Optimal Urban Transportation Policy: Evidence from Chicago[†]

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

FEBRUARY 16, 2024

Abstract

We characterize optimal urban transportation policies and evaluate their welfare and distributional effects. We present a framework of a municipal government that implements different transportation equilibria through its choice of public transit policies — prices and frequencies — as well as road pricing. The government faces a budget constraint that introduces monopoly-like distortions. 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 consumers (travelers), whose surplus falls even if road pricing revenues are fully rebated. The largest losses are borne by middle income consumers, who are most reliant on cars. We find that the optimal price of public transit is close to zero and goes along with a reduction in the frequency of buses and an increase in the frequency of trains. Combining these transit policies with road pricing eliminates budget constraints. This allows the government to implement higher transit frequencies and even lower prices, in which case consumer surplus can increase with rebates.

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

Keywords: Urban transportation, Public Transit Subsidy Design, Road Pricing, Spatial Equilibrium, Ramsey Pricing.

[†] We thank seminar 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 Fall Meeting, Philadelphia Fed, UCLA IO/Spatial Conference, Berkeley, Toronto, Georgetown, Johns Hopkins, Maryland, Wharton Real Estate, UIUC, UNLP, Washington University in St Louis, St Louis Fed, IIOC, LSE, Oxford, Warwick, IFS/LSE/UCL IO workshop, Cowles Models and Measurement, SED, ERWIT/CURE, Sciences Po, Chicago-Princeton Spatial Conference, Stanford Cities Workshop and Northwestern Interactions Conference. We are grateful to Panle Jia-Barwick, Isis Durrmeyer, Matthew Freedman, Clara Santamaria, and Daniel Sturm for their insighful discussions. We also thank Steve Berry, Gilles Duranton, Jessie Handbury, Gabriel Kreindler, Jing Li, Jim Poterba, and Chris Severen for helpful comments. We are also grateful to Gilles Duranton and Prottoy Akbar for helping us obtain travel time data, and to Anson Steward, Jinhua Zhao, and Xiaotong Guo for helping us get access to CTA data. Melissa Carleton, Diego Gentile Passaro, and Shreya Mathur provided outstanding research assistantship. We acknowledge financial support from the NSF (Award No. 1559013, Supplement), through the NBER's "Transportation Economics in the 21st Century" initiative and from the John S. and James L. Knight Foundation through a grant to the University of Pennsylvania Center for Technology, Innovation & Competition.

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

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

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

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

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

1 Introduction

Since the 1950s, urban transportation in the U.S. has been characterized by the overwhelming use of personal cars, while public transit accounts for only 3.4% of the 850 million city trips that Americans undertake every day.[1] This current transportation landscape poses significant challenges for city governments as the economic cost of road-congestion in the U.S. alone is estimated at $87 billion annually, which are in addition to the significant environmental impacts.[2,3]

Cities' efforts to mitigate their effect on the climate and reduce inequality have led to a renewed discussion about the right mix of urban transportation policies.[4] Some argue that public transit should be cheaper, and indeed several municipalities have recently introduced free public transit.[5] Others suggest that, instead of reducing fares, cities should provide more frequent, higher-quality public transit.[6] Although these proposals appeal to different types of riders, it may not be feasible to pursue both because of stressed municipal budgets. By contrast, some cities have recently started taxing private forms of transportation. For example, London enacted a £15 congestion surcharge during the daytime and New York City recently approved a cordon tax below 60th street in Manhattan.[7] These forms of road pricing not only steer consumers away from private transportation towards public alternatives but also help financially strained municipalities to collect tax revenues. Importantly, all of the above policy levers cannot be evaluated in isolation — they are linked through mode substitution on the demand side, technologically through road congestion, and fiscally due to municipal budget constraints.

Given these interactions, what is the right mix 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? Are these two policy measures complements or substitutes in achieving cities' objectives? In this paper, we characterize the optimal mix of policies for a budget-constrained municipal government. Given prevailing difficulties to build new

[1] This is despite generous subsidies for public transit, which account, on average, for 3/4 of marginal trip costs. See https://www.bts.dot.gov. and newgeograghy.com

[2] See World Economic Forum— US Traffic Congestion Cost in 2018.

[3] A private car emits 0.96 pounds of CO2 per passenger-mile whereas public transit emits 0.45, even with low utilization rates.

[4] See Brookings — U.S. Transportation policy and HKS — Free Public Transit

[5] See NYT — "Should Public Transit Be Free? More Cities Say, Why Not?"

[6] See The Conversation — Low-cost, high-quality public transportation.

[7] For a comprehensive list of congestion taxes see https://ops.fhwa.dot.gov.

transit infrastructure in the US (Brooks and Liscow, 2021), we focus on three interventions that do not require new infrastructure: changing the fares of public transit, adjusting their service frequency, and road pricing.[8]

In order to measure the welfare effects of these interventions, we must account for several ingredients. First, we need to understand how people substitute across transportation modes as prices and travel times change. Second, we need to account for important forces that arise in transportation markets—traffic and environmental externalities as well as scale economies. Third, we have to account for the interaction of travel decisions across different locations, which are linked through roads that can become congested. To achieve this, we contribute to the literature both theoretically and empirically.

We first formulate a framework of a municipal government who wants to maximize welfare. The government chooses the prices and quality (in terms of frequencies) of different modes of transportation, subject to a budget constraint that accounts for operational costs, fare revenue, and taxes. Given these government choices, travelers in the market adjust and reach an equilibrium. We first find that an unconstrained social planner would set price minus marginal cost equal to the marginal externality—a canonical result in the theory of optimal taxation. However, our social planner deviates from this solution because budget considerations introduce two monopoly-like distortions. First, the planner charges markups that downwards-distort quantities. Second, given the ability to choose the service frequency of public transit, the planner directly affects the quality that travelers experience. Budget considerations distort quality towards the marginal consumer, as in Spence (1975).

Next, we move to an empirical application of this framework to Chicago, home to the second largest public transit system in the US.[9] Chicago serves as an ideal setting for our purposes for several reasons. First, public and private transportation are both important, with 69% of trips made by car and 25% by public transit.[10] Therefore, both changes in public transit provision and road pricing could be important policy levers. Second, Chicago is a city with large economic disparities, which means that it is important

[8] We focus on short-run adjustments, keeping residents' and firms' locations as well as infrastructure fixed. We thus measure welfare gains in the short run, before firms and residents readjust.

[9] In Chicago, during Q4 2019, the number of unlinked bus passengers was 237,276,500 and the number of unlinked heavy rail passengers was 218,467,000, which sum up to a total of 455,743,500 public transit passengers. Source: American Public Transportation Association (APTA), Public Transportation Ridership Report, Fourth Quarter 2019.

[10] The remaining trips are taxi and ride-hailing trips.

to account for the distributional effects of those policies.

Beyond these economic considerations, Chicago is also particularly good in terms of data availability. We combine several data sources to construct a rich dataset of travel flows, travel times, and prices for all relevant modes at a high temporal and spatial resolution. We have access to the near-universe of public transit through records from the CTA and the universe of ride-hailing trips, which are made public by the city of Chicago. One challenge we face is that there are no official records of private car trips. To overcome this problem, we determine mobility patterns from individual cellphone location records, which we use to construct the total number of trips in the city. We then recover the number of car trips by subtracting public transit and livery-vehicle (ride-hail and taxi) trips from the total number of trips.

Our data reveals that both mode choice and access to public transportation vary by travelers income levels and location. Car ownership is less common for both high and low income travelers relative to those with median incomes. These patterns together motivate a mode choice (demand) model that can account for these sources of heterogeneity.

The richness of our data allows us to designate granular transportation markets—people traveling from one community area (CA) to another during one hour of the week—and still conduct our analysis with aggregate market shares (Berry *et al.*, 1995).[11] This approach has the advantage that we can use standard inversion techniques to address endogeneity concerns. While the cost of operating a car and public transit prices are invariant to demand shocks, both the travel times of road-based modes and the prices of ride-hail are endogenous.

We instrument endogenous travel times using free-flow traffic speeds and argue that those capture long-run cross-sectional differences in infrastructure that are independent of the demand shocks that would bias our estimated model coefficients. To address the endogeneity of ride-hailing prices, we use price variation introduced by a surcharge that is levied on ride-hailing trips that either start or end downtown between 6AM and 10PM.[12] We construct indirect-inference moments that capture the deviation between the model-predicted demand response and the (difference in difference) treatment effect of the surcharge policy. The resulting model allows for heterogeneous substitution patterns across locations, income, and car ownership. Our estimates reveal substantial hetero-

[11] There are 77 community areas in the City of Chicago, with an average population of 35,600.
[12] The details of this policy are laid out here: https://www.chicago.gov/...congestion_pricing.html

geneity in the value of time across travelers and time of day, ranging from $4 to $45 per hour.[13]

Another key component of our model is a congestion technology that allows us to map traffic flows into speeds. To estimate the model, we exploit variation across hours of the day in travel speeds and in the number of vehicles traveling between adjacent Community Areas, following Akbar and Duranton (2017) and Kreindler (2023). We find an average elasticity of travel speed with respect to traffic flows between -0.12 to -0.19, comparable to existing estimates in the literature (Akbar and Duranton, 2017; Couture *et al.*, 2018). We find substantial heterogeneity across different areas of the city: traffic has a larger effect on travel times in central areas than in more peripheral areas.

We model wait times for public transit as a function of their frequency (the "fleet size" of buses and trains), taking into account the empirical frequency of delays and resulting changes in headway.

With our estimates of preferences and the congestion technology, we proceed to quantify the welfare effects of different transportation policies. We first explore the cases where the government only adjusts transit prices and frequencies or only implements road pricing. To explore the interactions between these two we then compute a counterfactual where the government does both.

We find that a planner who only controls public transit would like to reduce prices of both buses and trains by slightly more than fifty percent. The changes in frequency are different for buses and trains. The frequency of the former is reduced by 28% while that of trains increases by 10%. The main reason for this difference is that trains serve higher value of time consumers. Together, these changes lead to an increase in overall welfare by $1.45 million per week and an increase in consumer welfare by $740,000 per week with the remainder being due to the reduction in externalities. The consumer welfare gain under this policy are progressive and benefits lower income consumers more, mostly because they are very price elastic. [Say something about market shares].

The optimal road-price, when this is the only instrument of the planner, is 37 cents per kilometer. This leads to almost three times the welfare gains of the optimal transit policies, or about $4.2 million per week. However, most of the welfare gains come from a reduction in externalities equal to $3.2 million per week. Without rebating the resulting government

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

surplus, consumer surplus were to fall by $25.1 million per week. Thus, while this policy dominate on efficiency grounds, it hurts consumers. If the government were to fully rebate the surplus, most of those losses would be offset but consumers would still be worse off by $200,000 per week.

This policy leads to welfare gains of $0.9M per week. In a second scenario, the planner only controls prices on private cars. We find that optimal road prices should be of the order of $0.42/km, leading to an increase in welfare of $2.5M per week. Considering that the average commute trip distance is 16km, this tax on cars imply a daily congestion charge of $13.4 for commuters.[14] However, even though optimal road pricing return the largest efficiency gains relative to the status quo, they cause a large, regressive decrease in consumer surplus. When we put the two policies together, we find similar adjustments in prices and service levels, which lead to a total welfare gain of $2.5M per week. Finally, as a benchmark, we let the social planner set all prices including ride-hailing prices and the frequency of public transit. We find that she wants to reduce them by 45% because market power keeps prices high, implicitly acting as a Pigouvian tax.

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

Some papers analyze 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 search and matching model that he calibrates to the New York City taxi market. More recent empirical work includes Frechette *et al.* (2019) and Buchholz (2021), who also both model the New York City taxi market, as well as Brancaccio *et al.* (2020a), who model the dry bulk shipping industry. Castillo (2023) and Rosaia (2023) study ride-hailing platforms. Kreindler (2023) combines a structural approach with experimental evidence to study the effects of congestion on the well-being of travelers. Fuchs and Wong (2022) also study a multi-modal transportation model in the context of freight transportation. Like Brancaccio *et al.* (2020b), who derive optimal policies for frictional transportation markets, we derive them for urban transportation markets with a budget-constrained social planner.

Within this strand of literature, Durrmeyer and Martínez (2023) and Kreindler *et al.*

[14] For reference, the cordon price currently implemented in London is 15 sterling pounds.

(2023) are most closely related to our work. Durrmeyer and Martínez estimate an equilibrium model of mode substitution and investigate the welfare effects of private car restrictions and road pricing. We depart from Durrmeyer and Martínez mainly in two ways. First, we introduce a framework to design optimal government policies that allows for different modes of transportation to interact through various channels: consumer choices, the congestion function, and the budget constraint of the planner. Second, we also consider public transit policies, their potential interaction with road pricing, as well as the distributional implications of such interventions—for which our granular data is particularly well suited. Kreindler *et al.* (2023) study optimal transit policies but focus on the optimal network configuration for buses. While our policy simulations are less granular in terms of network planning, we incorporate the trade-off that the social planner faces when setting policies for both public transit and private modes of transportation through congestion surcharges.

There are several papers that investigate the long run effects of urban transportation policy, such as new rail infrastructure and interactions with residential location choices. Severen (2023) measures the effects on commuter welfare and productivity of the Los Angeles Commuter Rail. Tsivanidis (2023) quantifies the equilibrium effects of new public transit infrastructure in Bogotá. Brinkman and Lin (2022) study the welfare implications of highway construction during the mid-1900s. Fajgelbaum and Schaal (2020), Allen and Arkolakis (2022), and Bordeu (2023) characterize the optimal allocation of road infrastructure. These papers focus only on one mode of transportation and abstract away from incorporating rich heterogeneity in substitution patterns. We depart from their work by abstracting away from the residential and firm location choice models but allowing for rich demand substitution patterns across modes and heterogeneity, which is a key ingredient to understand distributional effects of transit policies. More recently, Barwick *et al.* (2021) explores the interaction between mode choices and residential location choices and Herzog (2021) quantifies the equilibrium effects of London's downtown tolls. Whereas these papers focus on individual policies—road pricing, driving restrictions, and subway expansion—we allow the government to choose the optimal combination among a portfolio of policies: public transit prices and frequencies, and road pricing.

We also build on a classic theoretical literature in transportation economics that develop models that capture the interaction between schedule constraints and congestion (Small, 1982; Arnott *et al.*, 1990, 1993; Small *et al.*, 2005). We enrich these models by com-

bining a congestion model with the demand approach used in industrial organization (Berry, 1994; Berry *et al.*, 1995), which allows us to model rider heterogeneity and account for the endogeneity of travel times and prices.

Finally, our work relates to several other works in empirical transportation economics. The estimation strategy we follow for the traffic congestion technology builds on the literature that measures the effects of traffic congestion (Akbar and Duranton, 2017; Akbar *et al.*, 2023; Couture *et al.*, 2018; Kreindler, 2023). This paper also relates to works that analyze different forms of road pricing. Hall (2018) shows theoretically that pricing some highway lanes ("Lexus lanes") can lead to Pareto improvements. Cook and Li (2023) empirically explore the distributional effects of dynamic tolling, and Yang *et al.* (2020) exploit the variation that is induced by driving restrictions in Beijing to derive the optimal road congestion surcharge. The last two papers abstract away from mode substitution and the interaction between public and private transportation. Parry and Small (2009) derive theoretical expressions for the optimal prices of public transit. 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 the many thousands of markets in a city. Lastly, Leccese (2021) and Leccese (2022) investigate the pass-through and the distributional consequences of the ride-hailing surcharge in downtown Chicago that we use to identify price elasticities.

2 Background and Data

2.1 Background

Chicago is the third largest city in the U.S. and its public transit system, which is operated by the Chicago Transit Authority (henceforth CTA), is the second largest after New York's. It includes a bus network of 152 routes with more than two thousand buses, and a train rapid transit system, which is known as the "Chicago L," that has eight routes and 144 stations Prices are per ride and independent of distance. The full fare for bus and the L are \$2.25 and \$2.50 respectively.[15] The CTA has a history of budget shortfalls, which suggests

[15] Reduced fares exist for students and seniors and a monthly pass for unlimited rides can be obtained at \$75. There are also daily, 3-day, weekly, and monthly passes. See CTA Fares for additional details.

that it is important to account for budget considerations (see, for instance, Ramos (2019)). Lastly, passengers may travel by private for-hire-vehicles (FHV) in the form of taxis and ride hail. Taxis have a regulated rate of $\$2.25$ per mile or $\$0.2$ per 36 seconds as well as a \$3.25 base fare.[16] Ride hailing companies adjust prices dynamically according to market conditions.

2.2 Data description

We define a market as an origin-destination-time tuple. Chicago's Community Areas (CAs) serve as our origin and destination locations. In total there are 77 CAs, with an average size of three square miles and an average population of 36,000 people.[17] We define a unit of time as an hour of the day, distinguishing between weekdays and weekends. Thus, we have 48 time periods. Multiplying the number of time periods by the number of possible origins and destinations, we obtain a total of 284,592 different markets. Our main dataset consists of travel flows, prices, and travel times for every mode in every market during January 2020.

To construct this dataset, we rely on a variety of raw data sources. First, we use administrative public transit microdata from the CTA. We have access to the near-universe of individual public transit trips for both buses and the Chicago L train. We observe the station or bus stop of origin, the time when the passenger tapped in, and an inferred drop-off L station or bus stop (Zhao *et al.*, 2007).[18]

The second data source, which is published by the City of Chicago, contains the universe of de-identified taxi and ride hailing trips.[19] It includes prices, drop-off and pick-up locations, trip length, trip duration, and the number of pooled riders in the case that the trip is a pool ride.

The third source is based on Veraset mobile-phone location data, which keeps track of people's movements over time.[20] This dataset records a device ID and a sequence of GPS coordinates and timestamps, for approximately 40% of active cell-phone devices in the

[16] See Chicago Taxi Fare Regulation.
[17] See Community Areas in Chicago.
[18] Unfortunately, we do not have data on Metra rides as this is managed by the Regional Transit Authority (RTA) rather than by the CTA. There were 74 million Metra rides in 2019 in Cook, DuPage, Kane, Kendall, Lake, McHenry, and Will counties combined, corresponding to less than 1% of trips in the region before the pandemic. See My Daily Travel and Annual/Monthly Ridership for details.
[19] Source: Chicago Data Portal, Transportation Network Providers - Trips
[20] Source: Veraset: Location Data Provider

US. We infer all motorized trips from the sequence of GPS coordinates for each individual device. Appendix A.3.1 describes this process in detail. The frequency with which records are generated depends on the applications installed by the user, so we restrict our analysis to devices with frequent location information. As we explain at the end of this section, this restriction results in a sample of trips and travelers that is representative across many dimensions. We thus multiply the cell-phone trips by a common inflation factor to arrive at the total number of trips.

We combine these three data sources to construct the near-universe of market flows across all modes: private car, taxi, ride-hailing, and public transit (buses and subway). While the CTA data allow us to observe the total number of trips for buses, trains, taxis, and ride-hailing, we do not have official records to directly measure car trips. Given that the cell-phone data covers all motorized trips, we can recover car trips by subtracting public transit, taxi, and ride-hailing trips from the cell-phone trips. Finally, we measure the overall size of the market as twice the number of trips that we observe in the cell-phone data for that specific market. We believe this margin of adjustment is important for our counterfactual simulations as some policies may induce travelers not only to switch across modes but also to stop or start traveling.[21] In what follows, we define a traveler's outside option as staying put or walking.

To observe the counterfactual travel times and routes that would have taken place if travelers had chosen a different mode, we use Google Maps data. We query all 30,796,848 possible trips by (origin census tract, destination census tract, hour of the week) triple, which we then combine with our flows data.

Our comprehensive dataset has several advantages over survey data. First, while survey data is designed to be representative at the city level, it often falls short of being representative at a more granular spatial or temporal resolution. Even worse, the coverage of surveys diminishes at higher resolutions, leading to sparse data. We show in Appendix A.5 that over 60% of the origin-destination CA pairs have zero trips—a problem that is exacerbated when we further break down the data by time period. This means that, although surveys are useful to understand aggregate movement patterns, they are ill-suited to studying fine-grained movement patterns or policy interventions that are het-

[21] For example, Rosenblum *et al.* (2020) shows that a $0.50 bus subsidy in Cambridge, MA increased ridership by 30%, without decreasing the shares of other modes for their surveyed population. Similarly, Brough *et al.* (2023) using experimental evidence show that the benefits of public transit subsidies primarily accrue from the creation of new trips to access services.

erogeneous across space and across time. Our detailed data also allows us to estimate the relationship between vehicle flow and traffic speed throughout the city. Lastly, the richness of our data also has econometric advantages. We can invert market shares (Berry (1994), Berry *et al.* (1995)) at a very granular level, which allows us to construct moment conditions based on high-resolution instruments to address the endogeneity of prices and travel times.

