Online Appendix

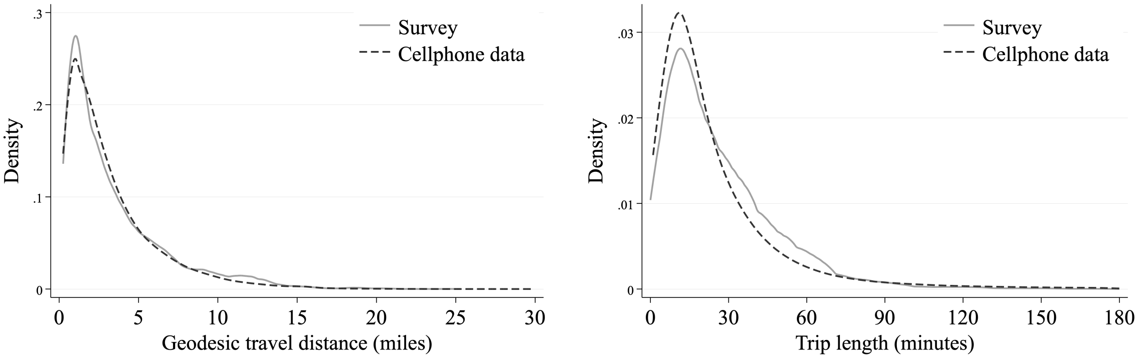
A Data Construction and Validation

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

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

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

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



(a) Travel distance

(b) Travel time

Figure A1: Representativeness of travel patterns

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

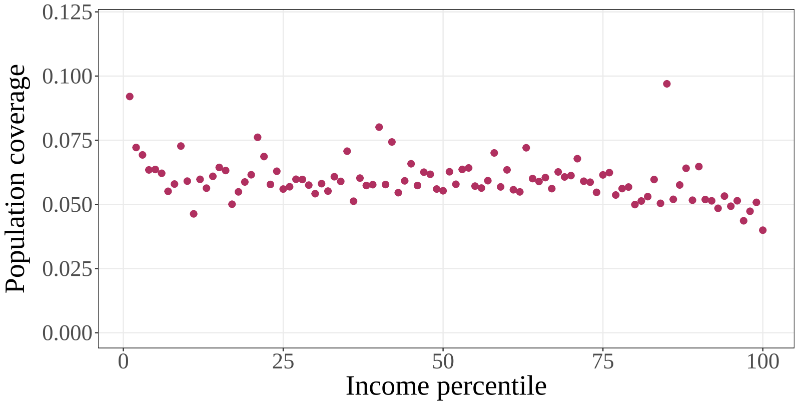


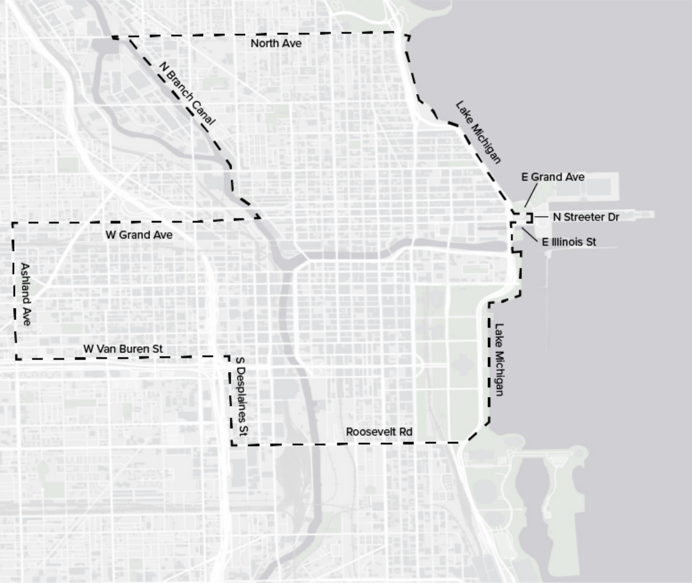
Figure A2: Representativeness across income groups

