Appendix A. Analytical Derivations

Equation (10): Agency optimization over route density (D) and vehicle size (n)

Combining (1), (4), and (5), the household's indirect utility function in (7) is defined by

$$\text{(A1)} \quad \tilde{U} = \tilde{u}\left(\{p^{ij}, t^{ij}, w^{ij}, a^{ij}, c^{ij}\}, TAX\right) - Z$$

$$= \underset{X,\{M^{ij}\},\lambda}{Max} \left\{ \begin{array}{l} u\left[X,\ M(\{M^{ij}\}),\ \Gamma\left(\Sigma_{ij} t^{ij} M^{ij},\ \Sigma_{ij} w^{ij} M^{ij}, \Sigma_{ij} a^{ij} M^{ij}, \Sigma_{ij} c^{ij} M^{ij}\right)\right] - \Sigma_{ij} z^{ij} V^{ij} \\ + \lambda \cdot \left[I - TAX - X - \Sigma_{ij} p^{ij} M^{ij}\right] \end{array} \right\}$$

From the agency's point of view, (A1) can be transformed into a social utility function by substituting the various definitions and constraints of the system, namely, (2), (3), (8), and (9). In doing so, the revenues $\Sigma_{ij} p^{ij} M^{ij}$ in the government's budget constraint (9) cancel those in the individual's budget constraint in the last term of (A1); prices appear only insofar as M^{ij} depends on them through the consumer's demand functions. The resulting social utility function can be optimized by setting $\lambda = u_X$ (the first-order condition for X) and then by setting to zero its partial derivatives with respect to D, n, V, and either M or p. (Henceforth we omit the *ij* superscripts for simplicity and understand the preceding statement to apply to each i and j.) Here it is convenient to use M as the agency's choice variable; that is, we hold M constant in taking the other three derivatives. We consider two of those in this subsection, deferring the third (V) till later. Each is a partial derivative, holding the other three variables constant. Thus, in optimizing route density and vehicle size, we hold constant M and V, which implies also that occupancy $o \equiv M/V$ is constant.

Route density affects user waiting and access costs, and vehicle size affects user crowding costs and agency operating costs OC. Thus each first-order condition for optimization has two terms, and each term involves only the same i and j, so we can continue to omit the *ij* superscripts without ambiguity:

$$\text{(A2)} \quad 0 = \frac{\partial \tilde{U}}{\partial D} = \tilde{U}_w \frac{\partial w}{\partial D} + \tilde{U}_a \frac{\partial a}{\partial D} = -\lambda M \rho^W w_f \frac{\partial f}{\partial D} - \lambda M \rho^A a_D$$

$$\text{(A3)} \quad 0 = \frac{\partial \tilde{U}}{\partial n} = \tilde{U}_c \frac{\partial c}{\partial n} - \tilde{U}_I \frac{\partial OC}{\partial n} = -\lambda M \rho^C c_l \frac{\partial l}{\partial n} - \lambda V t \frac{dK}{dn}$$

where w_f, a_D, and c_l are derivatives of the functions defined in (3a), and we have used the definitions of ρ^k from (6b). The partial derivatives on the right-hand sides of (A2) and (A3) can be computed using definitions (3) and (8), holding V, M, and o constant. This yields $\partial f/\partial D = -f/D$, $\partial l/\partial n = -l/n$, and $dK/dn = k_2$. Inserting these and dividing each equation by λM yields (10).

Equation (11). Marginal welfare effects of reduction in peak-rail fare

Partially differentiating (A1) and applying (6b) gives

(A2a) $\tilde{U}_{p^{ij}} \equiv -\lambda M^{ij}$; $\tilde{U}_{t^{ij}} = -\lambda\rho^T M^{ij}$; $\tilde{U}_{w^{ij}} = -\lambda\rho^W M^{ij}$; $\tilde{U}_{a^{ij}} = -\lambda\rho^A M^{ij}$; $\tilde{U}_{c^{ij}} = -\lambda\rho^C M^{ij}$;

(A2b) $\tilde{U}_{TAX} = -\lambda = -u_X$; $\quad \tilde{U}_Z = -1$

Totally differentiating (A1) shows that when the agency changes peak-rail price p^{PR}, utility changes according to

(A3)
$$\frac{d\tilde{U}}{dp^{PR}} = \tilde{U}_{p^{PR}} + \sum_{ij}\left\{\tilde{U}_{t^{ij}}\frac{dt^{ij}}{dp^{PR}} + \tilde{U}_{w^{ij}}\frac{dw^{ij}}{dp^{PR}} + \tilde{U}_{a^{ij}}\frac{da^{ij}}{dp^{PR}} + \tilde{U}_{c^{ij}}\frac{dc^{ij}}{dp^{PR}}\right\}$$
$$+\tilde{U}_{TAX}\frac{dTAX}{dp^{PR}} - \sum_{ij} z^{ij}\frac{dV^{ij}}{dp^{PR}}$$