We add demographic information to our main dataset using census data from the 2016-2020 American Community Survey (ACS) sample. We use information on income and car ownership rates at the census tract level. We match devices to census tracts by inferring a device user's home location based on the modal GPS location during nighttime hours. We partition devices into two groups: those who spend at least three nights of the month in their modal night location, which we call *residents*, and those who spent at most two nights, which we call *visitors*.[22] Trips made by 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. Lastly, for validation of our data construction we use travel survey data from the Chicago Metropolitan Agency for Planning (CMAP) 2019 Household Travel Survey.[23]

Since the cellphone data only covers forty percent of devices, and among those we select the ones that provide sufficient location data, one may wonder whether some demographic groups are over- or underrepresented. We test the representativeness of our data based on census covariates assigned to devices by our inferred home location. We find that income groups are equally represented in the data (see Appendix A.1.3). We also test whether our cellphone trips are representative of travel patterns by comparing the distribution of the time and distance travelled in the cellphone data and in the survey data, and find that both aggregate distributions align quite well (Appendix A.1.4).

[22] Figure 13 in Appendix A.1.3 shows the average share of visitors by origin location. As expected, visitors are highly concentrated in the city center and in the airports.

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

2.3 Summary Statistics and Descriptive Results

We present descriptive evidence in four parts: First, we explore the characteristics and usage of different transportation modes.[24] Next, we analyze rider characteristics and how they correlate with mode choice. We then document evidence on the low utilization rates of buses. Finally, we present evidence of traffic congestion in the raw data.

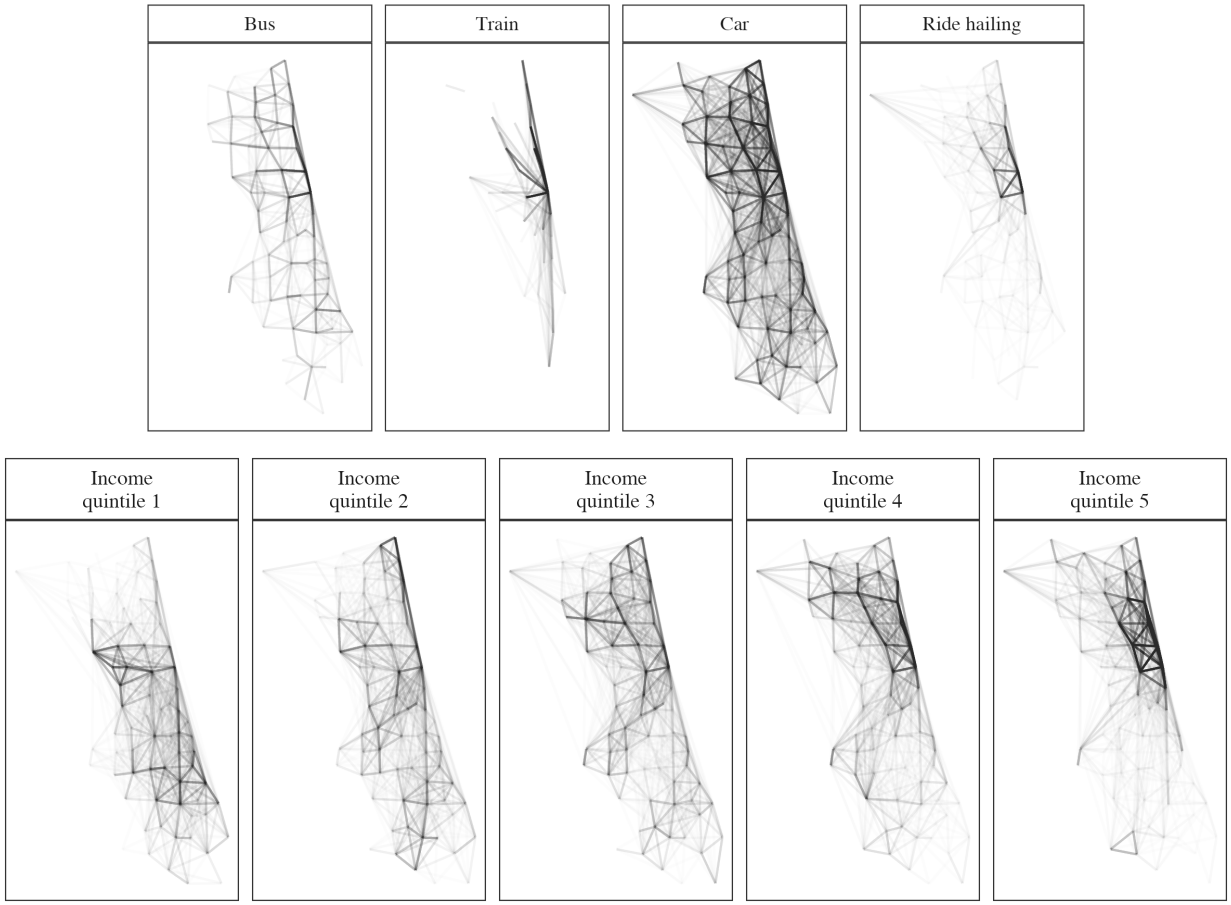


Figure 1: Trips by mode and by income

Notes: These figures show a random sample of 20,000 trips. A line connects the origin and destination of every trip. The top figure splits trips by mode. The bottom figure splits them by the income quintile of the traveler.

Although Chicago has one of the most extensive public transit systems in the US, about 65% of trips are taken by car. Public transit accounts for 30% of trips, with buses

[24] We exclude biking but these modes only represent 1.8% and 2% of the overall trips.

taking slightly more than half of this share. Ride-hailing accounts for 4% of trips. Lastly, taxis accounts for 0.5% of trips—and, hence, we exclude them for most of our analysis. Figure 1 shows that the characteristics of trips differ across modes. Bus and car trips are spread throughout the city. Train trips are concentrated along the corridors connected by the L lines. Ride hailing mostly accounts for short trips downtown or north of downtown, 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 has stark income differences that 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.

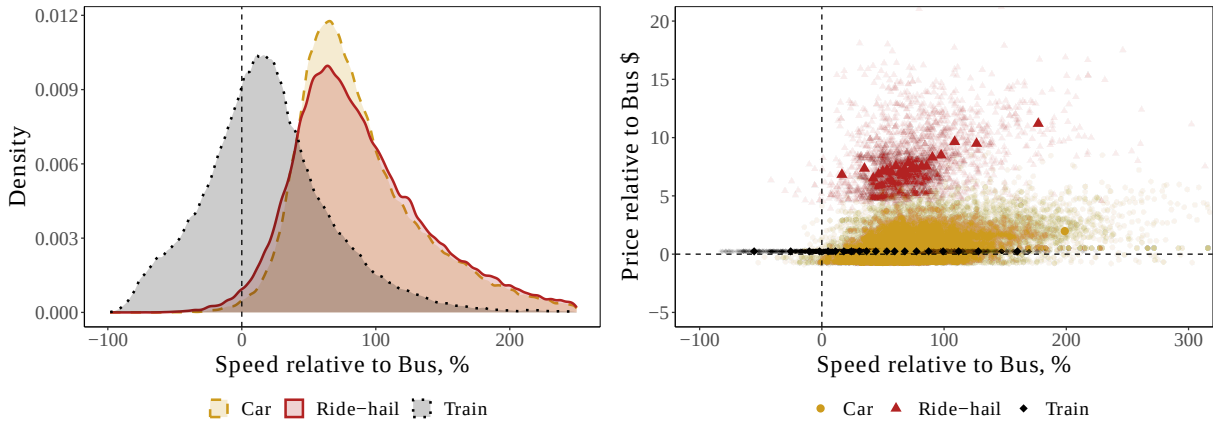


Figure 2: Speed and price differences across modes of transportation

Notes: The left panel reports density of speeds across four modes of transportation. The right panel presents scatter and binscatter plots of prices and speed by mode. Observations are at the market level, weighted by the total number of trips in the market.

Figure 2 shows the differences in speed, travel time, and prices across modes. The left panel shows the distributions of the speed of all modes relative to the speed of buses. Trains are, on average, 10% faster than buses, and cars and ride-hailing are on average almost twice as fast as buses. The right panel shows that buses are also the cheapest option for most trips, but car prices are lower for some trips, which are mostly short, inexpensive trips with a low price. We can also see that almost all of the comparisons with respect to buses lie on the upper-right quadrant, indicating that modes that are faster

also tend to be more expensive relative to travelling by bus.[25] This price-speed trade-off will therefore be a key margin of decision for our travellers when choosing their preferred mode of transportation.

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

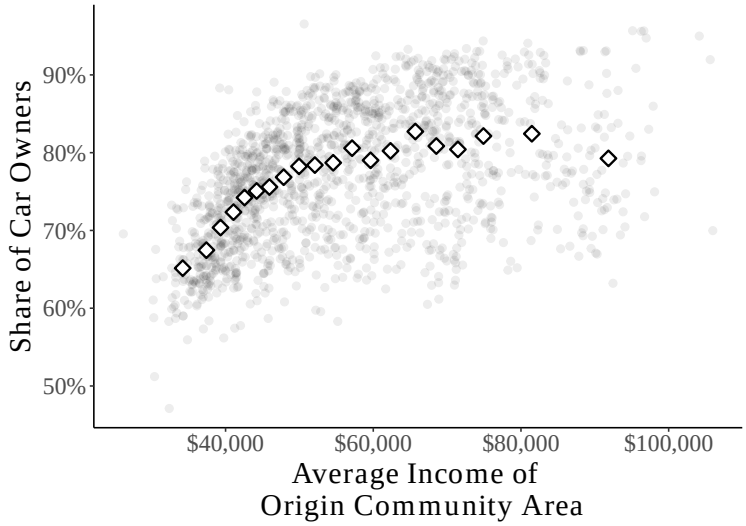


Figure 3: Car ownership by travelers' income

Notes: This figure plots a scatterplot and a binscatter of car ownership against median income. Our level of observation is a census tract.

Figure 4 shows how mode choices vary by travelers' income, which we proxy using the average income of the origin CA. Lower income travelers are more likely to travel by bus. Perhaps surprisingly, higher income people are more likely to use trains than lower income people (second panel): as we can see in Figure 1, train trips are concentrated along L lines, which tend to be located in higher income areas. Consistent with car ownership patterns, car usage follows an inverted-U shape: middle income people are most likely

[25] The positive relationship between speed and price arises from the fact that longer trips—those that are more expensive by construction—are also more likely to travel through highways rather than secondary roads.

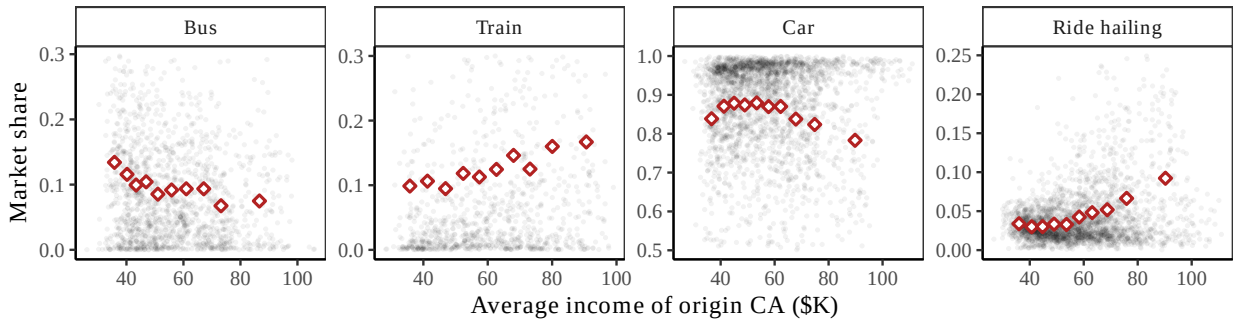


Figure 4: Mode market shares by travelers' income

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

to use cars—and, thus, they are likely to be hurt the most by road pricing. Finally, ride-hailing is overwhelmingly used by the highest income people.

We now turn to data on the occupancy of buses. Figure 5 shows that even though buses are the cheapest available mode of transportation, many of them are running empty. Even during the morning and afternoon rush hours, median utilization rates stay below 20%. If we look at the distribution within hour, we see that 90% of buses are at a utilization below 75% during the morning rush hour. Moreover, this graph also shows that, while certain buses may travel at full capacity at peak times, those events are extremely rare. Only the 99th percentile bus reaches full capacity at any time during the day, and only during the morning and afternoon rush hours. These low utilization rates suggest that resources may not be optimally allocated, implying that there is scope for the city government to optimize on the supply of public transit.

Lastly, we present raw data patterns that show how traffic congestion impacts travel times. We examine the relationship between the number of vehicles moving between pairs of adjacent CAs and travel times. The left panel of Figure 6 shows this relationship, after residualizing on market fixed effects. The right panel shows two separate markets. In both panels, the data exhibit a "hockey-stick" pattern: the trend is flat at lower vehicle counts and then, past a certain threshold, it begins to rise at a rate that is close to linear in log space. This means that when there are not many cars on the road, speeds are close to constant. However, beyond a certain traffic volume, additional vehicles decrease travel speeds with an approximately constant elasticity. These patterns guide our selection of

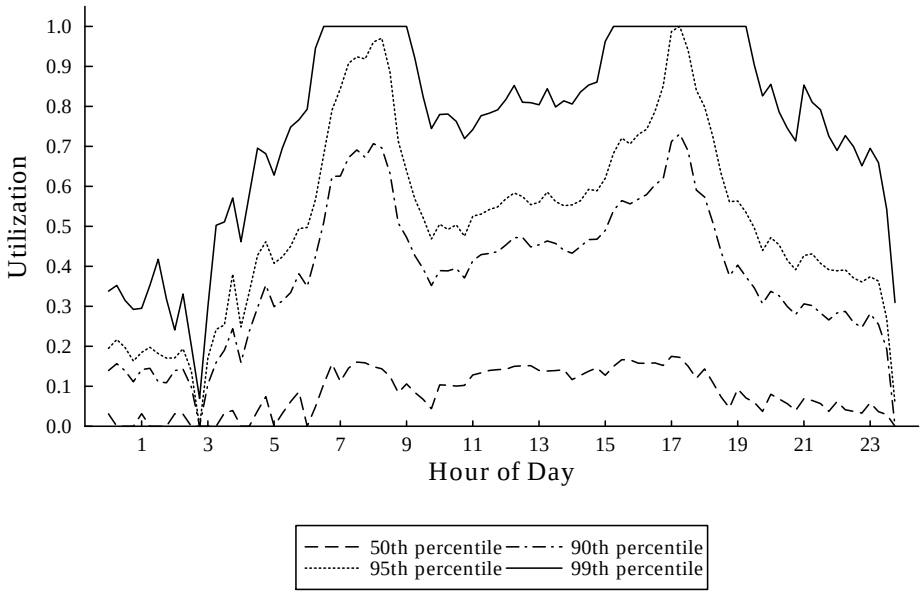


Figure 5: Bus utilization rates

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.

the congestion function's form and the instrument that we use for travel times when estimating demand.

3 Model

Our model consists of three parts. First, there are travelers with fixed origin and destination who choose either one of the available modes or not to travel at all. Second, there is a transportation technology that captures the relationship between the number of people who use a mode and its travel time. Third, there is a social planner that maximizes welfare subject to a budget constraint.

3.1 Setup and Equilibrium Definition

We first present a simple version of our model, in which we focus on only one market and describe traveler preferences and the transportation technology in the abstract. This

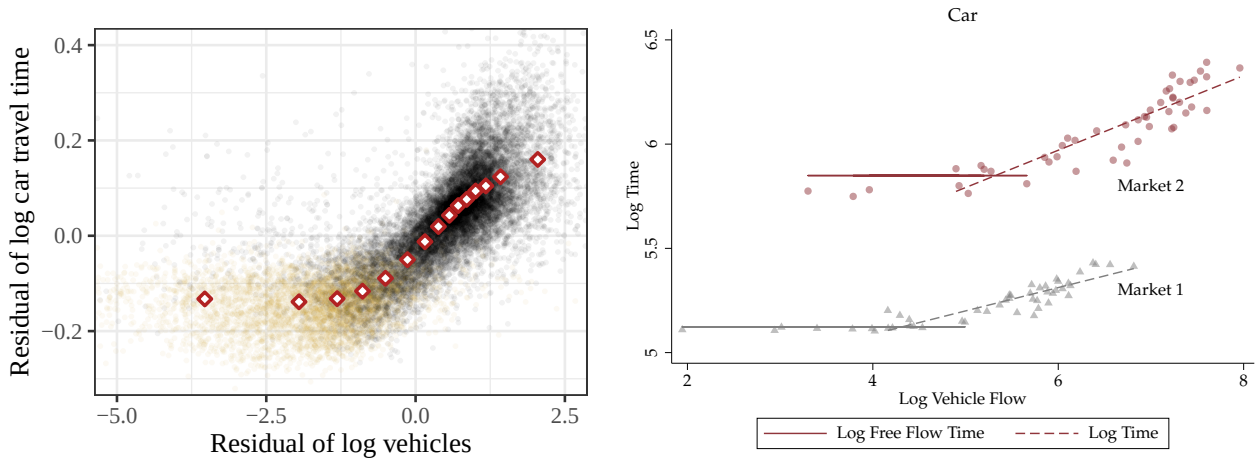


Figure 6: Congestion Graphs

Notes: These figures present the relationship between travel times and the number of vehicles, using data at the level of a pair of adjacent community areas during an hour of the week. Yellow points represent observations between midnight and 5 am. The left panel shows log-travel times against the logarithm of of vehicles on the road, where both variables are residualized by market fixed effects. The right panel singles out two random markets and shows the same relationship.

will allow us to present theoretical results in Section 3.2 that highlight the general forces in the social planner's objective function. In Section 3.3 we present the empirical version of our model, which accounts for inter-temporal and spatial variation in supply and demand as well as for for the spatial linkages across markets due to congestion and driver movements. In Sections 3.3.1 and 3.3.2 we provide the details on how we model traveler demand and the transportation technologies for different modes.

The first component of our model are travelers, who differ across two observable dimensions: their income and whether they own a car. We denote a traveler's type by $\theta \in \mathbb{R}^n$, with density $f(\cdot)$, which captures observable characteristics as well as unobservable preferences for modes.

A traveler of type θ decides which transportation mode j to take to her destination. She can choose among the set $\mathcal{J}(\theta)$, which varies depending on whether public transit is easily accessible and on whether she owns a car. She can also choose the outside option of not taking a trip, which we denote by 0. Thus, her choice set is given by $\mathcal{J}(\theta) \cup \{0\}$. She 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.[26] We normalize the utility of the outside option to zero. Therefore, $u_j(t_j, \theta) - p_j$ is the utility measured relative to not taking a trip.

The traveler chooses the mode that maximizes her utility among her choice set:

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

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 choosing mode j at $(\mathbf{p}, \mathbf{t})$.

We refer to the vector $\mathbf{q}$ as trips. Gross consumer utility and consumer surplus are given by

$$U(\mathbf{p}, \mathbf{t}) = \sum_j \int_{\Theta_j(\mathbf{p},\mathbf{t})} u(t_j, \theta) f(\theta)\, d\theta \quad \text{and} \quad CS(\mathbf{p}, \mathbf{t}) = \sum_j \int_{\Theta_j(\mathbf{p},\mathbf{t})} (u(t_j, \theta) - p_j) f(\theta)\, d\theta. \tag{3}$$

Travel times are determined by a transportation technology that captures the dependence 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 how congested roads are.

Accounting for all these considerations, we can compactly write the vector $\mathbf{t}$ of travel times for all modes as

$$\mathbf{t} = T(\mathbf{q}, \mathbf{k}), \tag{4}$$

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 phys-

[26] Observe that we can also incorporate heterogeneity in traveler's sensitivity to prices $u_j(t_j, \theta) - \theta_p \cdot p_j$. Note that a re-scaled version of utility, namely $(u_j(t_j, \theta) - \theta_p \cdot p_j)/\theta_p$ leads to the same optimal choices. Using $u_j(t_j, \theta) - p_j$ has the advantage of measuring utility directly in monetary terms for all consumers $\theta \in \Theta$, where the Value of Time (VOT) can be computed as $\partial u_j(t_j, \theta)/\partial t_j$.

ical costs, such as fuel and vehicle depreciation. Additionally, there is an environmental externality $E_j(q_j, k_j)$ that is borne by society. We assume that both functions are increasing and convex in both arguments.