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

B Downtown Surcharge

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

Figure A3: Downtown TNC surcharge area



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

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

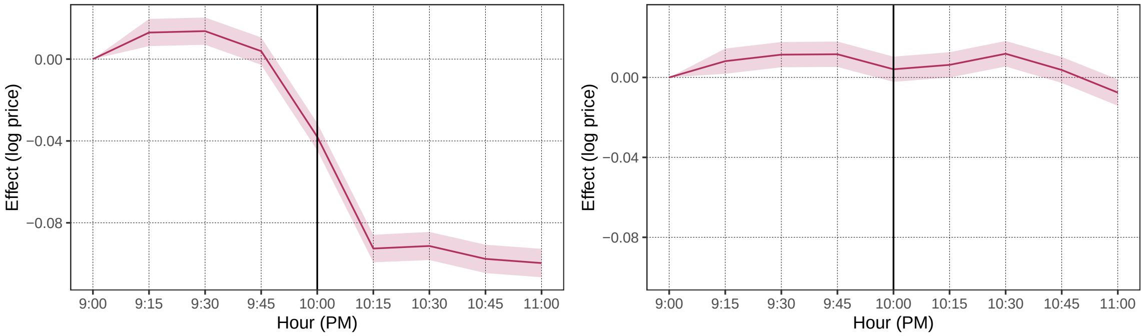
[37] See on the website of the City of Chicago.

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

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

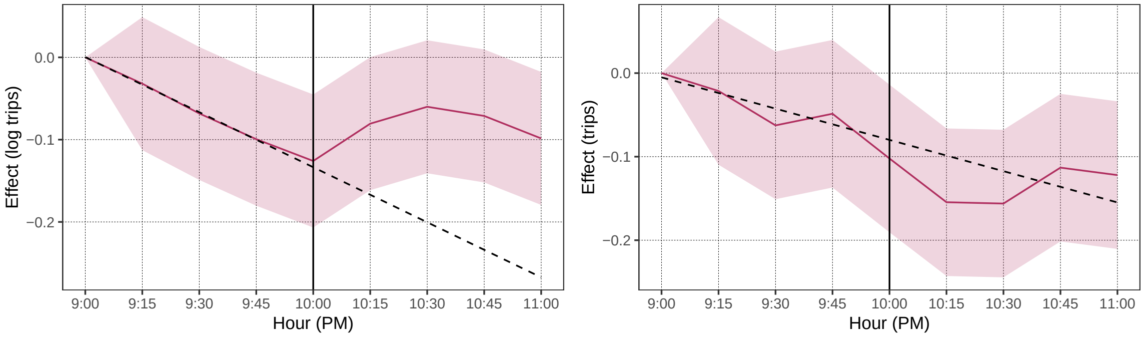
where $y_{o,d,t}$ is either log price or log trips, o, d refers to origin/destination CA. Time t is measured in 15-min intervals. $treat_{o,d}$ refers to all trips $o \to d$ trips subject to surcharge. We plot the coefficients of these treatment effects in Figure A4 and Figure A5. Taking both estimates,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

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



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

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



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

C Proofs and Additional Theoretical Results

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

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

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

C.1 Optimality condition for fleet size

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

$$\begin{aligned}\overbrace{u_j^T \check{T}^k_{jj}}^{\substack{\textit{Direct benefit}\\ \textit{of fleet size}}} = \overbrace{C_j^k}^{\substack{\textit{Mg. cost}\\ \textit{of fleet size}}} + \overbrace{E_j^k}^{\substack{\textit{Mg. env. externality}\\ \textit{of fleet size}}} - \overbrace{\sum_l u_l^T \cdot \tilde{T}^k_{lj}}^{\substack{\textit{Network}\\ \textit{effects}}} + \overbrace{M_j^k}^{\textit{Diversion}} + \\ \frac{\lambda}{1+\lambda}\Bigg(E_j^k + \underbrace{\sum_k (\tilde{u}_k^T - u_k^T) \cdot T^k_{lj}}_{\substack{\textit{Spence}\\ \textit{distortion}}} + \underbrace{\tilde{M}_j^k - M_j^k}_{\substack{\textit{Diversion}\\ \textit{distortion}}}\Bigg), \qquad (15)\end{aligned}$$

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

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

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

Proof. See Appendix C.2 □

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

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

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

C.2 Proof of Propositions 1 and 2

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

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

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

The first order condition for p_j is:

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

The first order condition for k_j is:

