).

Weyl, G. and White, A. (2010). Imperfect platform competition: A general framework. *Net Institute Working Paper* .

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.[1]

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.[2] 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.[3] We define a stream of pings as a

[1] Source: My Daily Travel survey (website)

[2] 40 meters per second, i.e. about 145 kilometers per hour

[3] In particular, we only keep pings that satisfy the following two conditions: (i) no more than ten

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. 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.

- 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 S8 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.[4] Cellphones without an assigned home location (visitors) are not assigned an income at this stage.

Next, for each market, we estimate travelers' income distribution.[5] First, we take median income by tracts and divide tracts according to Chicago-level income quintiles.[6] 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. We end up, for each market, with the share of travelers in each of the five income quintiles, plus a share of visitors. For each market, we then reassign visitors proportionally across the five income quintiles so that their shares sum to

[4] We compute the census-tract median income percentile using the 2010 Census data.

[5] Recall, a market is defined as an (origin CA, destination CA, hour of the week)-tuple.

[6] For 2010, income quintiles are defined using the following cut-offs: \$34,875, \$46,261, \$60,590 and \$85,762.

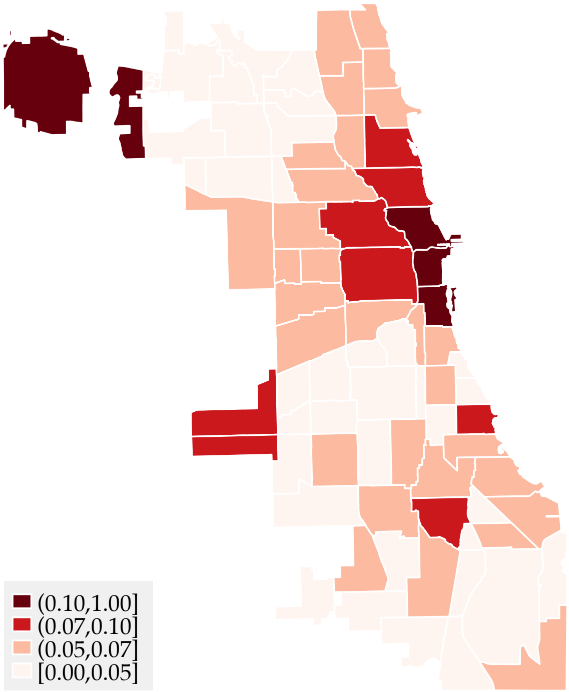


Figure S8: Share of visitors by origin location

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.

one. As a result, in the estimation, we work with five traveler types, corresponding to five income quintiles. 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.

S1.1.1 Survey Data Sparsity

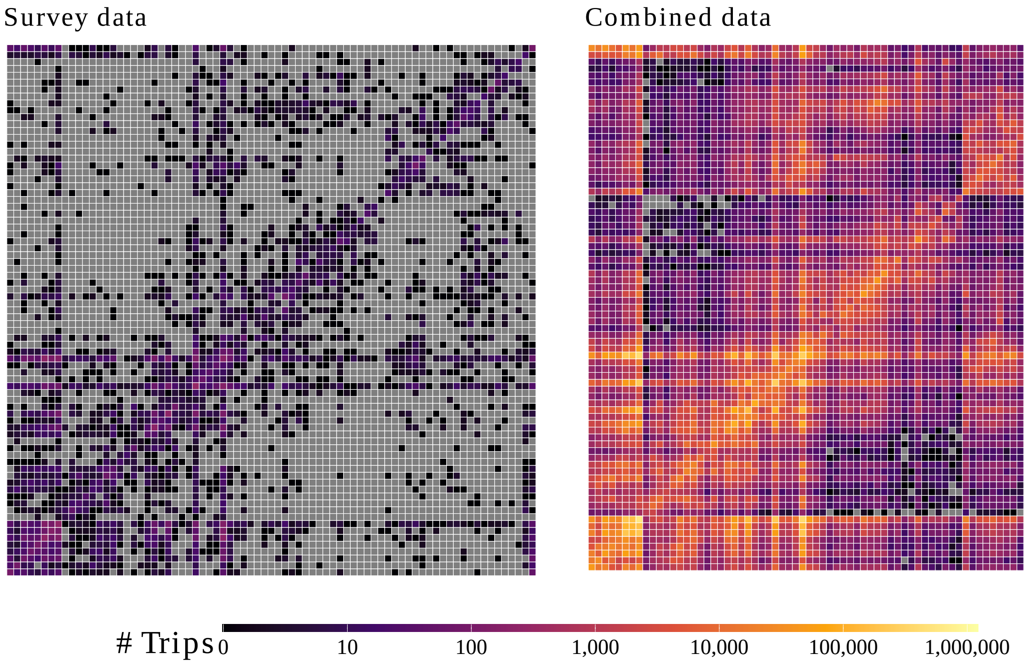


Figure S9: Combined vs. Survey Data: Flows Across Community Areas

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.[7] Importantly,

[7] 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

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 walk times from the origin latitude/longitude to the "best" train or bus station.[8]

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 September 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.[9]

Chicago Data City portal website.

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

[9] 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.

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 vary by the neighborhood of the trip's origin.[10] 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.[11]

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 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 were imputed by Zhao et al. (2007).[12]

These data notably exclude 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

[10] Each neighborhood is a group of about 8-9 CAs. The exact make-up of neighborhoods can be found on Wikipedia.