With this notation we can now define an equilibrium.

Definition 1 (Transportation Equilibrium)**.** *Given a set of prices* $\mathbf{p}$ *and fleet sizes* $\mathbf{k}$*, a market equilibrium is a set of trips* $\mathbf{q}$ *and travel times* $\mathbf{t}$*,* $\left(\mathbf{q}^*(\mathbf{p}, \mathbf{k}), \mathbf{t}^*(\mathbf{p}, \mathbf{k})\right)$ *such that* (2) *and* (4) *hold.*

Travel times and quantities are the equilibrium objects. Prices and fleet sizes are exogenous and the social planner is able to determine a subset of them. Therefore, for any given set of fleet size and prices, travel times adjust to bring the market into equilibrium.

3.2 The Social Planner's Problem

The city government chooses the prices and fleet sizes of buses and trains as well as the cost of traveling by car through a congestion surcharge. We denote these modes as $\mathcal{J}_G$. The city government's goal is to maximize welfare subject to a budget constraint.

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

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

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

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

If we let $\mathbf{p}_G$ and $\mathbf{k}_G$ be the vectors of prices and capacities set by the government, the objective function is:

$$\begin{aligned} \max_{\mathbf{p}_G, \mathbf{k}_G} \quad & U(\mathbf{q}, T(\mathbf{q}, \mathbf{k})) - C(\mathbf{q}, \mathbf{k}) - E(\mathbf{q}, \mathbf{k}) \\ \text{s.t.} \quad & \sum_j p_j(\mathbf{q}, T(\mathbf{q}, \mathbf{k})) \cdot q_j - C(\mathbf{q}, \mathbf{k}) \geq -B \end{aligned} \tag{5}$$

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

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

Proposition 1. *Prices under the solution of the social planner's problem* (5) *are given by:*

$$p_j = \overbrace{C_j^q + E_j^q}^{\text{Mg. cost and env. externality}} - \overbrace{\sum_l u_l^T \cdot T_{lj}^q}^{\text{Network effects}} + \overbrace{M_j^q}^{\text{Diversion}} + \frac{\lambda}{1+\lambda} \cdot \Bigg(\underbrace{\sum_{k \in \mathcal{J}_G} q_k \cdot \Omega_{kj}}_{\substack{\text{Market power} \\ \text{markup}}} + \underbrace{\tilde{M}_j^q - M_j^q}_{\substack{\text{Diversion} \\ \text{distortion}}} - E_j - \underbrace{\sum_l (\tilde{u}_l^T - u_l^T) \cdot T_{lj}^q}_{\substack{\text{Spence} \\ \text{distortion}}} \Bigg) \quad (6)$$

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 the total time if pickup times for mode j increase by 1%, and M_j^q and $\tilde{M}_j^q$ (which we discuss below) are defined as:*

$$M_j^q \equiv \sum_{k \neq j} D_{kj} \left(C_k^q + E_k^q - \sum_l u_l^T \cdot T_{lk}^q - p_k \right) \quad (7)$$

$$\tilde{M}_j^q \equiv \sum_{l \in \mathcal{J}_G \setminus j} D_{lj} \left(C_l^q + \sum_{k \in \mathcal{J}_G} q_k \cdot \Omega_{kj} - \sum_m \tilde{u}_m^T \cdot T_{ml}^q - p_l \right). \quad (8)$$

Proof. See Appendix C.1 □

To understand expression (6), think first of an unconstrained social planner, in which case $\lambda = 0$ and the final term drops out. Prices are equal to the Pigouvian solution—in addition to the direct costs, they are set to internalize environmental externalities, network effects, and an additional term that we call the *diversion term*, which we explain below.

Network effects are equal to the sum over modes of the product of $\bar{u}_k^T$, the derivative of gross utility with respect to time, and $\tilde{T}_{kj}$, the change in total time given an additional

trip using mode j. For $j \neq k$, $\tilde{T}_{kj}$ is positive due to congestion (or zero if the two modes are not interrelated), so those terms lead to a Pigouvian tax. But $\tilde{T}_{jj}$ can be negative. For livery vehicles, for instance, increasing the number of trips and fleet size by the same factor results in a reduction in waiting times due to returns to scale (i.e., economies of density) in the matching process. In that case, there should be a Pigouvian subsidy, as noted by Arnott (1996).

We now explain the diversion term M_j^q. It accounts for the fact that modes that are not priced by the planner are suboptimally priced. If there is no road pricing, for instance, the price of driving a car will be too low and too many people will drive their car. As a second best, the government would want to lower the price of public transit to induce substitution away from cars. In the expression for M_j^q (equation 7), the term in brackets can be thought of as deviations of prices from a standard Pigouvian solution—marginal costs plus marginal externalities plus network effects. M_j^q is therefore the diversion-ratio weighted sum of these deviations for all other modes. It captures the extent to which an increase in the price of j induces substitution towards modes that are overpriced and, thus, are chosen by fewer travelers than is socially optimal. This form of substitution increases welfare if the diversion term is positive, and, hence, the optimal price is higher. In particular, note that this term is zero whenever all other modes are already priced at the Pigouvian solution.

We now also take into account the budget constraint. To stay on budget, the social planner behaves qualitatively like a monopolist because it needs to raise revenue. This introduces a market power markup in equation (6). For the same reason, the diversion term is distorted towards its revenue-motivated equivalent $\tilde{M}_j^q$. It is very similar to the social welfare-motivated diversion term M_j^q, except that it captures the extent to which an increase in the price of j induces substitution towards modes that are chosen by fewer travelers than is optimal *to maximize government revenues*, rather than to maximize social welfare.[27] The social planner now also under-weights environmental externalities. Finally, there is a Spence distortion: the government internalizes effects on other travelers' utility, but imperfectly: it accounts for changes in the utility of marginal travelers, rather

[27] In this term, instead of the effect of a price change on all other modes, the city now only considers modes $\mathcal{J}_G$ from which it can raise revenue. And the term in brackets now represents deviations in prices from the revenue maximizing prices. It no longer includes the marginal externality but instead includes a markup term $\sum_{k \in \mathcal{J}_G} q_k \cdot \Omega_{kj}$. Furthermore, rather than considering the effect of travel times on the average traveler u_m^T, the city now cares about the marginal traveler $\tilde{u}_m^T$.

than that of all travelers. As $\lambda \to \infty$, the social planner no longer cares about welfare and it only cares about its budget. These two expressions then become the first order conditions for profit maximization: terms related to environmental externalities cancel out, and there is a full markup and a full Spence distortion.

In Appendix C we derive a similar decomposition for travel times, which follows the same logic as the price decomposition.

In our counterfactuals, we will come back to these results and empirically decompose the optimal prices and travel times into these constituent parts to explain the planner's competing motives when setting optimal prices and travel times.

3.3 Empirical Model

Here we describe the empirical version of our demand model and of the transportation technology, which we use for $U(\mathbf{p}, \mathbf{t})$ and $\mathbf{t} = T(\mathbf{q}, \mathbf{k})$ in the above expression. The city is composed of Community Areas $a \in \mathcal{A}$, the main level of spatial aggregation we consider. We will also refer to them as *locations*. Time is divided into hours $h \in \mathcal{H}$.[28]

3.3.1 Demand

First, we define a market $m = (a, \tilde{a}, h) \in \mathcal{M}$ as the collection of people who make travel decisions from Community Area a to Community Area $\tilde{a}$ at a particular time h.[29] In each market there is an exogenous number of potential travellers λ_m. They decide which mode $j \in \mathcal{J}_m^i \cup \{0\}$ to use, where $j = 0$ denotes the outside option of staying put, by solving the following problem:

$$U_m^i = \max_{j \in \mathcal{J}_m^i \cup \{0\}} \delta_{mj}^i + \varsigma_{mg(j)}^i + (1-\rho)\epsilon_{mj}^i \quad = \\ \max_{j \in \mathcal{J}_m^i \cup \{0\}} \quad \xi_{mj} + \alpha_T \cdot T_{mj} + \alpha_p^i \cdot p_{mj} + \varsigma_{mg(j)}^i + (1-\rho)\epsilon_{mj}^i \quad (9)$$

[28] Specifically, we define h as hour-of-the-week. We consider 48 = 2×24 hours-of-the-week, corresponding to an average weekday day and an average weekend day. Daily variation is aggregated by taking averages across dates for the same hour-of-the-week.

[29] Given that our level of temporal variation averages across dates, traveling decisions should be thought as the choice over average trips at a given hour-of-the-week h rather than stemming from very short-run temporal variation coming from shocks or special occasions.

where T_{mj} denotes the sum of the waiting and travel times for mode j, p_{mj} is the price for mode j, α_T is the preference parameter over travel times, and α_p^i is the person i-specific price coefficient.[30] Motivated by the disparities in mode choice across the income distribution from Section 2.3, we allow α_p^i to vary by income y_i according to $\alpha_p^i = \alpha_p / y_i^{1-\alpha_{py}}$, so that α_{py} captures the extent to which the price coefficient varies with income.[31] Differences in this coefficient then capture differences in the marginal utility of money across the income distribution, which lead to heterogeneity in the trade-off between time and money.

The parameters $\varsigma_{mg(j)}^i$ and ϵ_{mj}^i are idiosyncratic unobserved taste shocks. The taste shock $\varsigma_{mg(j)}^i$ is common to all goods within the group $g(j) \equiv \mathcal{J}_m^i$, defined as all modes of transportation excluding the outside option. This error allows the unobserved utility shocks of all inside options to be correlated. The taste shock ϵ_{mj}^i is specific to mode j. This model can be thought of as a sequential choice problem: First, travellers choose whether they want to travel or not, and second, conditional on travelling, travellers choose which mode they use.

We allow the choice sets $\mathcal{J}_m^i$ to vary across markets and across consumers within the same market to capture that some modes of transportation may not be available within a market or to specific consumers. For instance, some Community Areas cannot be reached by train and some consumers do not own a car.

We assume that ϵ_{mj}^i follows a Type I Extreme Value distribution and that the iid shock for group $g(j)$ follows the unique distribution such that $\varsigma_{mg(j)}^i + (1-\rho)\epsilon_{mj}^i$ is also an extreme value random variable. The parameter $\rho \in [0, 1]$ governs the within group correlation; in our case this means that a larger ρ implies that all inside options (e.g. all modes) are closer substitutes. The probability that person i chooses mode j in market m is therefore given by:[32]

$$\mathbb{P}_{mj}^i = \frac{\exp\left(\frac{\delta_{mj}^i}{1-\rho}\right)}{\sum_{j \in g} \exp\left(\frac{\delta_{mj}^i}{1-\rho}\right)^{\rho} \cdot \left[\sum_g \sum_{j \in g} \exp\left(\frac{\delta_{mj}^i}{1-\rho}\right)^{(1-\rho)}\right]}. \tag{10}$$

[30] In this model, the VOT of consumer i is computed as α_T / α_p^i.

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

[32] This is the product of the conditional probability $\mathbb{P}_{mj|g}^i = \frac{\exp\left(\frac{\delta_{mj}^i}{1-\rho}\right)}{D_g}$ and $\mathbb{P}_g^i = \frac{D_g^{(1-\rho)}}{\sum_g D_g^{(1-\rho)}}$, where $D_g = \sum_{j \in g} \exp\left(\frac{\delta_{mj}^i}{1-\rho}\right)$.

Integrating over α_p^i, which follows a market-specific distribution F_m, we obtain that market shares and trips for mode j in market m are given by:

$$\mathbb{P}_{mj} = \int \mathbb{P}_{mj}^i \, dF_m(\alpha_p^i) \qquad q_{mj} = \lambda_m \cdot \mathbb{P}_{mj}.$$

3.3.2 Transportation Technology

In this section, we describe the transportation technology, which determines travel times as a function of trips and fleet sizes (which determine frequencies). We describe the (effective) travel time as the sum of several components that vary by mode: walk time, wait time, and in-vehicle time:

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

The parameter γ represents the relative distaste for time spent walking or waiting relative to in-vehicle time. Throughout the empirical model, we set $\gamma = 2$ following the survey on the empirical values of the different components of the value of time by Small (2012). For ride-hailing, walk times are assumed to be zero; for cars, both walk time and wait time are assumed to be zero. Walk times are taken as exogenous and given by market-level averages of the walk times according to our Google Maps queries. We now describe how we model in-vehicle time T_{mj}^{vehicle} as a function of vehicle flows, which affects all modes but trains. Then we describe how we model wait times for public transit and ride-hail.

To model how traffic congestion affects travel times, we think of the city as a directed graph in which nodes are CAs and edges connect CAs that are spatially adjacent. Edge $e = (a, a')$, for instance, connects CAs a and a'. We assume that routes are exogenous and pre-determined. If a traveler uses mode j in market $m = (a, \tilde{a}, h)$, she follows a directed path $r_{a,\tilde{a},j} = (e_{a,a_1,j}, e_{a_1,a_2,j}, \ldots, e_{a_{N_m-1},\tilde{a},j})$ over edges that connects a with $\tilde{a}$. During hour h, the the total vehicle flow on edge e is

$$Q_{eh} = \sum_j w_j \cdot q_{ehj}, \tag{11}$$

where q_{ehj} is the total number of vehicles for mode j that go through e and can be defined

as the sum across all routes that go through edge e, $\mathcal{R}^e_{hj}$:

$$q_{ehj} \equiv \sum_{r \in \mathcal{R}^e_{hj}} q_{rhj}.$$

The weight w_j on vehicle flows captures the fact that some modes congest more than others—buses congest more than cars, and trains do not congest at all. We assume that the travel time over edge e at time h for mode j is given by the following function:

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

This functional form is directly motivated by the observed empirical patterns in Figure 6 in our descriptive section. For every segment in our directed graph, there is a range with low vehicle flows for which, regardless of the number of vehicles that are on the road, the time that it takes to travel through the edge is T^0_{ej}. We call this the *free-flow* time. As we have shown in Figure 6, free flow times vary by markets. Variation in free-flow traffic depends on existing infrastructure, for example the number of traffic lights, road quality, or the existence of a highway.

The second part of the function represents the range in which travel times increase with the flow of vehicles. A_{ehj} is a segment-mode specific parameter that describes the underlying road infrastructure, and $Q^{\beta_j}_{eh}$ captures how traffic increases travel times.

Now that we have described how traffic flows map into travel times at the edge level, we define the travel time in market m as the sum of the travel times over all edges in the path r_{mj}:

$$T^{\text{vehicle}}_{mj} = \sum_{e \in r_{mj}} T^{\text{vehicle}}_{ehj}. \tag{13}$$

Next, we define how public transit wait times are determined. Suppose the frequency of public transit is k_{mj}, which implies an average time between vehicles—buses or trains– of $1/k_{mj}$. However, we would also like to capture variability in realized wait times because of delays and traffic. We thus assume that the time between vehicles follows some distribution with density $\phi(\cdot)$ that has mean $1/k_{mj}$ and variance η^2/k_{mj}. The variance of this distribution captures the degree of unpredictability in arrivals.

The density of travelers arriving between two subsequent buses or trains with a time difference of t is $t \cdot k_{mj} \cdot \phi(t)$: the density $\phi(t)$ is multiplied by $t \cdot k_{mj}$ because the longer

the gap between vehicles, the more riders arrive between them.[33] If the time difference is t, a rider arriving between these two vehicles needs to wait $t/2$ in expectation until the second vehicle. Therefore, the expected waiting time is given by

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

This expression captures the fact that, for a given frequency, the expected waiting time increases with the unpredictability η of public transit . If vehicles are always on schedule, $\eta = 0$ and waiting times are uniformly distributed between 0 and $1/k_{mj}$, with mean $1/(2 \cdot k_{mj})$. In the opposite extreme, vehicle arrivals are a Poisson process—arrivals are always equally likely, regardless of whether the last vehicle arrived recently or not—in which case $\eta = 1$ and the expected waiting time is $1/k_{mj}$. Our functional form therefore nests both of these cases. Our data allows us to observe the deviations from schedules, which we use to estimate η. Trains follow schedules closely, so we set $\eta = 0$. For buses, we find that $\hat{\eta} = 0.194$, which means that there is substantially more variability than for trains.[34]

In the case of ride-hailing, wait time T_{mj}^{wait} arises from a simple model of matching and of driver movements that captures three main forces. First, waiting times are lower when many drivers are working: there will be a large number of idle drivers and, thus, the nearest driver will be close to the location where the rider requested a trip. Second, waiting times are higher when many travelers demand ride-hailing trips, depleting the market of available drivers. Third, at a given point in time, waiting times are lowest in those areas where more idle drivers are located. We set up a model of driver movements that parsimoniously accounts for the fact that there will mechanically be more idle drivers in neighborhoods with a net inflows of trips, as well as for the fact that drivers tend to relocate towards areas with higher earnings opportunities. Appendix D.1 presents further details, and explains how we estimate our model using data on Uber waiting times.

3.4 Costs and environmental externalities

We assume that costs and environmental externalities are proportional to the number of miles driven by the vehicles involved in each mode. For cars and ride hailing, the

[33] This implicitly assumes that travelers arrive to the stop or station at times that are uniformly distributed.
[34] To estimate this number, we compute actual times between buses and divide them by their average at the hour by route level. The variance of this ratio is our estimate $\hat{\eta} = 0.194$.

number of miles driven depends on how many passengers choose to travel using these modes. For buses and trains, on the other hand, the number of miles driven depends on the frequency with which routes run; hence, the marginal cost and externality of an additional passenger is zero.

For all modes, the cost per mile accounts for fuel or energy, vehicle depreciation, and maintenance. For buses and trains, it also includes labor costs.[35] Environmental externalities account for the social cost of carbon, for which we use latest EPA proposal of $190 per tonne as the baseline number, as well as for the social cost of local pollutants, which we obtain from Holland *et al.* (2016).[36] Appendix D.2 describes in detail the numbers that we use for all costs and externalities.

4 Estimation and Computation

4.1 Demand model

To estimate the model we need to address two important sources of endogeneity. First, we need to grapple with the issue of price endogeneity. Public transit prices and the cost of operating a car are fixed and therefore do not respond to specific time-varying market demand-shocks. We can therefore use these prices directly as instruments, following the logic pointed out by DellaVigna and Gentzkow (2019). Ride-hailing prices, on the other hand, adjust to demand conditions, making them correlated with the demand shocks ξ_{mj}. To address this endogeneity concern, we use price variation introduced by a ride-hailing surcharge that applies to all trips that either originate or end in a downtown zone between 6AM and 10PM. In Appendix B we show the area affected by this policy and how we compute difference-in-difference estimates from this policy, which imply an own-price elasticity of -1.42. We then construct moments that capture the difference between the model-implied demand response and those derived by the treatment effect estimates.

The second endogeneity issue concerns travel times. A positive demand shock ξ_{jm} for road based modes of transportation leads to more travel in this market, which in turn induces congestion and increases travel times. As it would for prices, this type of endo-

[35] For ride hailing, labor costs depend on the number of drivers that are working, which is an exogenous quantity that is independent of the number of people that request ride-hailing trips.

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

geneity biases the travel time coefficient, which we expect to be negative, towards zero. The idea for our instrument is based on our observation of a hockey-stick functional form for congestion in Figure 6. There are large differences in free flow speed across markets, as captured by the flat portion of this function. These differences are not driven by intertemporal variation in vehicle flows but by permanent differences in road infrastructure. For instance, between two edges that are only connected by small roads with many traffic lights, traffic will flow slower than if they were connected by a highway, regardless of the number of vehicles. We therefore use free flow times T^0_{mj} as an instrument for T_{mj}.[37] Importantly, free flow times are not affected by demand shocks, and therefore are not subject to the endogeneity concerns we are worried about.

We implement the estimation using a two-step GMM procedure with moment conditions in the form of:

$$\mathbb{E}[\xi_{mj} \cdot \mathbf{Z}_{mj}] = \mathbf{0},$$

for a given vector of instruments $\mathbf{Z}_{mj}$. Following the discussion above, for non-ride hail trips we use prices directly as instruments and free-flow times to instrument travel times. We also need additional instruments to identify the non-linear parameters α_{py} and ρ that respectively govern income heterogeneity[38] and the correlation of taste shocks for the inside options. As suggested in Gandhi and Houde (2019), we construct both local and quadratic differentiation instruments based on free-flow times:

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

With the idea that ride-hailing prices must be lower in markets where other modes offer similar travel times, we also interact these with an indicator for whether a mode is not ride-hail: $\mathbb{1}^{rh}_j = \mathbb{1}\{j \neq \text{ride-hail}\}$. Lastly, we interact the exogenous non-ride hail prices with income quintiles π^y_m in each market:

$$\mathbf{Z}_{mj} = (T^0_{mj}, Z^{\text{local}}_{mj}, Z^{\text{quad}}_{mj}, Z^{\text{local}}_{mj}\mathbb{1}^{rh}_j, Z^{\text{quad}}_{mj}\mathbb{1}^{rh}_j, p_{mj}\mathbb{1}^{rh}_j, \pi^y_m \cdot p_{mj}\mathbb{1}^{rh}_j\}).$$

Our GMM objective function also includes the following indirect inference moments,

[37] Specifically, we construct market-specific free-flow times for each mode by taking the minimum observed time for that mode across all hours for that origin-destination pair.