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

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

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

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

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

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

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

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

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

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

First, by Leibniz's integral rule,

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

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

Also by Leibniz's integral rule,

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

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

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

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

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

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

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

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

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

C.3 Generalizing Propositions 1 and 2 to multiple markets

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

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

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

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

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

where

$$C^{k,\sigma} = \sum_r \frac{\partial C}{\partial k_r}\frac{dk_r}{d\sigma} \qquad E^{k,\sigma} = \sum_r \frac{\partial E}{\partial k_r}\frac{dk_r}{d\sigma} \qquad U^{A,k,\sigma} = \sum_{nkr} \frac{\partial U}{\partial t_{nk}}\frac{\partial t_{nk}}{\partial k_r}\frac{dk_r}{d\sigma}$$
$$\Delta q_k^\sigma = \sum_m \frac{dq_{mk}}{d\sigma} \qquad \tilde{U}^{M,k,J,\sigma} + U^{M,k,J,\sigma} = \sum_{nkolr} q_{nk}\Phi_{nkol}^{J,\sigma}\frac{\partial t_{ol}}{\partial k_r}\frac{dk_r}{d\sigma}$$
$$M^{W,\sigma} = \sum_{k,m} \Delta q_k^\sigma (C_k^\sigma + E_k^\sigma - U_k^{A,\sigma} - p_k^\sigma) \qquad M^{\Pi,\sigma} = \sum_{k\in J,m} \Delta q_k^\sigma (C_k^\sigma + \mu_k^\sigma - U_k^{M,\sigma} - p_k^\sigma)$$
$$\Delta U^{k,\sigma} = U_j^{M,k,J,\sigma} - U_j^{A,k,\sigma} \qquad \Delta M^{k,\sigma} = M_j^{\Pi,k,\sigma} - M_j^{W,k,\sigma},$$

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

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

D Model Details

D.1 Model of Waiting Times for Public Transit

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Estimation We first estimate the parameters A^W_{ahj} and ϕ_j that map the number of idle drivers into waiting times. Consider CA a. We make the simple assumption that the I_{ahj} available drivers are distributed homogeneously across a and that the pickup time conditional on distance is $t(x) = M_{aj} x^{c_j}$. That implies that the pickup time has a distribution whose expectation is[40]

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

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

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

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

Table A1: Driver Movement Estimates

	Coefficient	Standard Error
λ	0.0419	0.00007
ρ	-0.1312	0.0101

Notes: Standard errors are computed using a sandwich estimator.