[11] Source: My Daily Travel survey (website)

[12] The inference relies on two observed patterns: a high percentage of riders begin their next trip at the destination of their previous trip, and many complete their final trip of the day at the same station where they began their first. These patterns were validated using travel diary data collected by the New York Metropolitan Transportation Council (Barry et al., 2002).

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 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.[13] 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 occu-

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

pancy by scaling trip numbers and prices using the average vehicle occupancy for that mode, which we report in Appendix E.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 cellphone trips.[14] 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.[15] 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.

[14] If the residual is negative we assume that there are no car trips.

[15] Source: AAA brochure "Your driving costs".

S2 Additional Results

S2.1 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 S10 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.

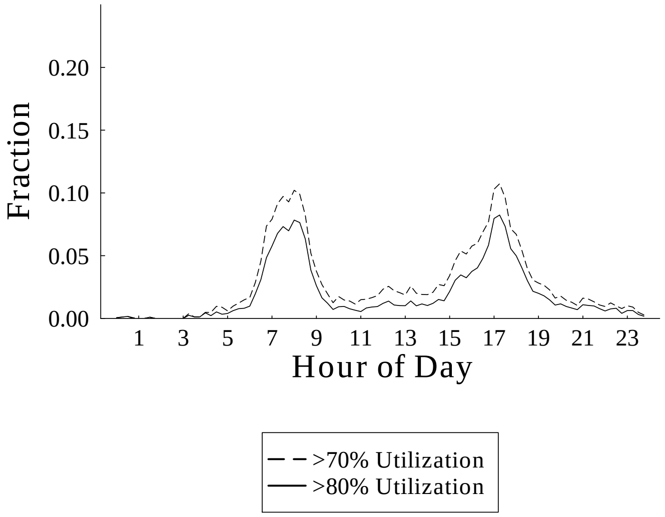


Figure S10: Bus Capacity

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

S2.2 Decomposition of welfare effects

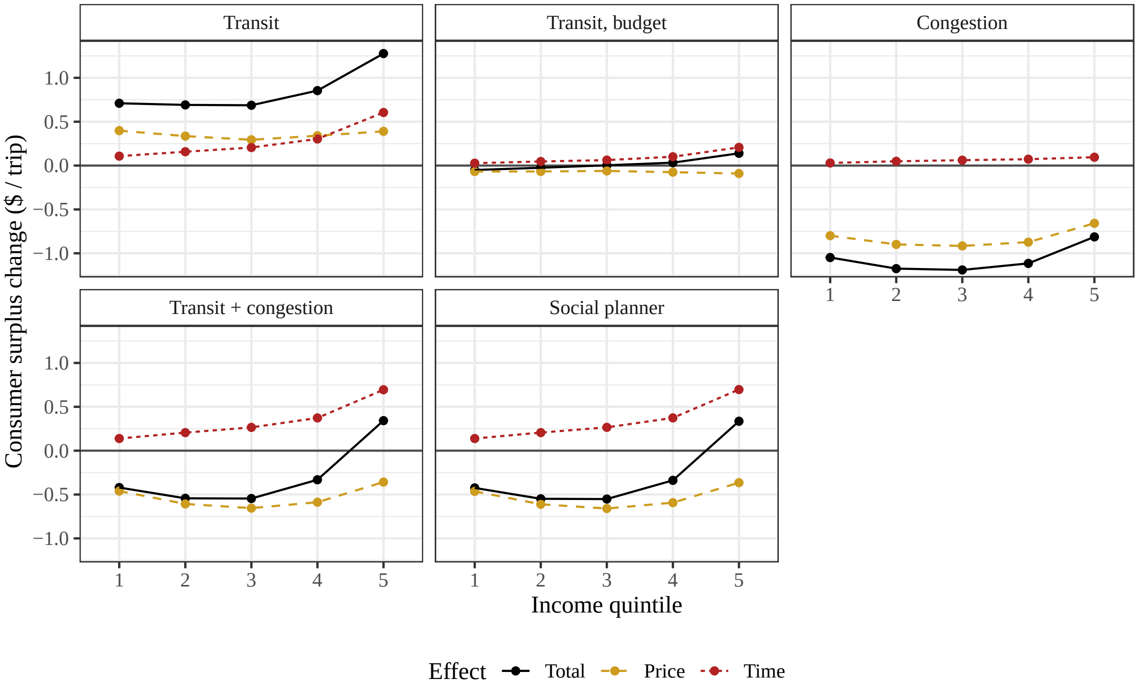
Table S3 decompose changes in consumer surplus into effects from prices and changes in environmental externalities into effects from frequencies and from travelers' substitution. Figure S11 shows how the decomposition of consumer surplus looks for different levels of income.

Table S3: Decomposition of Consumer Surplus and Environmental Externalities

		Status quo	Transit	Transit, budget	Road pricing	Transit + Road pricing	Social planner
Δ CS ($M/week)	Total	0	26.869	0.945	-32.320	-7.837	-8.021
	Price	0	14.475	-2.923	-34.746	-22.547	-22.776
	Time	0	12.394	3.869	2.426	14.710	14.755
	Capacity	0	9.756	3.764	0	10.069	10.056
	Substitution	0	2.637	0.105	2.426	4.641	4.699
Δ Externality ($M/week)	Total	0	-0.167	0.238	-2.717	-2.444	-2.455
	Capacity	0	0.398	0.122	0	0.452	0.450
	Substitution	0	-0.565	0.116	-2.717	-2.896	-2.905
Δ Avg. Speed (km/h)		0.00%	0.79%	-0.04%	1.98%	2.49%	2.50%

Notes: This table represents 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 S11: 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.