[38] For computational simplicity, we divide the population into five income bins according to their income quintile, with income levels set according to the median income of the corresponding quintile.

which match the demand response due to the surcharge:

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

where η_{mj} is the measured elasticity for ride-hailing trips as a response to the surcharge-induced change in prices and $\Delta\tilde{\eta}_{mj}$ is the model-implied elasticity.

To compute the GMM objective function, we follow the nested fixed-point algorithm outlined in Berry *et al.* (1995). First, we guess values for the parameter vector $\theta \equiv (\alpha_p, \alpha_T, \alpha_{py}, \rho)$. We can then recover mode-market mean utilities $\hat{\delta}_{mj}(\theta) = \xi_{mj} + \alpha_T \cdot T_{mj}$ using an iterative contraction mapping.[39] Second, we construct the residuals $\hat{\xi}_{mj}(\theta) = \hat{\delta}_{mj}(\theta) - \alpha_T \cdot T_{mj}$.[40] Finally, we compute the value of the GMM objective function

$$J(\theta) = (\mathbf{Z} \cdot \hat{\xi}(\theta))' \cdot W \cdot (\mathbf{Z} \cdot \hat{\xi}(\theta)),$$

where $\mathbf{Z}$ is the matrix whose columns are $\mathbf{Z}_{mj}$ and $\hat{\xi}(\theta)$ is the vector whose elements are $\hat{\xi}_{mj}(\theta)$.

In Table 1 we show estimates for several specifications for our model, gradually building up from a simple logit model to the main specification that we outline in Section 3.3. Columns (1) and (2) show estimates from a model without a nest or heterogeneity across consumers—a standard logit model—for which we show both OLS and IV estimates. These specifications also do not yet impose the indirect-inference moments that make use of the ride-hail surcharge policy. As expected, we see that the IV specification increases both the travel time and price elasticity in absolute value, roughly doubling both of them. Specification (3) makes use of the indirect-inference moments, which leads to small changes in the elasticity but a one dollar increase in the value of time.

Starting with specification (4), we incorporate heterogeneity in price sensitivity according to income. Introducing this heterogeneity leads to overall larger price and travel-time sensitivities and increases the value of time. In column (5) we also allow for random choice set variation, based on census car ownership data of travelers. In particular, for each consumer type i in a given market m we compute the probability $p^i_{car,m}$ of owning a car. Then, if cars are an available mode in market m, we assume that cars are in the

[39] Note that due to the Box-Cox functional form the price coefficient α_p and the income-heterogeneity α_{py} are not additively separable and so price enters only into the non-linear part of utility.

[40] In specifications with fixed effects we also concentrate out any fixed effects.

Table 1: Demand Estimation Results

	Pooled						Peak	Off-Peak
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
α_T	-1.068	-1.692	-1.535	-2.345	-2.415	-1.928	-1.824	-1.872
	(0.011)	(0.022)	(0.018)	(0.023)	(0.023)	(0.018)	(0.022)	(0.027)
α_p	-0.058	-0.155	-0.129	-8.461	-3.416	-2.078	-2.388	-1.657
	(0.001)	(0.002)	(0.001)	(0.492)	(0.111)	(0.09)	(0.225)	(0.068)
α_{py}	·	·	·	-1.262	-0.588	-0.414	-0.696	-0.152
				(0.039)	(0.02)	(0.022)	(0.048)	(0.022)
ρ	·	·	·	·	·	0.262	0.376	0.162
						(0.012)	(0.017)	(0.017)
Estimator	OLS	IV	GMM	GMM	GMM	GMM	GMM	GMM
Policy Moment			✓	✓	✓	✓	✓	✓
Car Ownership					✓	✓	✓	✓
Nest						✓	✓	✓
Avg. VOT	18.41	10.89	11.9	23.88	14.65	13.47	19.47	9.81
VOT (Bot. Quintile)	·	·	·	2.44	3.26	3.62	3.9	3.42
VOT (Top Quintile)	·	·	·	64.24	32.36	27.94	45.32	18.09
Avg. Price Elast.	-0.2	-0.53	-0.44	-0.5	-0.61	-0.65	-0.55	-0.72
Avg. Time Elast.	-0.58	-0.91	-0.83	-1.26	-1.27	-1.29	-1.44	-1.07
M	92,284	92,284	92,284	91,908	91,561	91,561	42,989	48,572
N	281,755	281,755	281,755	281,042	280,185	280,185	136,337	143,848

Notes: This table presents demand estimation results from the specifications outlined in section 4.1. 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. Markets for which consumer incomes are missing are dropped in specifications that include income heterogeneity.

choice set $\mathcal{J}_m^i$ of consumers of type i only with probability $p_{car,m}^i$. Specification (6) also introduces a nest for all modes relative to the outside option, which is not to travel.

In the last two columns we present our main specification. We estimate the model from column (6) separately for peak hours–those during which the ride-hail surcharge is active—and non-peak hours. The value of time ranges from \$3.90 to \$45.30 when moving from the lowest to the highest income quintile during peak hours, and from \$3.40 to \$18 during off-peak hours. This large variation emphasizes the importance to account for income heterogeneity. When we weight peak and off-peak hours by trips, we find an average value of time of about \$15. Overall, the value of time is therefore quite stable across different specifications, ranging from \$10.89 to \$23.88.

We now present some of the key patterns implied by our main estimates. First, we examine how the value of time is distributed spatially. Second, we explore whether the model generates sensible diversion ratios across different modes of transportation.

Figure 7 shows the value of time that we infer for people in Chicago's different Community Areas. For the North Side, characterized by higher incomes, we infer values of time between $12 and $18, on average. In the South Side we mostly observe values of time below $10, with a few exceptions: Midway airport (center left) as well as the neighborhoods of Beverly, Mount Greenwood, and Morgan Park (bottom left), which were popular white-flight destinations during the 1950s and 1960s. These patterns correlate closely with the movement patterns in the bottom panel of Figure 1.

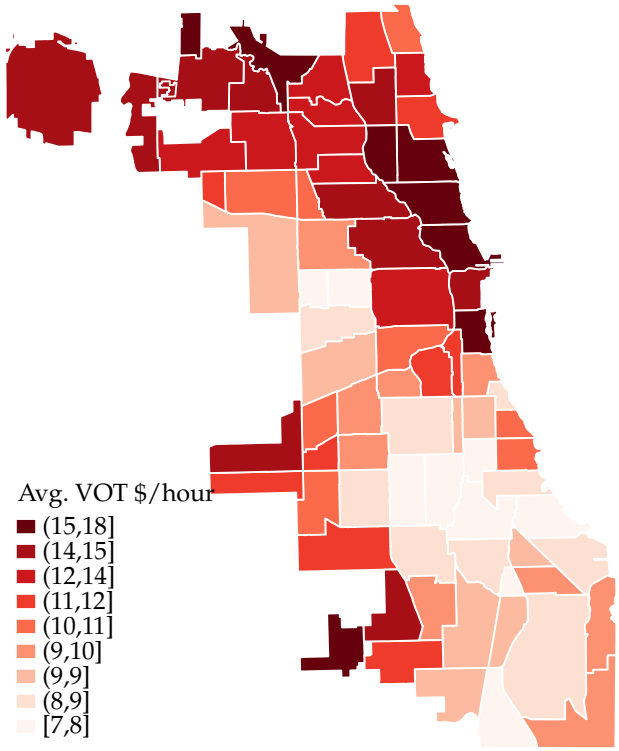


Figure 7: Value of Time across Space

Notes: This figure shows the VOT estimates across different regions in the city of Chicago under specification (6) of Table 1, which includes income heterogeneity, car ownership rates, and a nest for all inside goods. Income heterogeneity is driven by differences in the price coefficient in the utility function given by equation (9).

Table 2 presents substitution patterns in the form of diversion ratios. We can see that there is substantially more substitution to the outside option, which is is not to travel

Table 2: Diversion Ratios

(a) Overall

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

(b) Car

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

(c) No Car

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

(d) Bottom Income Quintile

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

(e) Top Income Quintile

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

Notes: These tables present average diversion ratios for various consumer types for the demand estimates in specification (6). To construct them, the diversion ratios are first averaged across markets, weighted by market size. Then, to account for the fact that not every mode is present in every market, they are normalized so that each row sums to 1. Element (m, m') of each table gives the diversion ratio from mode m to mode m'.

at all, among travelers who own cars. Furthermore, travelers in the top income quintile substitute more often to cars and ride-hail than those in the bottom income quintile. By contrast, those in the lowest income quintile are more likely to substitute towards buses or the outside option. Among cars owners, car trips are twice as likely to substitute to the outside option as the remaining modes.

4.2 Traffic congestion

In this section we show how we estimate the traffic congestion model outlined in Section 3.3.2. Recall that we model time in vehicle for edge e, at hour h, for mode j as follows:

$$T_{ehj}^{\text{vehicle}} = \max\{T_{ej}^{0}, A_{ehj} \cdot Q_{eh}^{\beta_j}\}, \tag{14}$$

where $Q_{eh} = \sum_j w_j q_{ehj}$.[41]

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 mode j and edge e to be the average travel time between these early morning hours.

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. Since $T_{ehj}^{\text{vehicle}} \geq T_{ej}^{0}$ for this subsample, our model becomes $T_{ehj}^{\text{vehicle}} = A_{ehj} \cdot Q_{eh}^{\beta_j}$. We take logs and assume that $a_{ehj} = \log A_{ehj} = a_e + \varepsilon_{ehj}$. Hence, our estimation equation is

$$\log T_{ehj}^{\text{vehicle}} = a_e + \beta_j \log Q_{eh} + \varepsilon_{ehj}. \tag{15}$$

The edge fixed effect a_e captures any edge-specific differences in infrastructure that determine travel speed outside of the free-flow region. The remaining error ε_{ehj} captures unobservable traffic shocks that vary across hours of the week h within edge e.

Tables 3 and 4 present results of regressions of the form (15) for car and bus in-vehicle travel times, respectively. These regressions include edge fixed effects, so the identification assumption is that, within an edge, shocks to the traffic congestion technology are

[41] In our data, we observe edge travel times τ_{ehj} as well as mode-specific vehicle flows q_{ehj}. To construct total-vehicle flow, Q_{eh}, we assume that $w_{car} = w_{ride-hail} = 1$ and we calibrate $w_{bus} = 2$, a parameter that we take from the Traffic Modelling Guidelines for London.

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 visibility due to day and night. When we control for such variables (temperature, visibility, and precipitation) in column (2), we estimate a similar elasticity. In column (3), we follow an instrumental variables approach to address the concern that people may re-optimize their route choices due to unobservable but expected traffic shocks, such as planned construction during certain hours of the day. Our instrument is thepotential number of travellers at the city level, which is exogenous as long as the total demand for travel is driven by daily patterns—commuting to work in the morning, going out for dinner in the afternoon, etc.—rather than by unobserved traffic shocks like anticipated construction or public transit disruptions. We find elasticities of travel time with respect to traffic flows between 0.05 and 0.07 for buses and between 0.10 to 0.17 for cars. The main elasticities that we use for our model are those in the third column.

Table 3: Effect of road congestion on bus travel times

	Dependent variable: Log travel time for bus		
	(1) Fixed effect	(2) + Weather Controls	(3) + IV
Log Flow	0.054^{***}	0.020^{***}	0.069^{***}
	(0.004)	(0.005)	(0.006)
Edge FE	✓	✓	✓
Weather controls	✗	✓	✓
IV	✗	✗	✓
within R^2	0.066	0.114	0.083
First-stage F			3229.320
Observations	9401	9401	9401

Notes: This table shows the regression estimates for the elastic portion of the congestion function separately for buses. The unit of observation is an edge. The dependent variable is the log of travel times for buses and the key independent variable is the log of all vehicles on the road. The first specification controls for edge fixed effects, the second specification adds weather controls and the third specification uses the potential market size as an instrument for potentially endogenous vehicle flows. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Effect of road congestion on car travel times

	Dependent variable: Log travel time for car		
	(1) Fixed effect	(2) + Weather Controls	(3) + IV
Log Flow	0.128***	0.100***	0.168***
	(0.005)	(0.005)	(0.004)
Edge FE	✓	✓	✓
Weather controls	✗	✓	✓
IV	✗	✗	✓
within R^2	0.411	0.529	0.444
First-stage F			4488.596
Observations	11739	11739	11739

Notes: This table shows the regression estimates for the elastic portion of the congestion function separately for cars. The unit of observation is an edge. The dependent variable is the log of travel times for cars and the key independent variable is the log of all vehicles on the road. The first specification controls for edge fixed effects, the second specification adds weather controls and the third specification uses the potential market size as an instrument for potentially endogenous vehicle flows. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

5 Solving for Equilibrium and the Planner's Problem

Before we move on to describe our counterfactuals, we spell out more concretely what an equilibrium for our model entails and how we compute both equilibria and the planner's problem. To remind you, a market equilibrium is a vector of travel times $\mathbf{T}_m = (T_{1m}, ..., T_{Jm})$ and quantities $\mathbf{q}_m = (q_{1m}, ..., q_{Jm})$ such that demand results under those travel times and quantities result from demand—i.e., equations (2) and (4) hold simultaneously. What does this mean concretely given our empirical model?

Definition 2 (Detailed Transportation Equilibrium)**.**

1. *Quantities result from the demand model:* $q_{mj} = N_m \cdot \mathbb{P}_{mj}(T_{mj}) \;\; \forall j \in \mathcal{J}, m \in \mathcal{M}$,
2. *Travel times for road based modes result from the congestion function:* $T_{mj}^{vehicle} = \sum_{e \in r_{mj}} T_{ehj}^{vehicle}$ $\forall m \in \mathcal{M}$
3. *Wait times for buses are given by:* $T_{mj}^{wait} = \frac{1}{2 \cdot k_{mj}}, \forall m \in \mathcal{M}$
4. *Wait times for ride-hail are given by:* $T_{mj}^{wait} = A_{mj}^{W} I_{mj}^{-\phi_j}, m \in \mathcal{M}$
5. *Idling drivers in market* m *are given by:* $I_{mj} = (L_{mj} - B_{mj}) \cdot \frac{\exp\left(\mu_m + \sum_b B_{mb} F_{mb}\right)}{\sum_k \exp\left(\mu_k + \sum_b B_{kb} F_{kb}\right)}$

In words, travellers take as given travel times, prices, and wait times for buses, taxis, and ride-hailing when making their travel decisions. These decisions generate flows across community areas and therefore congestion at a given edge, which is a function of all the vehicle flows through an edge. Road-based travel times have to be consistent with edge level congestion. Wait times for public transit have to be consistent with the fleet size of public transit and wait times for for hire vehicles with the driver movements. Driver movements in turn have to be consistent with consumer demand and road based travel times. The main computational challenge arises from requirements 2 and 5 in this list: given the linkages of markets through the congestion technology and driver movements, an equilibrium requires the plans of agents in thousands of markets to be consistent with each other.

To find an equilibrium, we write the two conditions equations (2) and (4) as a fixed point problem with prices $\mathbf{p}$ and capacities $\mathbf{k}$. Let $f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) = q(\mathbf{p}, T(\mathbf{q}, \mathbf{k}))$. The problem of finding an equilibrium is equivalent to finding a fixed point of $f^{\mathbf{p},\mathbf{k}}$. In words, q^* is an equilibrium if people would demand q^* trips when (wait and in-vehicle) times are those that arise when q^* trips take place. Naive algorithms to find fixed points—such as simple or dampened fixed point iteration—often diverge. We thus rewrite the fixed point problem as a root-finding problem (i.e., finding a root $f^{\mathbf{p},\mathbf{k}}(\mathbf{q}) - \mathbf{q} = 0$) and we use a limited-memory version of Broyden's method to solve it. Appendix D.3 describes in detail the algorithm that we use.

Once we find an equilibrium, we can compute all quantities that go into the city government's objective function. To solve the social planner's problem —which involves a budget constraint — we follow the augmented Lagrangian method (Nocedal and Wright, 2006), where we iteratively maximize problems that approximate the Lagrangian of the main problem until convergence. We provide further details in Appendix D.4.

6 Optimal Counterfactual Policy Design

In what follows, we explore counterfactual policy designs based on the optimality conditions that we described previously in equation (5). We first separately analyze the case where the planner only sets transit prices and frequencies. To quantify the additional distortions due to budget considerations, we compare the case of a budget constraint planner who can not exceed the CTA's operating deficit in 2019 (*Budget Constrained Transit*), which

is given by $-\$15.4$ million per week, to an unconstrained planner (*Transit*).[42] We then separately analyze the effect of *Road Pricing*. To explore whether these two different sets of policies act as complements or substitutes in the planner's goal to maximize welfare, we next analyze the case where the planner has all of these policy tools available at once (*Transit + Road pricing*). It turns out that in this case the budget constraint is never binding and we therefore drop it. Lastly, as a benchmark, we explore the case where the planner also controls the price of ride-hail services (*Social Planner*).

We now turn to a broader overview of these counterfactuals. The results obtained under the optimal schemes across different policy interventions can be found in Table 5, with results reported relative to the status quo in the first column. Figure 8 shows the resulting substitution patterns.

In all counterfactuals where the planner controls public transit, prices for both buses and trains are considerably reduced. In those scenarios, the planner also reduces the frequency of buses and increases the frequency of trains. The reason for this divergence is that buses serve more price sensitive travelers with lower value of time. Trains, on the other hand, are more likely used by higher income travelers with a high value of time. Since low income travelers are more reliant on buses, we explored under which welfare weights for different income groups one would obtain the observed frequencies.

Transit raises overall welfare by \$2.16 million per week whereas *Budget Constrained Transit* by only \$1.45 million per week. However, the welfare cost of budget distortions to travelers is large. Without a budget constraint, their surplus increases by $\$10.4$ million per week but only by $\$740,000$ when they are present. The reason for this discrepancy is that prices for both modes are about one dollar higher when the planner is constrained and the frequency of both buses and trains is reduced by ten percentage points. Lastly, while an unconstrained planner reduced the cost of externalities by \$900,000 per week, a constraint planner only does so by \$700,000 per week. From this exercise we conclude that to maintain the observed frequencies, the government must weight the lowest income quintile 4.4 times more than the highest income quintile, the second-lowest quintile 2.8 times more, the middle income group 2.1 times more, and the second-highest income quintile 1.6 times more.

In terms of the total number of trips taken, we find that when the city does not face

[42] For the operating expenses we use maintenance and vehicle operating expenses provided in https://www.transit.dot.gov/ntd/data-product/2019-operating-expenses.

Table 5: Counterfactual Results

		Status quo	Transit	Transit, budget	Road pricing	Transit + Road pricing	Social planner
		(1)	(2)	(3)	(4)	(5)	(6)
Avg. Price ($)	Bus	1.09	-0.33	0.44	1.09	0.07	-0.01
	Train	1.33	-0.45	0.66	1.33	0.10	-0.05
	Ride-hailing	9.10	9.10	9.10	9.10	9.10	8.09
Road Price ($/km)		0	0	0	0.367	0.337	0.313
Δ Frequency	Bus	0%	-19.55%	-28.4%	0%	-19.06%	-18.22%
	Train	0%	19.6%	9.17%	0%	19.24%	20.3%
Avg. Wait (min)	Bus	6.47	7.87	8.61	6.57	7.92	7.84
	Train	4.29	3.93	4.10	4.41	3.99	3.96
Δ Welfare ($M/week)		0	2.16	1.45	4.19	5.60	6.00
Δ CS ($M/week)		0	10.4	0.742	-25.1	-16.2	-12.2
Δ City Surplus ($M/week)		0	-8.56	0	24.9	17.4	14.9
Δ Transit Surplus ($M/week)		0	-8.56	0	0.656	-4.94	-6.04
Taxes ($M/week)		0	0	0	24.3	22.3	20.9
Externalities ($M/week)		17.3	16.4	16.6	14.2	13.8	14.0

Notes: This table presents the changes in prices, service levels, and welfare relative to the status quo (column 1) across different counterfactual scenarios. Column 3 only changes public transit prices and service levels without budget considerations. Column 3 repeats the same exercise subject to a budget constraint. Column 3 changes road pricing. Column 4 combines both. Column 5 also set optima ride-hailing prices.

a budget constraint, optimal transit policies increase the total number of trips taken by 800,000 per week and 250,000 per week when the city is budget-constrained.