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

D.3 In-vehicle time adjustment

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

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

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

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

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

D.4 Additional Parameters and Assumptions

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

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

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

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

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

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

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

D.5 Equilibrium computation

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

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

Algorithm 1 Equilibrium computation using Broyden's method

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

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

D.6 Optimization

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

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

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

With a budget constraint, the welfare maximization problem is

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

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

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

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

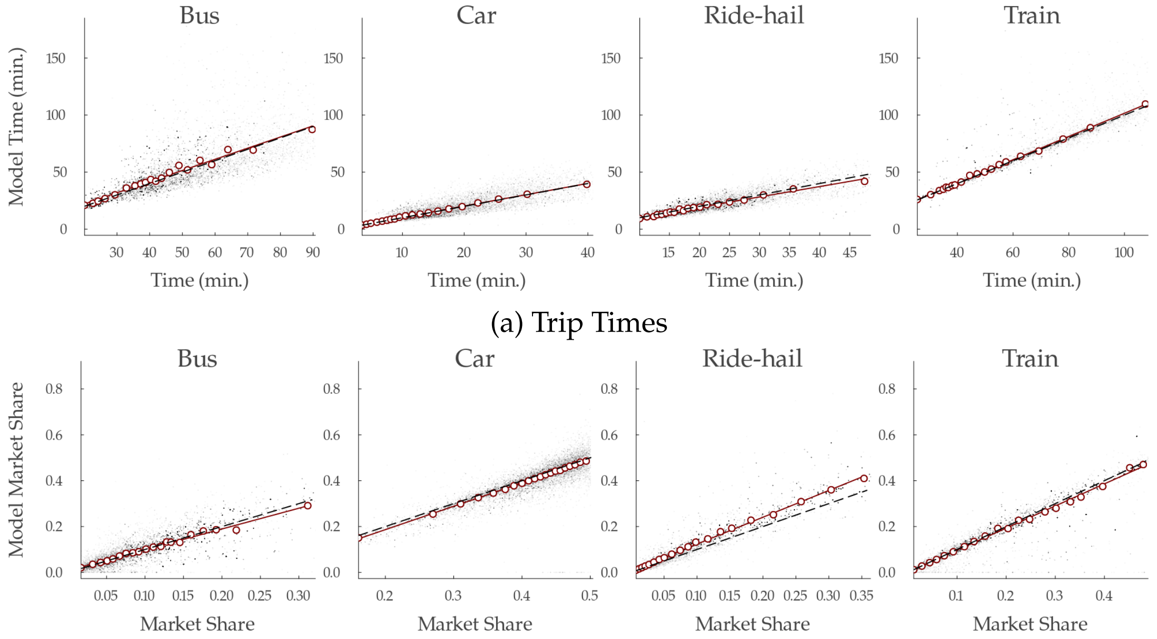
E Model Fit

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

F Sensitivity Analysis

Figure A7 shows the extent to which our main results are sensitive to some of the key parameters of our model, focusing on the *Transit + Road Pricing* counterfactual. Each panel shows how a 10% increase in several parameters of the model affect the five choice variables of the city government. In the first two panels, we see that the result indicating public transit prices should be close to zero is very robust: optimal prices are always within 1.2 cents of our baseline results. On the other hand, our results about optimal wait time for public transit are more sensitive to parameters, as can be seen in the third and fourth panels. For five of the six parameters (the marginal cost of public transit, the price and time sensitivity of travelers, the relative disutility of walking and waiting, and the variability of bus arrivals), a 10% change in the value of the parameter results in changes to the optimal bus wait time on the order of 0.6 minutes and in the optimal train wait time on the order of 0.2 minutes. These changes correspond to changes in frequencies of around 5%. Finally, the last panel shows that the road price is also quite robust: in every case, the optimal value is within 1.5 cents per km of the baseline value.

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



(a) Trip Times

(b) Market Shares

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

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

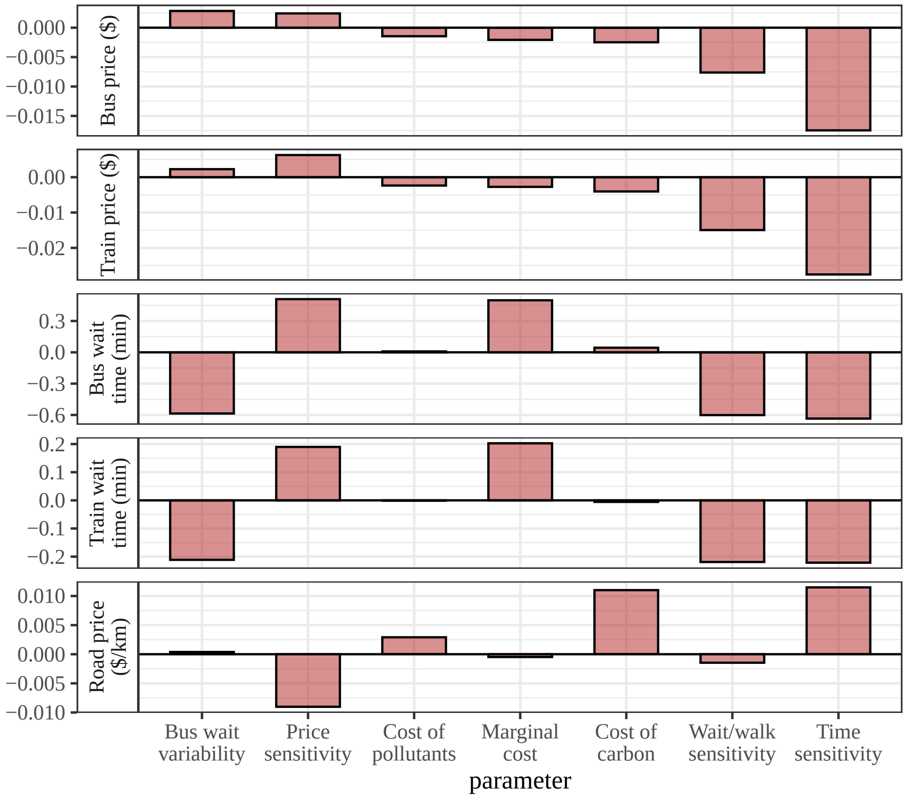


Figure A7: Robustness of counterfactual results

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

G Additional Results

G.1 Demand Robustness

To assess the robustness of our demand estimation results, we estimate a number of additional specifications. Results are shown in Table A2. We first relax the assumption that travelers have utility that is linear in time by including the square of time. The estimated coefficient on the square of time is $-.312$, implying that the disutility of travel time is increasing in the length of the trip. In particular, the marginal disutility of the first minute is about half as much as the marginal disutility of the 60th minute.[42] Measured at the average trip length, the average VOT and time elasticity are both lower than in our main estimates. The average price elasticity is similar. It follows from equation 5 that counterfactuals using this alternative specification would therefore lead to larger frequency reductions and higher congestion prices (to a first order). Therefore, our results should be interpreted as conservative lower bounds on frequency reductions and road prices.

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

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

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

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

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

Table A2: Demand Estimation Robustness

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

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

G.2 Decomposition of train prices and waiting times

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

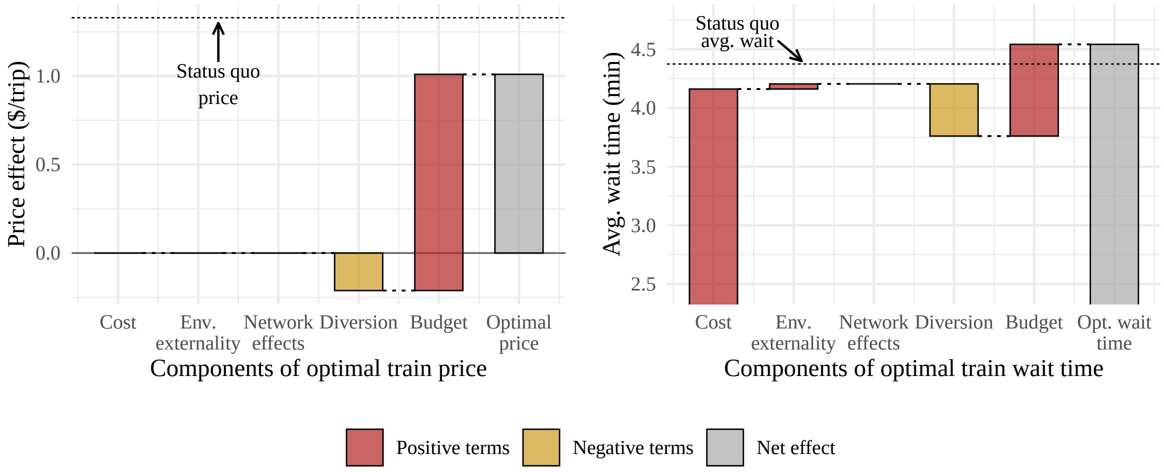


Figure A8: Decomposition of optimal price and waiting times for trains

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

Additional References

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

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

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

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

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

Supplementary Appendix

S1 Data Construction

S1.1 Cellphone location records

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

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

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

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

[45] 40 meters per second, i.e. about 145 kilometers per hour

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

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

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

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

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

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

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

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

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

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

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

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

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

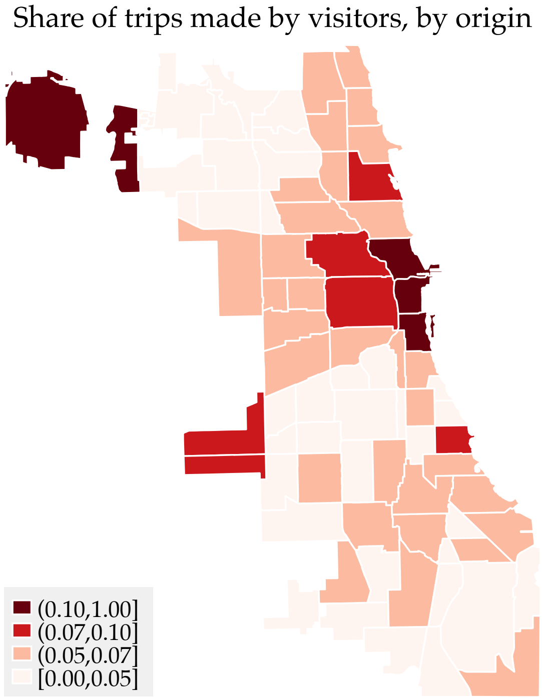
For all cellphones with an assigned home location, we impute their income by using the census tract median household income.[48]