We now discuss the effects of road pricing, which we explore both in isolation and in combination with transit policies. We find that the optimal per-km surcharge is 37 cents in the former and 34 cents in the latter case. The overall welfare gains in both of these counterfactuals are much larger than from transit policies alone and equal to $\$4.12$ and $\$5.6$ respectively. These welfare gains predominantly result from a reduction in externalities and hurt travelers in both case. However, in both cases the city now generates a surplus, which it can rebate to consumers. Even with rebates, consumers would still lose $\$200,000$ in weekly surplus if the city just enacts a surcharge. However, they would gain $\$800,000$ if the surcharge is combined with optimal transit prices and frequencies. Because road pricing relaxes the city's budget constraint, the frequency reductions in transit are now

close to what they are under the unconstrained version (2). Every policy that involves road pricing leads to a decline in the total number of trips taken across all modes (see Figure 8). When just the surcharge is enacted the total number of trips is reduced by 2.5 million per week. This effect is dampened to a 2 million reduction when transit policies are combined with the surcharge.

As a benchmark, we also explore the case where the planner can set prices for all modes, including ride hail. Interestingly, in this case the planner wants to only slightly reduce the price of ride-hail, which indicates that the markup that these platforms are charging is close to pricing their externality efficiently. This leads to a welfare gain of $\$6$ million per week but also an increase in the cost of externalities.

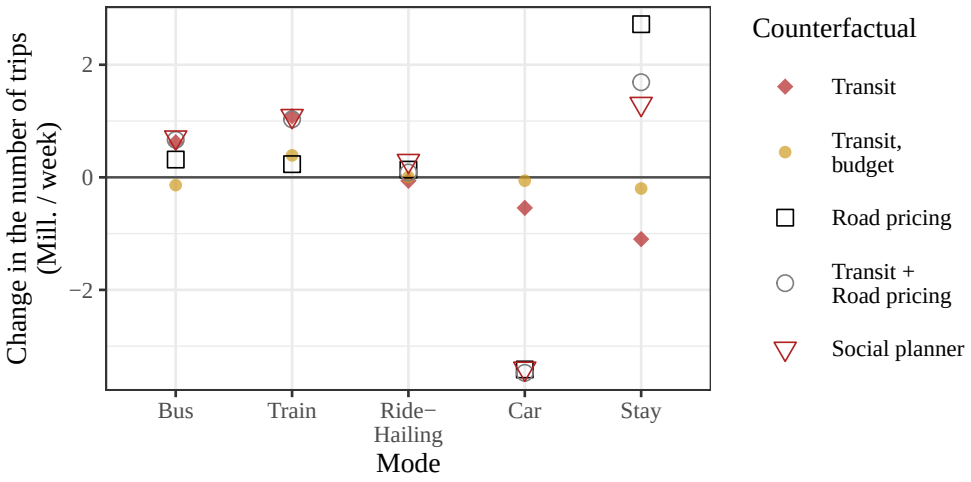


Figure 8: Substitution patterns across different policies

Notes: This figure presents changes in the number of trips by mode relative to the status quo under optimal policies across different counterfactual scenarios.

We now decompose the forces that give rise to these results and present the empirical analogue of our theoretical results in Section 3.2 by decomposition of optimal transit prices and frequencies and a similar decomposition for the per-kilometer cost of operating car when the planner can set a congestion surcharge (Equation 5). Recall that these prices account for the direct cost as well as environmental and network externalities. In addition, there is a misallocation term that arises due to the planner's desire to internalize externalities of modes not directly under her control. Lastly, there is a term that accounts for budget considerations, akin to a markup term. The budget term summarizes all terms that multiply $\lambda/(1+\lambda)$.

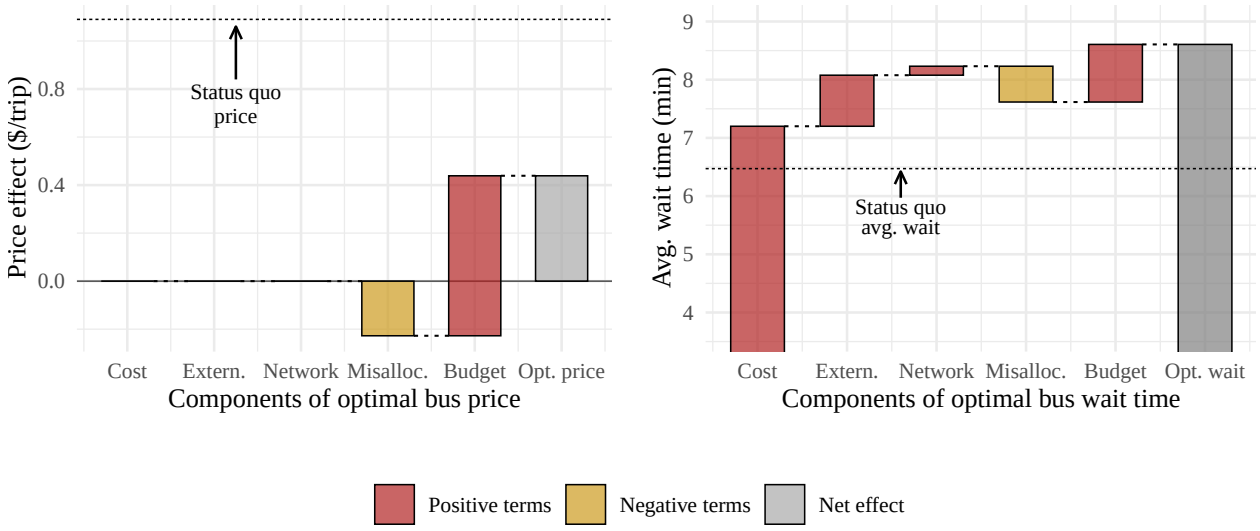


Figure 9: Decomposition of optimal price and waiting times for public transit modes

Notes: This graph shows the a decomposition of the optimal prices and travel times for buses corresponding to our theoretical decomposition in Section 5. Red bars indicate terms that lead prices and travel times to be higher and yellow bars indicate terms that lead prices to be lower.

Figure 9 shows the decomposition of the optimal prices and travel times for buses under *Budget Constrained Transit*. We observe that optimal prices are lower and wait times higher than at the status quo. For prices we observe a negative mis-allocation term. This is due to the planner's inability to affect prices of other road-based modes of transportation. The price of buses is therefore lowered to internalize the externalities of these modes. The budget term is positive, which shows that the only reason that prices are positive in this scenario if because of the necessity to raise revenues. While the government wants to lower prices, it wants to decrease the frequency of buses, which leads to an increase in wait times. This arises because the marginal cost of transporting an additional passenger is zero while the marginal cost of adding an additional bus is large. Consequently, both cost and budget considerations lead to higher wait times.

A similar decomposition is provided in Figure 10 for the price of operating a car. We observe that the optimal per-km price is about twice the price of the status quo. The status-quo is simply given by the cost of operating a car. About 50% of the price increase is to lower congestion. About 30% of the price increase is due to environmental externalities and the remaining 20% because of mis-allocation. The latter is effectively due to ride-hail,

which is a mode whose price the government has no direct control over. The budget-term is zero because the ability to lever a congestion-surcharge makes the government's budget constraint non-binding.

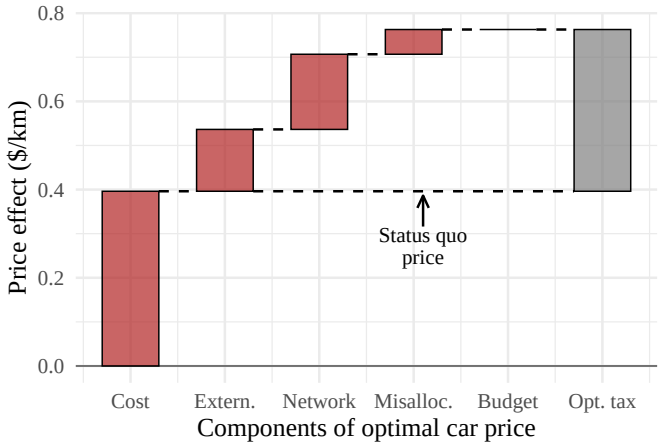


Figure 10: Decomposing the Optimal Price of Operating a Car

Notes: This figure shows the price decomposition for cars, following our theoretical derivations in Section 5. Red bars indicate terms that lead prices and travel times to be higher and yellow bars indicate terms that lead prices to be lower.

Robustness Figure 11 shows the extent to which our main results are sensitive to some of the key parameters of our model, focusing on the *Transit + Road pricing* counterfactual. Each panel shows how a 10% increase in several parameters of the model affect the five choice variables of the city government. In the first two panels, we see that the finding that the price of public transit should be close to zero is very robust: the optimal prices are always within 1.5 cents of our baseline results. On the other hand, our results about the optimal waiting time for public transit—in other words, for the frequencies of public transit—are more sensitive to parameters, as can be seen in the next two panels. For four of the six parameters (the marginal cost of public transit, the price and time sensitivity of travelers, the relative disutility of walking and waiting, and the variability of bus arrivals), a 10% change in the value of the parameter results in changes to the optimal bus wait time on the order of half a minute and in the optimal train wait time on the order of 0.2 minutes. These changes correspond to changes in frequencies of around 5%. Finally, the last panel shows that the road price is also quite robust: in every case, the optimal value is within 1.5 cents per km of the baseline value.

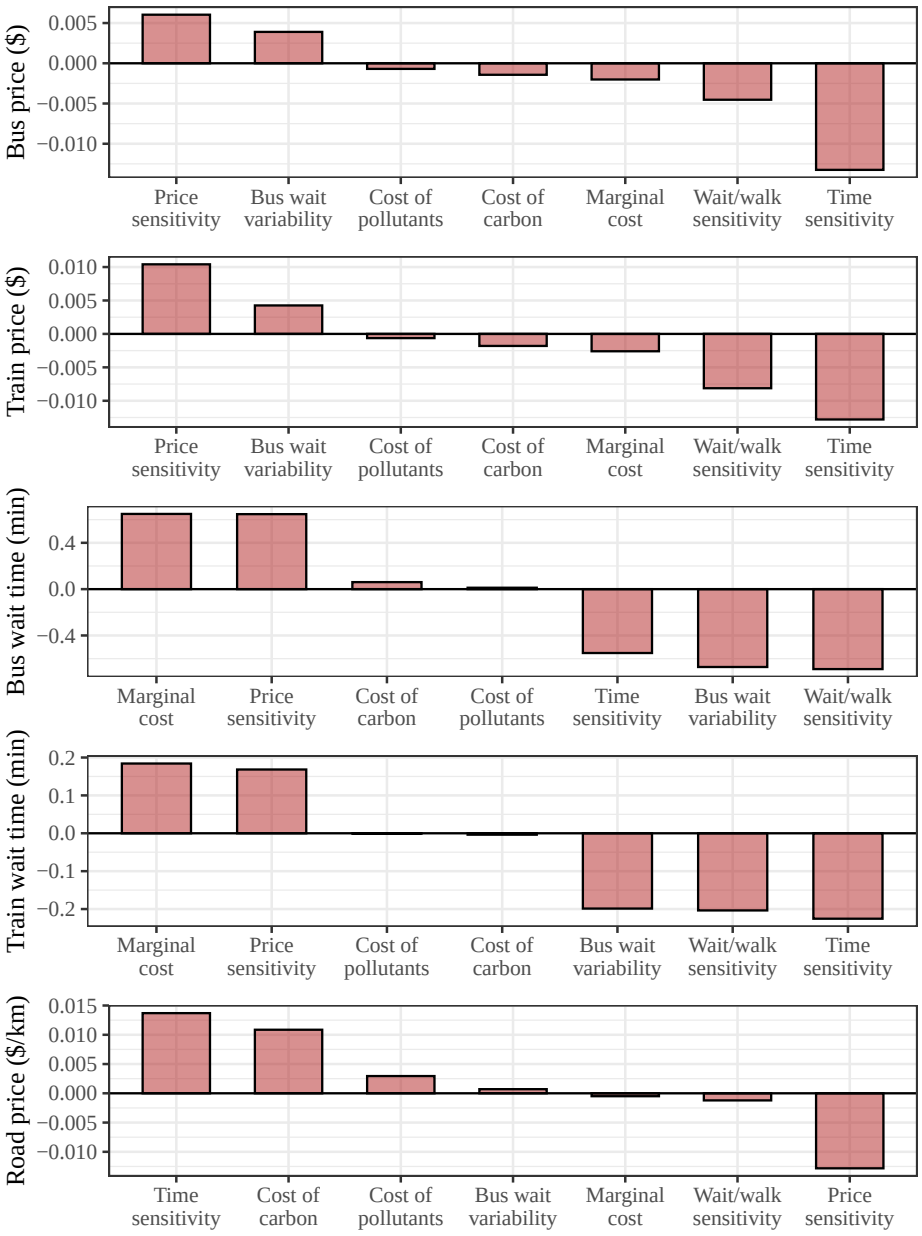


Figure 11: Robustness of counterfactual results

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

One important model parameter over which there is substantial disagreement in the literature is the social cost of carbon. This figure shows that the only result on which it has

an important impact is on the optimal road price. If instead of the latest recommendation from the EPA ($190 per tonne of CO_2) we used the previous figure of $51 per tonne, the optimal road price would drop by around 9 cents to $0.25 per km.

Distributional Effects We now investigate the distributional effects of these policies and compute consumer surplus changes across income quintiles. Results can be seen on Figure 12, which shows both absolute and percentage changes across the income distribution. Optimal public transit policies in isolation lead to gains in consumer surplus for all travelers except for the highest income group. They are also progressive both in absolute terms and in relative terms. The three remaining policies, which include a congestion surcharge all exhibit a similar pattern. All income groups are now worse off and absolute losses are U-shaped, which is related to the U-shaped pattern in car market shares across income levels as shown in panel (a) of Figure 4. However, when we measure those losses as a percentage of consumer surplus, we find that these policies are highly regressive. To summarize, our counterfactual analysis reveals that the policies that deliver the largest efficiency gains are also the ones that are the most regressive in terms of consumer surplus.

6.1 Discussion

Although our model incorporates many ingredients, such as rich heterogeneity in substitution patterns and differences in marginal congestion across vehicle types, we make some simplifying assumptions to keep our model tractable. In this section, we discuss these assumptions and their implications.

First, our model does not allow travelers to choose when to travel. Intertemporal substitution is, instead, modeled indirectly: decisions not to travel at certain times are captured as substitution towards the outside option. Although this allows us to model elasticities to own prices and own travel times correctly, we are not able to capture spillover effects between different hours of the week. A more realistic model would capture those spillover effects; however, Kreindler (2023) finds that, after accounting for individual constraints and heterogeneity, inter-temporal choices are rather inelastic and peak-spreading policies have a limited impact. For these reasons, we prefer not to model intertemporal substitution directly because it would add substantial complexity to our model.

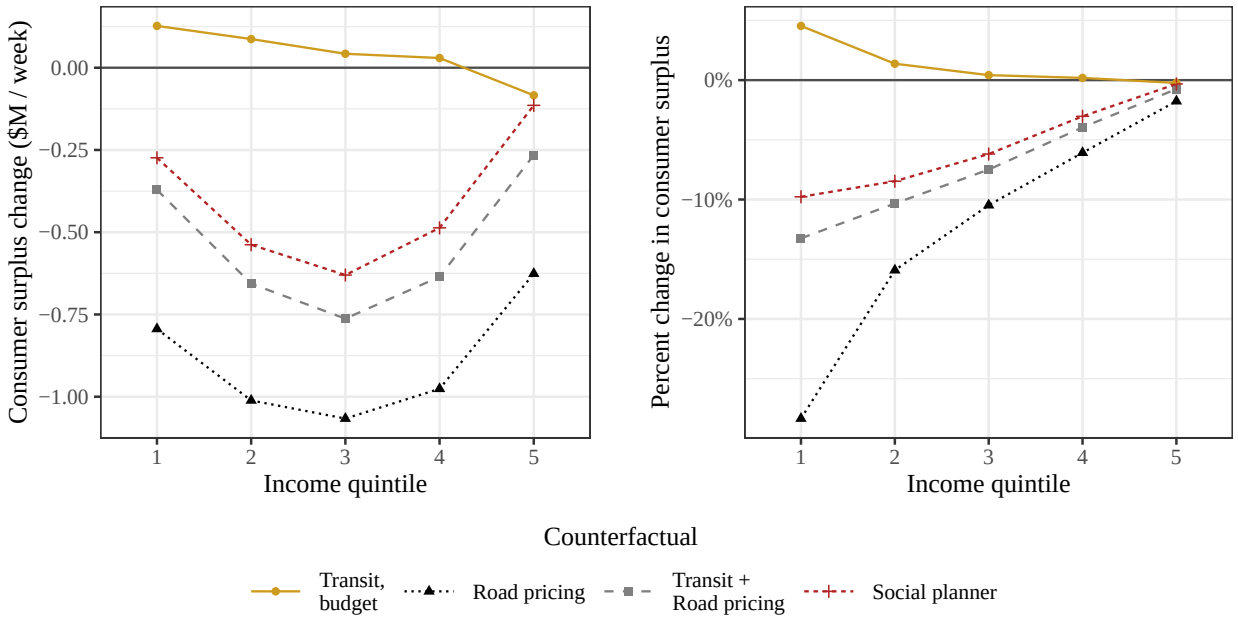


Figure 12: Change in Consumer Surplus Across Income Quintiles

Notes: This figure presents changes in consumer surplus relative to the status quo under optimal policies across different counterfactual scenarios. We calculate consumer surplus for five groups according to traveler's income quintiles. Panel (a) displays net changes per trip in dollar values across the five quintiles of the income distribution. Panel (b) displays changes as a percent.

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

Third, our model keeps the origin and destination of trips fixed. We think of our model as quantifying short-run responses, before the spatial distribution of the economic activity starts responding to policy interventions. One can thus take our model as the bottom layer of an economy that incorporates such long-run responses. Although we believe those longer-run responses are important, we defer their investigation for future research.

Finally, routes are fixed in our model. This is not a realistic assumption because congestion directly impacts the choice of routes. However, our spatial areas are relatively large—with a median size of 7 km^2. Therefore, our model does capture small re-routing

deviations, the most common ones in urban transit.

7 Conclusions

In this paper, we measure the welfare effects of urban transportation policies, and we explore how a budget-constrained planner should set optimal prices and public transit service levels. We first start by showing in a theoretical model that on top of congestion and environmental externalities, budget considerations introduce additional sources of inefficiency as the planner starts behaving somewhat like a monopolist.

Then, we move on to empirically quantify welfare effects. We focus on the city of Chicago due to data quality and the importance of its public transit system, the second largest in the nation. To do so, we construct a novel data set that comprehensively captures trip flows across locations, hours, and modes. Using this dataset, we estimate a model of mode demand that allows us to quantify how travelers substitute across modes. We also estimate a congestion technology that allows us to measure how vehicle flows map onto travel times.

Finally, we use our estimates to run counterfactual simulations and characterize optimal policies for a battery of counterfactual scenarios. We find that congestion prices deliver the largest welfare gains but they also come at the cost of being highly regressive.

References

AKBAR, P., COUTURE, V., DURANTON, G. and STOREYGARD, A. (2023). Mobility and congestion in urban india. *American Economic Review*, **113** (4), 1083–1111. 7, 53

— and DURANTON, G. (2017). *Measuring the cost of congestion in highly congested city: Bogotá*. CAF - Working paper N° 2017/04, CAF, https://scioteca.caf.com/handle/123456789/1028. 4, 7

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

ARNOTT, R. (1996). Taxi travel should be subsidized. *Journal of Urban Economics*, **40** (3), 316–333. 5, 20

—, DE PALMA, A. and LINDSEY, R. (1990). Economics of a bottleneck. *Journal of urban economics*, **27** (1), 111–130. 6

—, — and — (1993). A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *The American Economic Review*, pp. 161–179. 6

BARWICK, P. J., LI, S., WAXMAN, A. R., WU, J. and XIA, T. (2021). *Efficiency and equity impacts of urban transportation policies with equilibrium sorting*. Working paper, National Bureau of Economic Research, https://www.nber.org/papers/w29012. 6

BERRY, S., LEVINSOHN, J. and PAKES, A. (1995). Automobile prices in market equilibrium. *Econometrica*, **63** (4), 841–890. 3, 7, 10, 28, 70

BERRY, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pp. 242–262. 7, 10

BORDEU, O. (2023). Commuting infrastructure in fragmented cities. *Job Market Paper, University of Chicago Booth School of Business*. 6

BRANCACCIO, G., KALOUPTSIDI, M. and PAPAGEORGIOU, T. (2020a). Geography, transportation, and endogenous trade costs. *Econometrica*, **88** (2), 657–691. 5

—, —, — and ROSAIA, N. (2020b). *Search frictions and efficiency in decentralized transportation markets*. Tech. rep., National Bureau of Economic Research. 5

BRINKMAN, J. and LIN, J. (2022). Freeway revolts! the quality of life effects of highways. *The Review of Economics and Statistics*, pp. 1–45. 6

BROOKS, L. and LISCOW, Z. D. (2021). Infrastructure costs. *Available at SSRN 3428675*. 2

BROUGH, R., FREEDMAN, M. and PHILLIPS, D. C. (2023). Eliminating fares to expand opportunities: Experimental evidence on the impacts of free public transportation on

economic and social disparities. 9

BUCHHOLZ, N. (2021). Spatial Equilibrium, Search Frictions, and Dynamic Efficiency in the Taxi Industry. *The Review of Economic Studies*, **89** (2), 556–591. 5

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. 72

CASTILLO, J. C. (2023). *Who Benefits from Surge Pricing?* Working paper, University of Pennsylvania, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3245533. 5

—, KNOEPFLE, D. and WEYL, G. (2018). Surge pricing solves the wild goose chase. *Working Paper*. 69

COOK, C. and LI, P. Z. (2023). Value pricing or lexus lanes? the distributional effects of dynamic tolling. 7

COUTURE, V., DURANTON, G. and TURNER, M. A. (2018). Speed. *Review of Economics and Statistics*, **100** (4), 725–739. 4, 7

DELLAVIGNA, S. and GENTZKOW, M. (2019). Uniform pricing in us retail chains. *The Quarterly Journal of Economics*, **134** (4), 2011–2084. 26

DURRMEYER, I. and MARTÍNEZ, N. (2023). *DP18332 The Welfare Consequences of Urban Traffic Regulations*. Cepr discussion paper no. 18332, CEPR, https://cepr.org/publications/dp18332. 5, 6

FAJGELBAUM, P. D. and SCHAAL, E. (2020). Optimal transport networks in spatial equilibrium. *Econometrica*, **88** (4), 1411–1452. 6

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. 5

FUCHS, S. and WONG, W. F. (2022). *Multimodal Transport Networks*. Federal Reserve Bank of Boston. 5

GANDHI, A. and HOUDE, J.-F. (2019). *Measuring Substitution Patterns in Differentiated-Products Industries*. Working Paper 26375, National Bureau of Economic Research, http://www.nber.org/papers/w26375. 27

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. 7

HERZOG, I. (2021). *The city-wide effects of tolling downtown drivers: Evidence from london's congestion charge*. Tech. rep., Working Paper. 6

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. 26, 71

KREINDLER, G. (2023). *Peak-hour road congestion pricing: Experimental evidence and equilibrium implications*. Tech. rep., National Bureau of Economic Research. 4, 5, 7, 42

—, 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, http://www.nber.org/papers/w31369. 5, 6

LAGOS, R. (2003). An analysis of the market for taxicab rides in new york city. *International Economic Review*, **44** (2), 423–434. 5, 69

LECCESE, M. (2021). Asymmetric taxation, pass-through and market competition: Evidence from ride-sharing and taxis. *Pass-through and Market Competition: Evidence from Ride-sharing and Taxis (April 12, 2021)*. 7

— (2022). Do minorities pay more for congestion taxes? evidence from a tax on ride-sharing. *Evidence from a Tax on Ride-Sharing (January 30, 2022)*. 7

MIRAVETE, E., SEIM, K. and THURK, J. (2023). Elasticity and curvature of discrete choice demand models. 22

NOCEDAL, J. and WRIGHT, S. J. (2006). *Numerical Optimization*. Springer New York, NY. 35, 73

PARRY, I. W. H. and SMALL, K. A. (2009). Should urban transit subsidies be reduced? *American Economic Review*, **99** (3), 700–724. 7

RAMOS, M. (2019). Cta avoids service cuts, fare hikes under proposed $1.8 billion budget. *https://chicago.suntimes.com/2022/10/20/23413692/cta-proposed-budget-2023-chicago-transit-authority-no-fare-hikes-service-cuts*. 8

ROSAIA, N. (2023). *Who Benefits from Surge Pricing?* Working paper, Columbia Business School, https://drive.google.com/file/d/104F8R6yZ1jHyVuoBbGTtzKeYlMhM6xsx. 5

ROSENBLUM, J. L., ZHAO, J., ARCAYA, M., STEIL, J. and ZEGRAS, P. C. (2020). How low-income transit riders in boston respond to discounted fares: A randomized controlled evaluation. In *2020 APPAM Fall Research Conference*, APPAM. 9

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. 6

SMALL, K. A. (1982). The scheduling of consumer activities: work trips. *The American Economic Review*, **72** (3), 467–479. 6

— (2012). Valuation of travel time. *Economics of Transportation*, **1** (1), 2–14. 23

—, WINSTON, C. and YAN, J. (2005). Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica*, **73** (4), 1367–1382. 6

SPENCE, A. M. (1975). Monopoly, quality, and regulation. *The Bell Journal of Economics*, pp. 417–429. 2

TSIVANIDIS, N. (2023). *Evaluating the Impact of Urban Transit Infrastructure: Evidence from Bogota's TransMilenio*. Working paper, University of California, Berkeley, https://www.nicktsivanidis.com/s/TsivanidisTransMillenio_82023.pdf. 6

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. 7

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. 8, 55

A Data Appendix

A.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 share of car trips is consistent with what is reported by the Chicago Metropolitan Agency for Planning (CMAP) 2019 Household Travel Survey.[43]

A.1.1 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 months of July 2019 - March 2020 and for the months of June 2020 - August 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.[44] 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.[45] We define a stream of pings as a sequence of pings for the same cellphone identifier such that a ping always has another ping within the next 15 minutes and within 1,000 meters. We drop streams with less than 3 pings. Finally, we aggregate pings to the minute of the day by taking the average location and timestamp across pings within each minute for a given

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

[44] 40 meters per second, i.e. about 145 kilometers per hour

[45] In particular, we only keep pings that satisfy the following two conditions: (i) no more than ten minutes to either the next or the previous ping, (ii) no more than 5,000 meters to either the next or the previous ping.

cellphone identifier. In what follows, we focus on the remaining filtered pings aggregated at the minute level.

A.1.2 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).

A.1.3 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.

Home locations We start by assigning all cellphone pings to census block polygons for the subset of pings within Chicago during our sample period.[46] 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.
- 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.

[46] See Appendix A.1 for the sample restrictions.

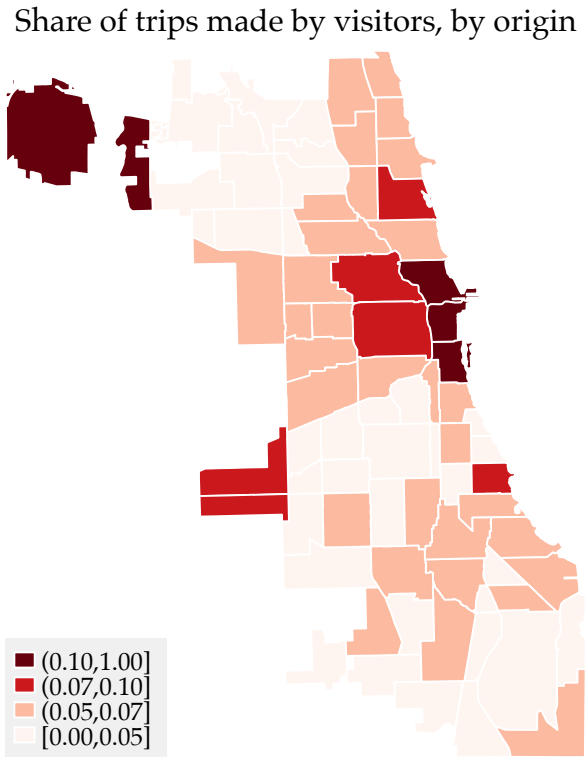


Figure 13: Share of visitors by origin location

Notes: This figure shows the share of trips at the origin Community Area 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 Community Area, using inside market size (number of cellphone trips per market) as weight.

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 13 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).

Individual cellphone's income For all cellphones with an assigned home location, we impute their income by using the census tract median household income.[47]

Market-level income quintile shares Next, for each market, we estimate traveler's income distribution.[48] First, we take median income by tracts and divide tracts according to quintiles.[49] Next, assign an income quintile to each device according to their home location. Since we can follow how devices travel across space and over time, for each market we can measure the quintile from each traveler departing from its destination. Finally, for each market, we construct shares of traveler's income quintile. For markets with less than 5 trips, we impute market-level income shares using the underlying distribution of census tract-level income for the origin Community Area of that market.

A.1.4 Cell-phone data validation

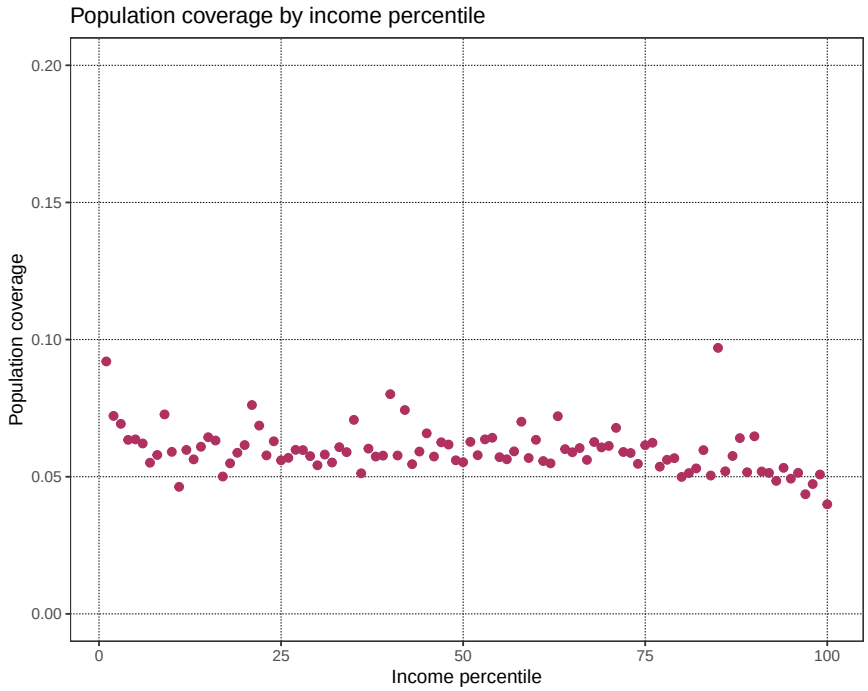
In this section, we provide evidence that the cell-phone geo-location data is representative of the population and that it measures traveling patterns with precision.

First, we measure the coverage of our cell-phone devices across the income distribution. To do so, for each census tract, we count the number of devices whose home location falls within the tract boundaries. We divide those counts by the census population. Figure 14 displays the share of the tract population covered by the cell-phone data. We order tracts by income percentiles. We see that our coverage is fairly constant and hovers 5% across all percentiles of the income distribution.[50] We take this as evidence that our cell-phone location records cover a representative sample of the population.

Next, we show that the cell-phone location records can accurately represent travelling patterns. To do so, we plot the travel (geodesic) distance distribution for the universe of trips for both the cell-phone data and the survey data. The two distributions, plotted in Figure 15, present a high degree of overlap and similarity. We conclude that movements measured using cell-phone location records with the procedure described in Section A.1 are a good representations of the travel patterns in the city of Chicago.

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

Figure 14: 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.

A.2 Travel times, routes, and schedules

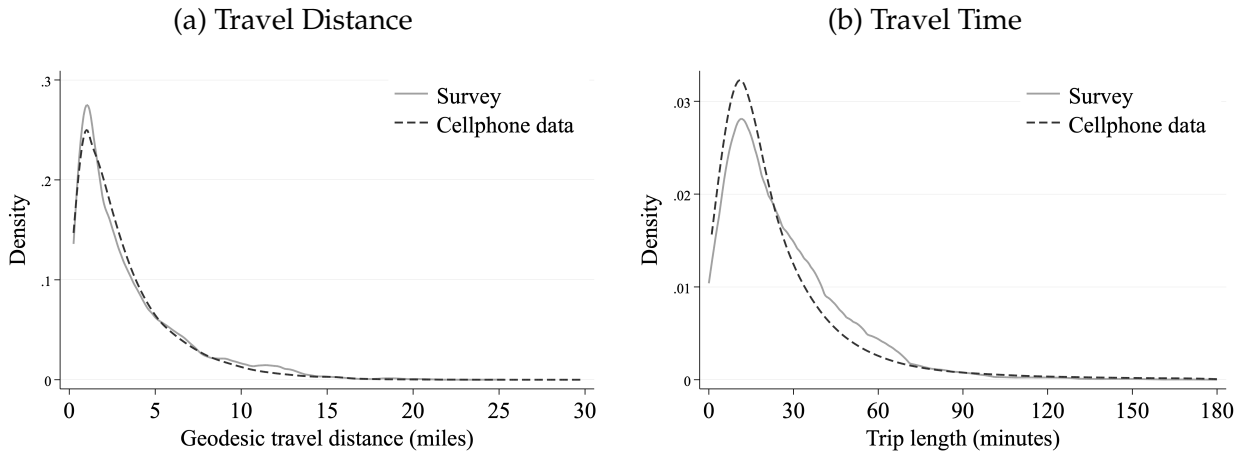
A.2.1 Travel times and routes

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 path.[51] Importantly, we can measure trip duration for the same origin-destination tuple over the time of the weekday (or weekend) and how this varies with traffic conditions. Moreover, using the detailed "steps" of the public transit Google Maps queries, we obtain walk times from the origin latitude/longitude to the "best" train or bus station.[52]

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

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

Figure 15: Travel Distance Histograms



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 cell-phone data. Our level of observation is a trip. Trips in the cell-phone data are constructed following the steps in Appendix A.1. Trips in the survey data do not include walking, biking or multi-modal trips.

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.

A.2.2 Public transit schedules

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

A.3 Constructing Market Shares

The market shares 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.

A.3.1 Raw data processing

Taxi and Transportation Network Provider (TNP) data

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

Cellphone trips data

We construct cellphone trips from cellphone pings using the procedure detailed in Appendix A.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.[54] Specifically, we calibrate these factors to ensure that the proportion of car trips beginning in each neighborhood in our dataset matches the corresponding proportion in the Chicago Metropolitan Agency for Planning (CMAP) Household Travel Survey.[55]

Public transit data

We obtain individual public transit trips via a partnership between the MIT Transit Lab and the CTA. Specifically, we use the near-universe of CTA bus and train trips within the city of Chicago. This data notably excludes trips taken via the Metra, which is a suburban rail system operating in and around Chicago. Metra is managed by a different agency, the Regional Transportation Authority. Each observations corresponds to a a passenger swiping in to access the bus or the train station. For buses, we observe the specific bus stop, bus line, and boarding time. For trains, we observe the station and swiping time. Drop-off locations are given to us and imputed following Zhao *et al.* (2007). Unfortunately, these data are missing 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 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

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

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

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

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.[56] In particular, we first assign each trip one of three time categorizations: weekend (if Saturday or Sunday), peak weekday (if between 5-9:59am or 2-6:59pm on a weekday), or non-peak weekday (otherwise). We then take the time to be the travel time of the matching train trip from the Google Maps data.

As an additional cleaning step, we drop any duplicate trips along with any trips that have missing coordinates, mode information, or time (for bus trips). We also compute travel distances for each trip. We use the Harversine 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. Finally, since there could be measurement error, we account for trips that may be missing in our dataset by assigning each trip an inflation factor. This inflation factor is computed at the day-mode level such that

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

where dm indexes the day-mode, T is the total number of observed trips, and R is the observed ridership, which we obtain from the City of Chicago's Data Portal.

A.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.

We calculate market shares at the (origin Community Area, destination Community Area, hour-of-the-week) level using a two-step process. First, we aggregate trips at the (origin Community Area, destination Community Area, hour-of-the-week) level. In par-

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

ticular, we compute trip counts by mode as

$$ntrips_{odtj} = \sum_{k \in \mathcal{I}(odtj)} infl_k,$$

where $\mathcal{I}(\cdot)$ gives the observations corresponding to the given origin, destination, time, mode tuple and $infl_k$ gives the inflation factor for observation k. Travel times, travel distances, and prices were computed analogously as weighted averages (with inflation factors as weights) of the times and prices for each individual trip in $\mathcal{I}(odtj)$. We let the number of car trips be the residual after subtracting public transit, taxi, TNP, and shared trips from the cellphone trips.[57] Car prices are computed as 0.6374 U.S. Dollars per trip mile, which is AAA's estimate of per mile driving costs for an average 2020 model.[58]

We then get the trip counts, prices, times, and distances for a given mode j for each origin Community Area, destination Community Area, and hour-of-the-week by averaging across the corresponding (origin Community Area, destination Community Area, hour-of-the-week)-level observations. Finally, the market size is taken to be twice the maximum number of total trips, where the max is taken over the (origin Community Area, destination Community Area, hour-of-the-week)-level observations.

To compute market shares, we need to take a stance on the size of the market, which captured how many people could be traveling at a given moment in time. We set the market size to two times the total number of trips observed in a given market. We then probe how sensitive our counterfactuals are to this assumption.

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

A.5 Sparsity of Our Data versus Survey

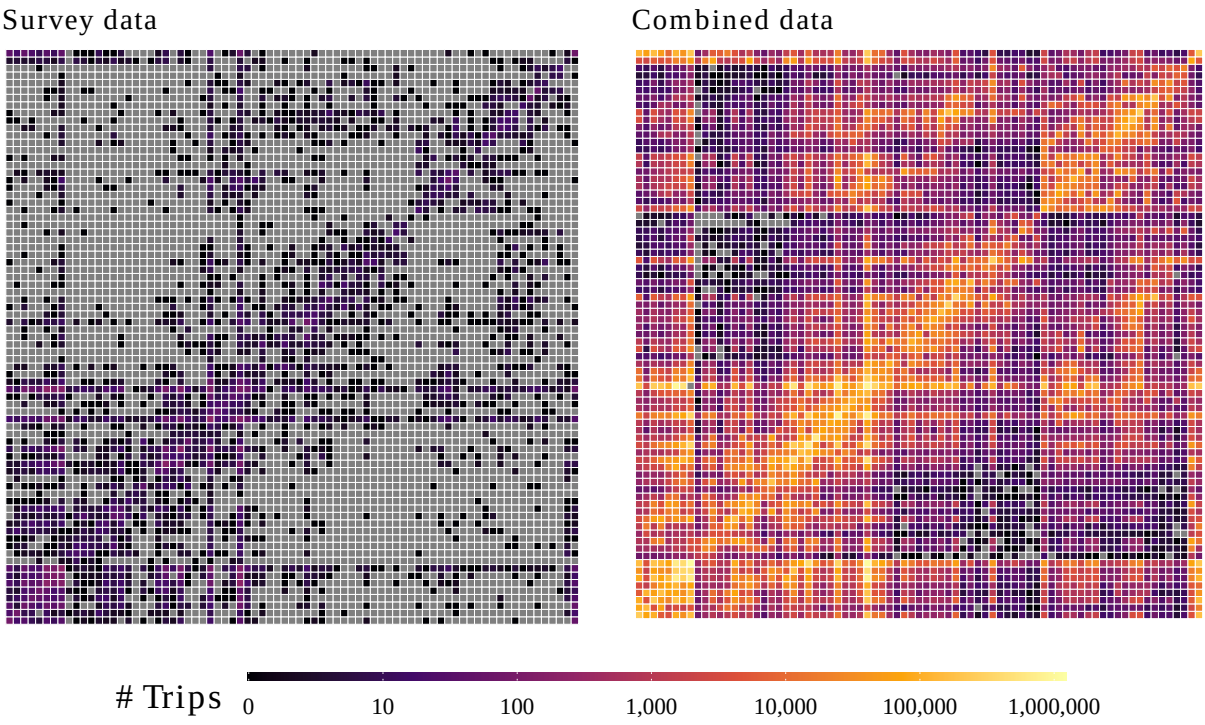


Figure 16: Combined vs. Survey Data: Flows Across Community Areas

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

B Downtown Surcharge

Effective January 6, 2020, the City of Chicago implemented a new Ground Transportation Tax structure for trips on Transportation Network Providers (TNPs) such as Uber, Lyft, and Via. The tax structure includes a $0.02 per trip administrative fee and a $0.10 per trip Accessibility Fund Fee, except for trips on Wheelchair Accessible Vehicles (WAVs).