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

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

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

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



Share of trips made by visitors, by origin

Figure S9: Share of visitors by origin location

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

S1.1.1 Survey Data Sparsity

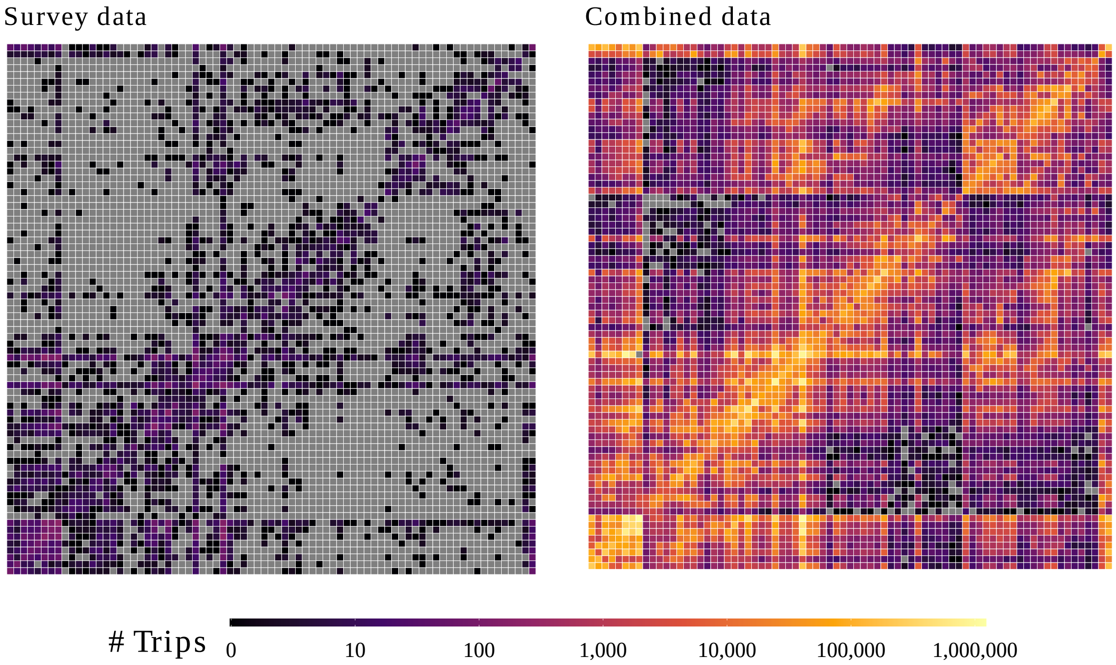


Figure S10: Combined vs. Survey Data: Flows Across Community Areas

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

S1.2 Travel times, routes, and schedules

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

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

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

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

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

S1.3 Constructing Mode-Specific Trips

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

S1.4 Market Share Calculations

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

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

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

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

S1.5 Market Size

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

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

S2 Additional Empirical Results

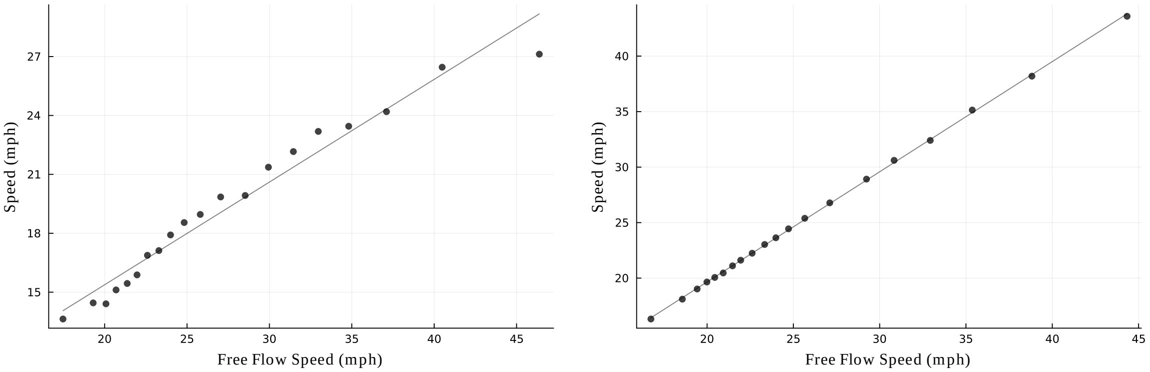
S2.1 First-Stage Coefficients

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

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

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

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



(a) 8am

(b) 12am

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

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

Table S3: First-Stage Coefficients

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

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

S2.2 Bus Utilization

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

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

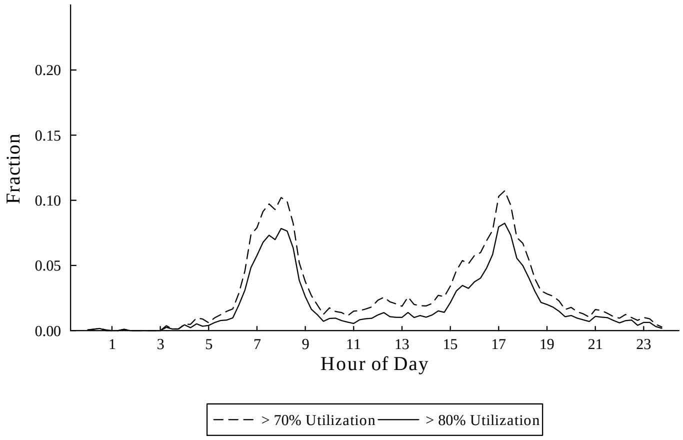


Figure S12: Bus Capacity

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

S3 Additional Counterfactual Results

S3.1 Decomposition of welfare effects

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

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

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

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

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

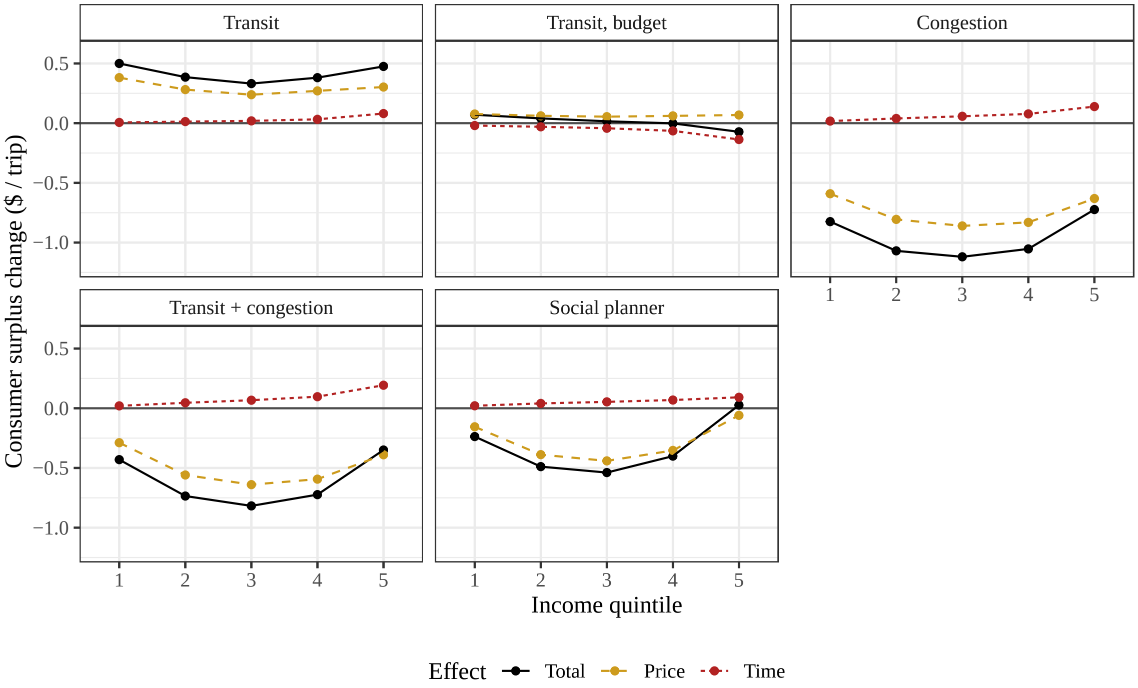
Next, we zoom in on how consumer surplus changes across the income distribution. The results, in percentage terms, can be seen in Figure S13. Observe that the absolute effect of prices for the public transit counterfactuals is larger for lower income consumers, as they are the ones who are more likely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

Table S4: Decomposition of Consumer Surplus and Environmental Externalities

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

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

Figure S13: Decomposition of consumer surplus through different channels



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