For single TNP trips, the tax is set at $1.25 without a Downtown Zone Surcharge and $3.00 with the surcharge. If a single trip starts or ends in designated special zones, which include airports, Navy Pier, and McCormick Place, the tax is $6.25 or $8.00, respectively. For trips on WAVs, the tax is $0.55, regardless of the zone in which the trip starts or ends.

The Downtown Zone Surcharge applies to any trip that starts or ends within a spec-

ified Downtown Zone Area during peak times, which are weekdays between 6 AM and 10 PM. For shared TNP trips, the tax is $0.65 without the Downtown Zone Surcharge and $1.25 with it. If a shared trip starts or ends in the designated special zones, the tax is $5.65 or $6.25, respectively.

Before January 6, 2020, the surcharge was $0.72 for all TNC rides, at all times and in all Chicago areas.[59] This implies a $0.53 basic increase in single rides, a $0.07 decrease in pooled rides outside the zone, and extra charges in in the surcharge zone at peak hours of $1.75 and $0.60 for single and pooled rides respectively .

[59] See https://abc7chicago.com/uber-lyft-chicago-congestion-tax-taxes/5818233/

Figure 17: Downtown TNC Surcharge Area

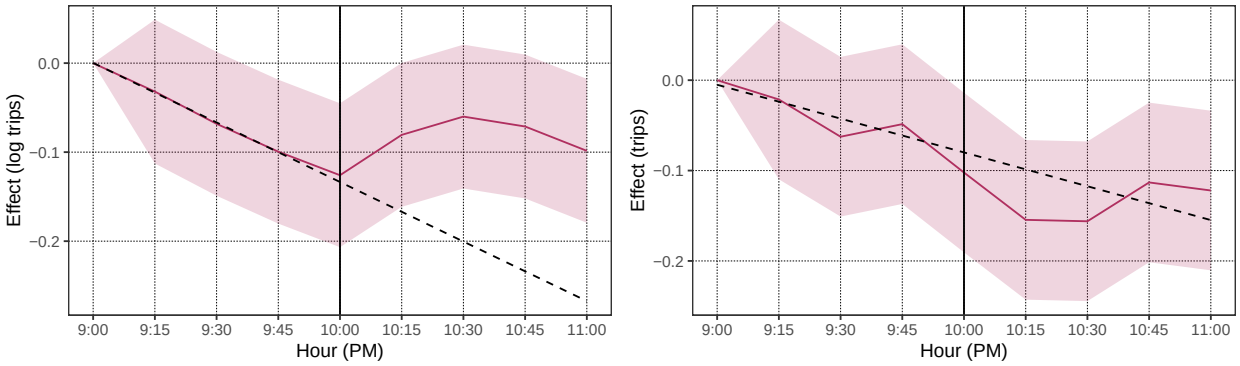
Notes: This figure shows the downtown surcharge zone. The surcharge of $3 applies to any trip that starts or ends within this zone and which are weekdays between 6 AM and 10 PM. For shared TNP trips, the tax is $0.65 without the Downtown Zone Surcharge and $1.25 with it.

We use the policy to identify a price elasticity by comparing trips that originate from end in the zone to those that originate from or end in adjacent areas around 10PM, when the policy is phased out. Concretely, our specification is

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

where is $y_{o,d,t}$ is either log price or log trips, o, d refers to origin/destination Community

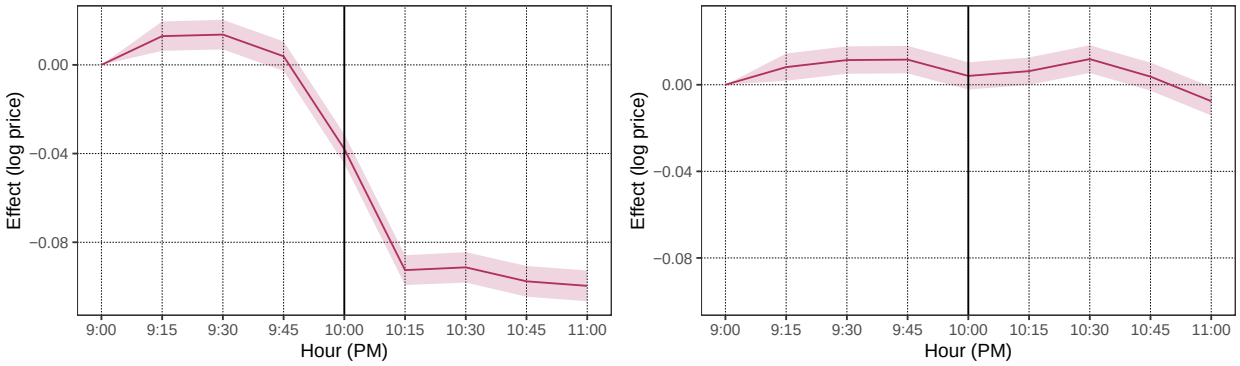
Figure 19: Evening Price Response, Surcharge Zone, 2020 (top) and 2019 (bottom)



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

Area. Time t is measured in 15-min intervals. $treat_{o,d}$ refers to all trips $o \to d$ trips subject to surcharge. We plot the coefficients of these treatment effects in Figure 18 and Figure 19. Taking both estimates, we recover an implied price elasticity of -1.42.

Figure 18: Evening Price Response, Surcharge Zone, 2020 (top) and 2019 (bottom)



Notes: The top panel shows ride hail prices of areas affected by the ride hail surcharge of $3 relative to unaffected adjacent areas around 10PM, after which the surcharge no longer applies. One can see a drop in prices, which implies an incidence of ... The bottom panel shows the same Figure in 2019, when the surcharge policy was not in place yet. We see that prices are flat.

C Proofs and Additional Theoretical Results

Proposition 2. *The first order conditions for the social planner's problem* (5) *can be written as*

$$p_j = C_j^q + E_j^q - \sum_l u_l^T \cdot T_{lj}^q + M_j^q + \frac{\lambda}{1+\lambda} \cdot B_j^q \tag{16}$$

$$u_j^T T_{jj}^k = C_j^k + E_j^k - \sum_l u_l^T \cdot \tilde{T}_{lj}^k + M_j^k + \frac{\lambda}{1+\lambda} \cdot B_j^k \tag{17}$$

$$M_j^q = \sum_{l \neq k} D_{lj} \left(C_l^q + E_l^q - \sum_m u_m^T \cdot T_{ml}^q - p_l \right)$$

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

$$B_j^q = \underbrace{\sum_{k \in \mathcal{J}_G} q_k \cdot \Omega_{kj}}_{\substack{\textit{Market power} \\ \textit{markup}}} - E_j - \underbrace{\sum_l (\tilde{u}_l^T - u_l^T) \cdot T_{lj}^q}_{\substack{\textit{Spence} \\ \textit{distortion}}} + (\tilde{M}_j^q - M_j^q)$$

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

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

$$B_j^k = E_j^k + \underbrace{\sum_k (\tilde{u}_k^T - u_k^T) \cdot T_{lj}^k}_{\substack{\textit{Spence} \\ \textit{distortion}}}$$

where λ is the Lagrange multiplier for the budget constraint and $\tilde{u}_j^T$ is a weighted sum of the derivative of gross utility among marginal travelers *with respect of the total time if pickup times for mode j increase by 1%.*

C.1 Proof of proposition 1

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

$$U(q,T(q,k)) - C(q,k) - E(q,k) - \lambda\left(C(q,k) - \sum_j p_j(q,T(q,k))q_j - B\right)$$

The first order condition for q_j is:

$$\frac{\partial U}{\partial q_j} + \sum_k \frac{\partial U}{\partial t_k}\frac{\partial T_k}{\partial q_j} - \frac{\partial C}{\partial q_j} - \frac{\partial E}{\partial q_j} + \lambda\left(p_j + \sum_k q_k \frac{dp_k}{dq_j} - \frac{\partial C}{\partial q_j}\right) = 0$$

The first order condition for k_j is:

$$\sum_k \frac{\partial U}{\partial t_k}\frac{\partial T_k}{\partial k_j} - \frac{\partial C}{\partial k_j} - \frac{\partial E}{\partial k_j} + \lambda\left(\sum_k q_k \frac{dp_k}{dk_j} - \frac{\partial C}{\partial k_j}\right) = 0$$

We now show that $\partial U/\partial q_j = p_j$. Let $\partial\Theta_j(p,t)$ be the boundary between $\Theta_j(p,t)$ and $\Theta_0(p,t)$, and let $\partial\Theta_{jk}(p,t)$ be the boundary between $\Theta_j(p,t)$ and $\Theta_k(p,t)$. Gross utility can be written as

$$U(q,t) = \int_{\Theta_j(q,t)} u(t_j,\theta) f(\theta)\, d\theta$$

Using the Leibniz integral rule, we get that

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

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

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

Since agents in the boundaries are indifferent between two choices, $u_k(t_k,\theta) = p_k$ for the first sum and $u_k(t_k,\theta) - u_l(t_l,\theta) = p_k - p_l$ for the second sum. After substituting and

rearranging terms, our main expression can be written as:

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

The term in parentheses is how much $\Theta_k(p,t)$ expands in total into all other regions, so it is equal to $\partial q_k / \partial q_j$. It is thus equal to 1 for j and 0 for $k \neq j$. We can thus conclude that $\partial U(q,t)/\partial q_j = p_j$. Substituting $\partial U/\partial q_j = p_j$, $dp_k/dq_j = \partial p_k/\partial q_j + \partial p_k/\partial t_j \cdot \partial T_k/\partial q_j$, and $dp_k/dk_j = \partial p_k/\partial t_j \cdot \partial T_k/\partial k_j$ on the first order conditions and a few steps of algebra yield

$$p_j = \frac{\partial C}{\partial q_j} - \frac{1}{1+\lambda} \sum_k \frac{\partial U}{\partial T_k} \frac{\partial T_k}{\partial q_j} + \frac{1}{1+\lambda} \frac{\partial E}{\partial q_j} - \frac{\lambda}{1+\lambda} \sum_k \left(q_k \frac{\partial p_k}{\partial q_j} + q_k \sum_{k'} \frac{\partial p_k}{\partial T_{k'}} \frac{\partial T_{k'}}{\partial q_j} \right), \quad (18)$$

$$\frac{\partial C}{\partial k_j} = \frac{1}{1+\lambda} \sum_k \frac{\partial U}{\partial T_k} \frac{\partial T_k}{\partial k_j} - \frac{1}{1+\lambda} \frac{\partial E}{\partial k_j} + \frac{\lambda}{1+\lambda} \sum_k q_k \sum_{k'} \frac{\partial p_k}{\partial T_{k'}} \frac{\partial T_{k'}}{\partial k_j}. \quad (19)$$

The final term in equation (18) can be written as

$$\sum_k q_k \sum_{k'} \frac{\partial p_k}{\partial T_{k'}} \frac{\partial T_{k'}}{\partial q_j} = \sum_k \Big(\sum_{k'} q_{k'} \frac{\partial p_{k'}}{\partial T_k} \Big) \frac{\partial T_k}{\partial q_j}$$

We now show that $\sum_{k'} q_{k'} \Big(\partial p_{k'} / \partial T_k \Big)$ can be written as a weighted average of the change in gross utility among marginal travelers. By Leibniz's integration rule,

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

where

$$W_j(p,t) = \int_{\partial\Theta_j(p,t)} v_j(\theta) \cdot \hat{n}_j(p,t,\theta) f(\theta)\, d\theta$$

and

$$W_{jk}(p,t) = \int_{\partial\Theta_{jk}(p,t)} v_{jk}(\theta) \cdot \hat{n}_{jk}(p,t,\theta) f(\theta)\, d\theta$$

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

the boundary $\partial\Theta_x(p,t)$ at θ. Also by Leibniz's integration rule,

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

where

$$V_j(p,t) = \int_{\partial\Theta_j(p,t)} \frac{\partial u_j(t,\theta)}{\partial t} v_j(\theta) \cdot \hat{n}_j(p,t,\theta) f(\theta)\, d\theta$$

and

$$V_{jk}(p,t) = \int_{\partial\Theta_{jk}(p,t)} \frac{\partial u_j(t,\theta)}{\partial t} v_{jk}(\theta) \cdot \hat{n}_{jk}(p,t,\theta) f(\theta)\, d\theta$$

These are similar integrals as before, only that the integrand is the density of riders that are willing to switch modes in response to an increase in pickup times.

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

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

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

$$\tilde{u}_j^T \equiv \sum_k q_k \frac{\partial p_k}{\partial t_j} = \sum_k q_k \Psi_{kj}$$

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

Substituting $\tilde{u}_j^T \equiv \sum_k q_k \partial p_k / \partial t_j$ and the definitions of $\epsilon_{jk}^{T,q}$, $\epsilon_{jk}^{T,k}$, $\bar{u}_j^T$, Ω_{kj}, $\tilde{C}_j^q$, and $\tilde{E}_j^q$ into expressions (18) and (19) yields expression (17) and

$$p_j = \underbrace{\frac{\partial C}{\partial q_j}}_{\substack{\text{Mg. cost} \\ \text{of a trip}}} + \underbrace{\frac{\partial E}{\partial q_j}}_{\substack{\text{Mg. env. ext.} \\ \text{of a trip}}} - \sum_k \underbrace{\bar{u}_k^T}_{\substack{\text{Mg. utility} \\ \text{of time}}} \underbrace{\frac{\partial T_k}{\partial q_j}}_{\substack{\text{Congestion} \\ \text{effect}}} + \frac{\lambda}{1+\lambda} \left\{ \underbrace{\sum_{k \in J} q_k \Omega_{kj}}_{\substack{\text{"Monopolist"} \\ \text{markup}}} - \frac{\partial E}{\partial q_j} + \underbrace{\sum_k (\tilde{u}_k^T - \bar{u}_k^T) \frac{\partial T_k}{\partial q_j}}_{\substack{\text{Spence} \\ \text{distortion}}} \right\}. \tag{20}$$

To obtain expression (6), multiply (17) by k_j/q_j and subtract it from (20). □

C.2 Generalizing Section 3.2 to multiple markets

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

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

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

$$\max_{\sigma} U(\mathbf{q}(\sigma), T(\mathbf{q}(\sigma), \mathbf{k}(\sigma))) - C(\mathbf{q}(\sigma), \mathbf{k}(\sigma)) - E(\mathbf{q}(\sigma), \mathbf{k}(\sigma)) + \lambda \left[\sum_{mj} p_{mj}(\sigma) q_{mj}(\sigma) - C(\mathbf{q}(\sigma), \mathbf{k}(\sigma)) \right] \tag{21}$$

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

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

$$0 = \sum_{mj} p_{mj} \frac{dq_{mj}}{d\sigma} + \sum_{nkmj} \frac{\partial U}{\partial t_{nk}} \frac{\partial t_{nk}}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} + \sum_{nkr} \frac{\partial U}{\partial t_{nk}} \frac{\partial t_{nk}}{\partial k_r} \frac{dk_r}{d\sigma} - \sum_{mj} \frac{\partial C}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} - \tag{22}$$

$$\sum_{mj} \frac{\partial E}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} - \sum_{r} \frac{\partial C}{\partial k_r} \frac{dk_r}{d\sigma} - \sum_{r} \frac{\partial E}{\partial k_r} \frac{dk_r}{d\sigma}$$

$$+\lambda \Bigg\{ \sum_{mjnk} q_{nk} \frac{\partial p_{nk}}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} + \sum_{mjnkol} q_{nk} \frac{\partial p_{nk}}{\partial t_{ol}} \frac{\partial t_{ol}}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} +$$

$$\sum_{mj} p_{mj} \frac{dq_{mj}}{d\sigma} - \sum_{mj} \frac{\partial C}{\partial q_{mj}} \frac{dq_{mj}}{d\sigma} - \sum_{r} \frac{\partial C}{\partial k_r} \frac{dk_r}{d\sigma} \Bigg\} \tag{23}$$

Suppose that σ is a price instrument, in which case $\frac{dk_r}{d\sigma}$ is equal to zero for all r. Then,

after some algebra, this first order condition can be written as

$$p_j^\sigma = C_j^\sigma + E_j^\sigma - U_j^{A,\sigma} + M_j^{W,\sigma} + \frac{\lambda}{1+\lambda}\left\{\mu_j^\sigma - E_j^\sigma - \Delta U_j^\sigma + \Delta M_j^\sigma\right\} \tag{24}$$

where

$$w_{mj}^\sigma = \frac{\frac{dq_{mj}}{d\sigma}}{\sum_n \frac{dq_{nj}}{d\sigma}} \qquad D_{kj}^\sigma = \frac{\sum_m \frac{dq_{mk}}{d\sigma}}{\sum_m \frac{dq_{mj}}{d\sigma}}$$

$$p_j^\sigma = \sum_m p_{mj} w_{mj}^\sigma \qquad C_j^\sigma = \sum_m \frac{\partial C}{\partial q_{mj}} w_{mj}^\sigma \qquad E_j^\sigma = \sum_m \frac{\partial E}{\partial q_{mj}} w_{mj}^\sigma$$

$$U_j^{A,\sigma} = \sum_{nkm} \frac{\partial U}{\partial t_{nk}} \frac{\partial t_{nk}}{\partial q_{mj}} w_{mj}^\sigma \qquad U_j^{M,J,\sigma} = \sum_{mnkol} q_{nk} \Phi_{nkol}^{J,\sigma} \frac{\partial t_{ol}}{\partial q_{mj}} w_{mj}^\sigma \qquad \mu_j^\sigma = \sum_{nkm} q_{nk} \Omega_{nkmj}^{J,\sigma} w_{mj}^\sigma$$

$$M_j^{W,\sigma} = \sum_{k \neq j} D_{kj}^\sigma (C_k^\sigma + E_k^\sigma - U_k^{A,\sigma} - p_k^\sigma) \qquad M_j^{\Pi,\sigma} = \sum_{k \in J \setminus \{j\}} D_{kj}^\sigma (C_k^\sigma + \mu_k^\sigma - U_k^{M,\sigma} - p_k^\sigma)$$

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

This equation resembles very closely equation (6). The key insight when generalizing it to this setting with many markets and a coarse policy lever σ is that the relevant price, marginal cost, marginal externality, network effects, and diversion ratios are weighted averages of individual-market quantities across markets. The weight given to market m is $w_{mj}^\sigma = \frac{\frac{dq_{mj}}{d\sigma}}{\sum_n \frac{dq_{nj}}{d\sigma}}$: to what extent does a change in σ affect the number of trips in market m.

For the specific case of a per-km congestion tax, which affects prices proportionally to travel distance, one can find a related expression that describes the tax in per-km form. The key point is to note that 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 follows:

$$\tau = \frac{1}{r_j^\sigma}\left(E_j^\sigma - U_j^{A,\sigma} + M_j^{W,\sigma}\right), \tag{25}$$

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

with the optimal congestion tax.

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

$$-\tilde{U}^{A,k,\sigma} = C^{k,\sigma} + E^{k,\sigma} - U^{A,k,\sigma} + M^{W,k,\sigma} + \frac{\lambda}{1+\lambda}\left\{-E^{k,\sigma} - \Delta U^{k,\sigma} + \Delta M^{k,\sigma}\right\} \tag{26}$$

where

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

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

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

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

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

Once again, this equation resembles equation (17) very closely. Quantities are also aggregated across markets through a weighted average in which the weight given to market m is $w_{mj}^\sigma = \frac{\frac{dq_{mj}}{d\sigma}}{\sum_n \frac{dq_{nj}}{d\sigma}}$.

D Model Details

D.1 Model of Waiting Times for Ride-hailing and Taxis

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

$$T_{ahj}^W = A_{aj}^W I_{ahj}^{-\phi_j} \tag{27}$$

A^W_{aj} is a scale factor that measures the overall matching inefficiency for mode j in location a. The parameter ϕ_j is an elasticity that determines how quickly waiting times decrease with the number of idle drivers. This flexible specification nests simple models of matching in taxi and ride hailing markets.[60]

To determine the number of idle drivers in every location, we assume that the distribution of drivers across the city arises from a parsimonious model that captures the spatial dynamics of the market. Let L_{hj} be the total number of drivers working for mode j during hour h. The number of drivers that are busy is given by $B_{hj} = \sum_{od} T^{\text{vehicle}}_{odh} q_{odhj}$, where T^{vehicle}_{odh} are the travel times from the traffic congestion model, and q_{odhj} is the number of people taking mode j from o to d. Thus, the total number of idle drivers is given by $I_{hj} = L_{hj} - B_{hj}$.

The probability that an idle driver is in location a during hour h is given by

$$\frac{\exp(\mu_a + \sum_b B_{ab} F_{hb})}{\sum_{a'} \exp(\mu_{a'} + \sum_b B_{a'b} F_{hb})}, \tag{28}$$

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

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

$$I_{ahj} = (L_{hj} - B_{hj}) \frac{\exp(\mu_a + \sum_b B_{ab} F_{hb})}{\sum_{a'} \exp(\mu_{a'} + \sum_b B_{a'b} F_{hb})}. \tag{29}$$

This expression, coupled with equation (27), determines the waiting times for taxis and ride hailing.

[60] In the taxi model in Lagos (2003), for instance, $\phi_j = 1$. In the simplest ride-hailing model described by Castillo *et al.* (2018), $\phi_j = 1/n$ in n-dimensional space.

Estimation We first estimate the parameters A^W_{ahj} and ϕ_j that map the number of idle drivers into waiting times. Consider Community Area a. We make the simple assumption that the I_{ahj} available drivers are distributed homogeneously across a and that the pickup time conditional on distance is $t(x) = M_{aj}x^{c_j}$. That implies that the pickup time has a distribution whose expectation is[61]

$$T^W_{ahj} = M_{aj}\Gamma\left(1+\frac{c_j}{2}\right)\left(\frac{1}{\pi I_{ahj}}\right)^{\frac{c_j}{2}}. \tag{30}$$

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

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

We then move on to estimate the parameters of the driver location model (μ_a, λ, and ρ). We do not observe drivers directly, but we use Uber data for the average waiting time at the Community Area by hour of the week level—i.e., T^W_{ahj}. Inverting equation (30) allows us to compute all values of I_{ahj}. We can then estimate (μ_a, λ, and ρ) by maximum likelihood, based on equation (28). Maximizing this likelihood is not a simple problem since the vector of μ_a has 77 elements. We simplify the task by splitting the problem into an inner loop that computes the optimal vector of μ_a given λ and ρ using a contraction mapping, as in Berry *et al.* (1995), and an outer loop that maximizes over λ and ρ. Table 6 presents our main estimates.

[61] With a density of idle drivers I_{ahj}, the pdf of the distance to the nearest driver is given by $2\pi x I_{ahj}e^{-\pi I_{ahj}x^2}$, a Weibull distribution with parameters $k = 2$ and $\lambda = 1/\pi L$. We integrate the travel time over this density to obtain equation (30).

Table 6: Driver Movement Estimates

	Coefficient	Standard Error
λ	0.0419	0.00007
ρ	-0.1312	0.0101

Notes: Standard errors are computed using a sandwich estimator.

D.2 Additional Parameters and Assumptions

- Social cost of carbon: $190 per ton
- Cost of local pollutants: 44.93 cents per gallon for gasoline, 41.32 cents per gallon for diesel, based on Holland *et al.* (2016).
- Mg. cost of cars: $0.35 / km, based on AAA cost of driving
- Uber/taxi wages: $10 / h
- Mg. cost of buses (including labor): $7.60 / km. This is the sum of
  - Capital costs of $900,000 per bus, each one of which lasts 250,000 miles
  - Fuel costs of $3.5 per gallon with a fuel efficiency of 3.38 mpg
  - Wages of $33 per hour at 20 km/h, multiplied by a factor of 2 to account for benefits and the wages of supervisors, schedulers, etc.
  - $2.51 per km of maintenance costs.
- Mg. cost of trains (including labor): $12.90 / km. This is the sum of:
  - Capital costs of $11M per train that lasts 2M miles
  - Energy costs twice the fuel costs of a bus
  - Wages of $33 per hour at 20 km/h, multiplied by a factor of 2 to account for benefits and the wages of supervisors, schedulers, etc.
  - $5 per km of maintenance costs.
- Number of passengers per private car: 1.5
- Number of ride-hailing passengers per trip : 1.3

D.3 Equilibrium computation

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

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

Algorithm 1 Equilibrium computation using Broyden's method

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

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

D.4 Optimization

Based on the procedure we describe in Appendix D.3, which we use to compute an equilibrium, we can compute welfare $W(\mathbf{p}, \mathbf{t})$ and the net revenue of the city $\Pi(\mathbf{p}, \mathbf{t})$. The welfare maximization problem is

$$\max_{\mathbf{p},\mathbf{t}} W(\mathbf{p}, \mathbf{t}). \tag{31}$$

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

The social planner's problem is

$$\max_{\mathbf{p},\mathbf{t}} W(\mathbf{p}, \mathbf{t}) \qquad \text{s.t.} \qquad \Pi(\mathbf{p}, \mathbf{t}) = -B, \tag{32}$$

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

$$\max_{\mathbf{p},\mathbf{t}} W(\mathbf{p}, \mathbf{t}) - \lambda_n \left(\Pi(\mathbf{p}, \mathbf{t}) + B\right) + \mu_n \left(\Pi(\mathbf{p}, \mathbf{t}) + B\right)^2. \tag{33}$$

We initialize this iterative procedure by setting $\mu_0 = 10^{-6}$ and $\lambda_0 = 0$. In every step n we use the method we described above to maximize the objective function, and we set $\mu_{n+1} = 2\mu_n$ and $\lambda_{n+1} = \lambda_n + \mu_n(\Pi^n + B)$, where Π^n is the net revenue at the n-th step optimum. In this algorithm λ_n converges to the Lagrange multiplier that results in the budget constraint being satisfied with equality (Nocedal and Wright, 2006). This means that (33) converges to the true Lagrangian plus an extra penalty for deviations from the budget constraint—and thus, the sequence of solutions converge to the solution of (32).

E Additional Empirical Results

E.1 First-Stage Coefficients

Table 7: First-Stage Coefficients

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

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

E.2 Demand Robustness

To assess the robustness of our demand estimation results, we estimate a number of additional specifications. Results are shown in Table 8. We first relax the assumption that travelers have utility that is linear in time by including the square of time. The estimated coefficient on the square of time is $-.312$, implying that the disutility of travel time is increasing in the length of the trip. In particular, the marginal disutility of the first minute is about half as much as the marginal disutility of the 60th minute.[62] Measured at the average trip length, the average VOT and time elasticity are both lower than in our main estimates. The average price elasticity is similar. It follows from equation 6 that counterfactuals using this alternative specification would therefore lead to larger frequency reductions and higher congestion prices (to a first order). Therefore, our results should be interpreted as conservative lower bounds on frequency reductions and road prices.

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

In our third robustness specification we additionally allow for heterogeneity in time sensitivity by income. We again adopt a Box-Cox function form: $\alpha_T^i = \alpha_T + \frac{\alpha_{Ty}}{y_i^{1-\lambda_T}}$.[63] The estimated coefficients imply a similar average VOT, price sensitivity, and time sensitivity as in our main results. However, the dispersion in VOT is compressed because we estimate that low-income individuals who are more price elastic are also more time elastic. While this would mute the dispersion in distributional consequences that we estimate in our counterfactual results, the results would remain qualitatively unchanged since lower-income individuals still exhibit significantly lower VOT than higher-income individuals.

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

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

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

Table 8: Demand Estimation Robustness

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

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

E.3 Bus Utilization

While our model does not consider capacity constraints for buses when solving for the optimal policy, we can consider *ex-post* whether this constraint might bind. Our results imply frequency reductions for buses ranging from approximately 20-30%. 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 20 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.

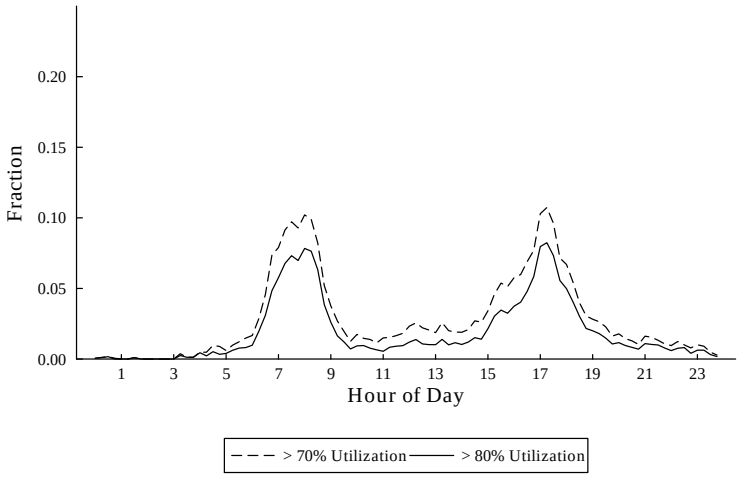


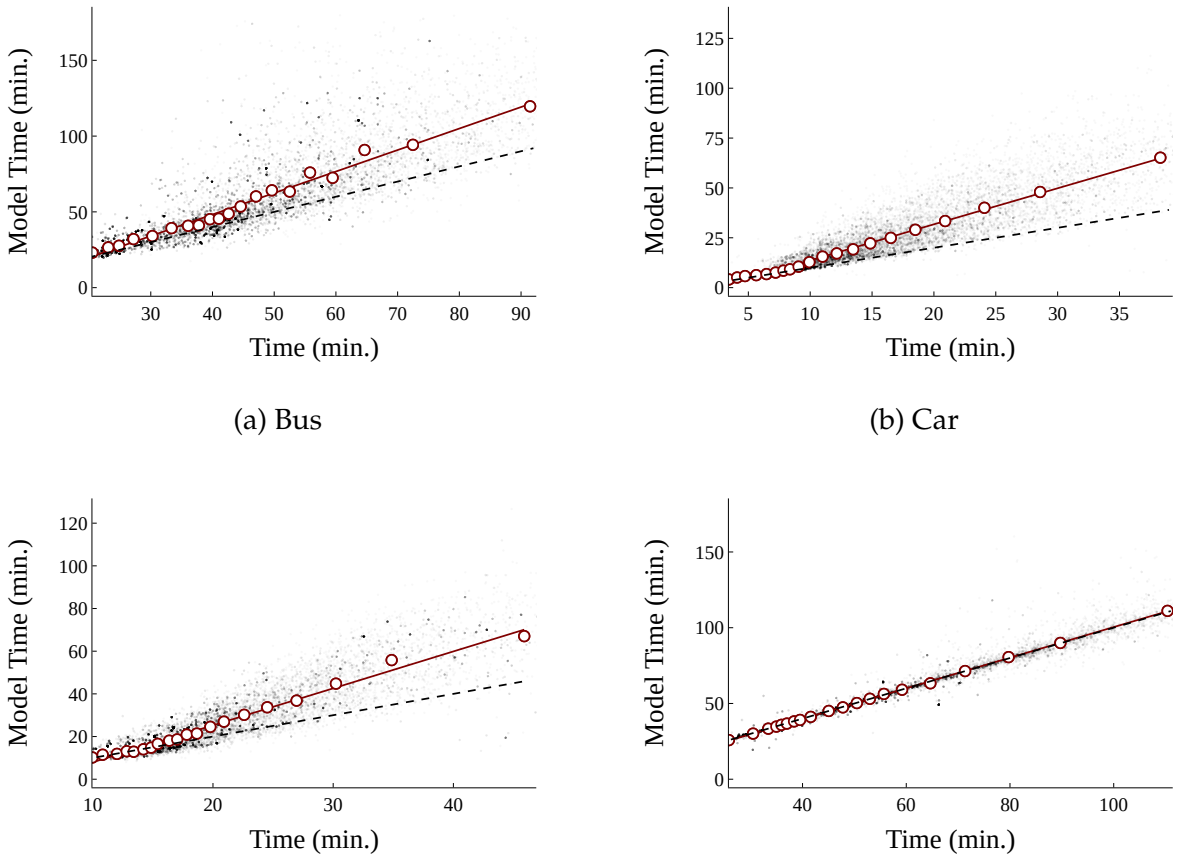
Figure 20: Bus Capacity

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

F Additional Counterfactual Results

F.1 Model Fit

F.1.1 Trip Times



(a) Bus

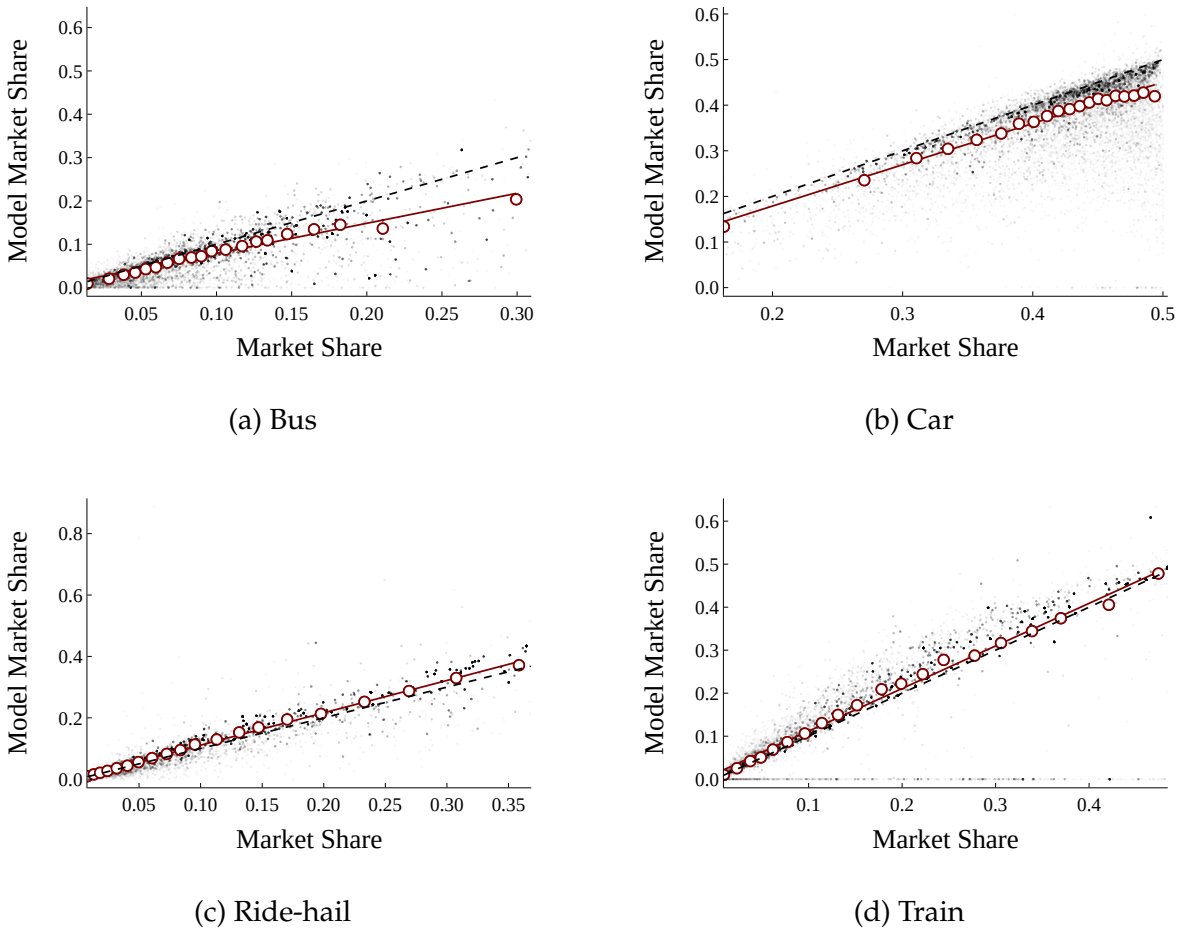
(b) Car

(c) Ride-hail

(d) Train

Notes: This figure compares observed trip times to model trip times separately for each mode. Each panel displays both a binscatter and a scatterplot of observed vs. model trip times 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.

F.1.2 Market Shares



(a) Bus

(b) Car

(c) Ride-hail

(d) Train

Notes: This figure compares observed market shares to model market shares separately for each mode. Each panel displays both a binscatter and a scatterplot of observed vs. model trip times 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.

F.2 Decomposition of welfare effects

In this section, we decompose the change in consumer surplus and in environmental externalities attributed to different channels. The change in consumer surplus is a product of two forces: the direct change in prices and the indirect effect in time due to changes in mode choices. The change in the environmental externalities is also due to two channels: the change in services levels (capacity) and the change in traveler's mode choices (substitution). Table 9 shows how each of these channels contribute to the overall aggregate effects across different scenarios.

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

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

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

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

Next, we zoom in on how consumer surplus changes across the income distribution. The results, in percentage terms, can be seen in Figure 23. Observe that the absolute effect of prices for the public transit counterfactuals is larger for lower income consumers, as

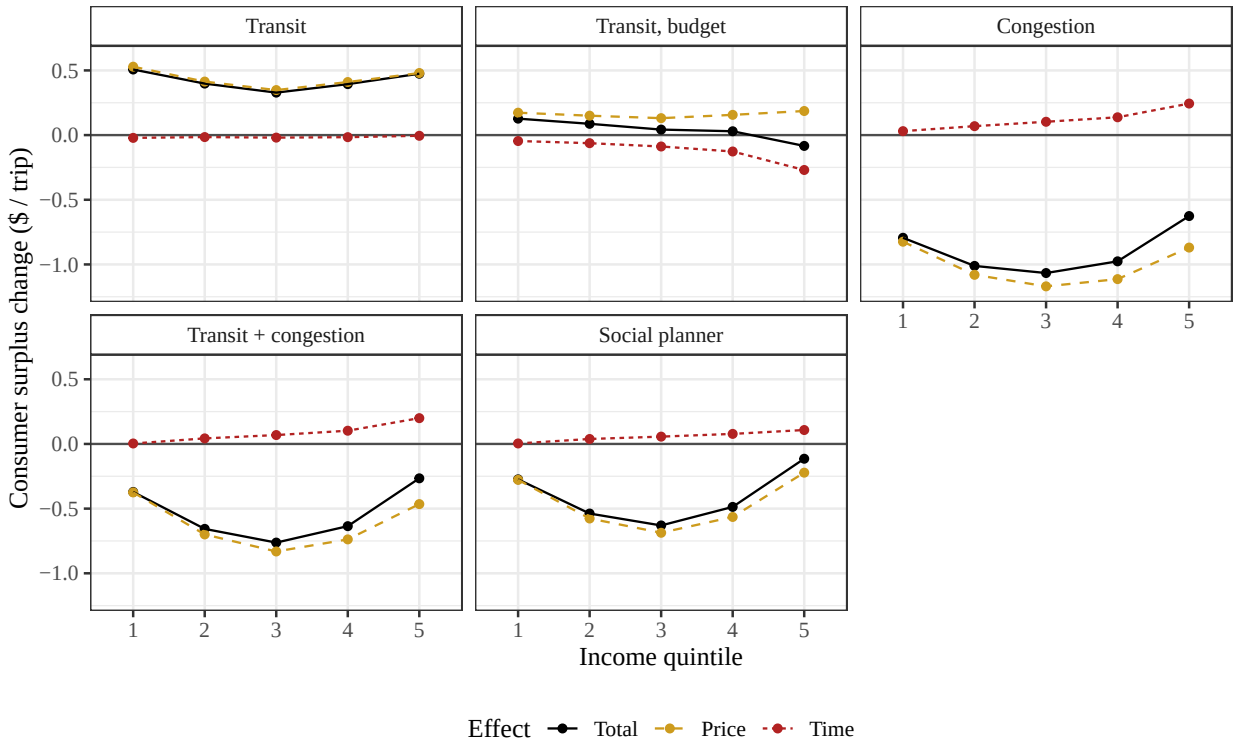
they are the ones who are more likely to use those modes of transportation. Conversely, the effect of time is more pronounced for higher income consumers, as they are the ones with the highest VOT.

Table 9: Decomposition of Consumer Surplus and Environmental Externalities

		Status quo	Transit	Transit, budget	Road pricing	Transit + Road pricing	Social planner
Δ CS ($M/week)	Total	0	10.369	0.742	-25.119	-16.198	-12.196
	Price	0	10.825	3.973	-28.058	-18.236	-13.118
	Time	0	-0.455	-3.231	2.939	2.038	0.922
	Capacity	0	-1.247	-3.325	0	-1.213	-1.004
	Substitution	0	0.791	0.094	2.939	3.251	1.927
Δ Externality ($M/week)	Total	0	-0.808	-0.686	-3.040	-3.450	-3.218
	Capacity	0	-0.374	-0.526	0	-0.365	-0.351
	Substitution	0	-0.434	-0.160	-3.040	-3.085	-2.867
Δ Average speed (km/h)		0.00%	0.52%	0.16%	2.68%	2.87%	2.59%

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

Figure 23: 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.