From (A2) and (A3), we obtain

(A4a)
$$MW^{PR} \equiv -\frac{1}{\lambda}\frac{d\tilde{U}}{dp^{PR}} = M^{PR} + \sum_{ij} M^{ij}\left\{\rho^T\frac{dt^{ij}}{dp^{PR}} + \rho^W\frac{dw^{ij}}{dp^{PR}} + \rho^A\frac{da^{ij}}{dp^{PR}} + \rho^C\frac{dc^{ij}}{dp^{PR}}\right\}$$
$$+\frac{dTAX}{dp^{PR}} + \frac{1}{\lambda}\sum_{ij} z^{ij} V_M^{ij}\frac{dM^{ij}}{dp^{PR}}$$

where $V_M^{ij} \equiv dV^{ij}/dM^{ij}$ is a constant ($1/o^{iCAR}$) for j=CAR and depends on the transit agency's operating policy for j=B, R. To keep track of its parts, we write the components of (A4a) as

(A4b)
$$MW^{PR} = M^{PR} + USERTIM + WAITACC + CROWD + \frac{dTAX}{dp^{PR}} + POLLACC$$

where *WAITACC* includes the terms involving ρ^W and ρ^A and *POLLACC*, the last term in (A4a), represents changes in pollution and accident externality costs.

We can compute $dTAX/dp^{PR}$ by rearranging (9) with only *TAX* on the left-hand side, differentiating it, and using (2a) and (8) to get

$$\frac{dTAX}{dp^{PR}} = -M^{PR} - \sum_i \tau^{iA} V_M^{iCAR} M_{PR}^{iCAR} - \sum_i \sum_{j \neq CAR} p^{ij} M_{PR}^{ij} + \sum_i \sum_{j \neq CAR} K^{ij} \cdot \left(t^{ij} V_M^{ij} M_{PR}^{ij} + V^{ij} \frac{dt^{ij}}{dp^{PR}} \right)$$

(A5a)

$$+ \sum_i \sum_{j \neq CAR} t^{ij} V^{ij} k_2^{ij} \frac{dn^{ij}}{dp^{PR}}$$

where we hold constant τ^{iA} and all transit prices other than p^{PR}. It is convenient to write the terms in (A5a) as changes in particular financial flows:

(A5b) $$\frac{dTAX}{dp^{PR}} = -M^{PR} - FUELREV - TRANSITREV + (OPSUPPLY + OPCONG) + VEHSIZE$$

where the first term is changes in peak-rail revenue from existing passengers; the second is changes in fuel tax revenue; the third is changes in transit fare revenue due to mode and time-of-day shifts; the fourth is changes in transit operating cost related to travel time (divided into two parts: changes due to shifts among different modes and times of day with different average supply costs, and effects of congestion); and the last is changes in transit operating cost related to vehicle size. Note that new revenues reduce the lump-sum *TAX* that must be levied, whereas new costs increase it.

Substituting (A5b) into (A4b), we see that the terms M^{PR} cancel, and we can rearrange the other parts into a more convenient order for further calculation, as follows:

$$MW^{PR} = (WAITACC - TRANSITREV + OPSUPPLY)$$

(A6) $$+ (USERTIM + OPCONG + POLLACC - FUELREV)$$

$$+ (CROWD + VEHSIZE)$$

It is useful to summarize the definitions of elasticities of bus and rail travel characteristics, recalling that all are defined so as to be positive:

(A7a) $\eta_w^{ij} = -\dfrac{f^{ij}}{w^{ij}} w_f^{ij}$, $\quad \eta_a^{ij} = -\dfrac{D^{ij}}{a^{ij}} a_D^{ij}$, $\quad \eta_c^{ij} = \dfrac{l^{ij}}{c^{ij}} c_l^{ij}$, $\quad \forall i, j \neq \text{CAR}$

(A7b) $\varepsilon_V = \dfrac{M^{ij}}{V^{ij}} V_M^{ij} = o^{ij} V_M^{ij}$, $\quad 1 - \varepsilon_V = \dfrac{M^{ij}}{o^{ij}} o_M^{ij} = V^{ij} o_M^{ij}$, $\quad \forall i, j \neq \text{CAR}$

We also define how service frequency and route density change with vehicle miles, and how vehicle size and load factors change with occupancy, as follows:

(A7c) $\varepsilon_f^{ij} = \dfrac{V^{ij}}{f^{ij}} f_V^{ij} = D^{ij} f_V^{ij}$, $\quad 1 - \varepsilon_f^{ij} = \dfrac{V^{ij}}{D^{ij}} D_V^{ij} = f^{ij} D_V^{ij}$, $\quad \forall i, j \neq \text{CAR}$

(A7d) $\varepsilon_n^{ij} = \dfrac{o^{ij}}{n^{ij}} n_o^{ij} = l^{ij} n_o^{ij}$, $\qquad 1-\varepsilon_n^{ij} = \dfrac{o^{ij}}{l^{ij}} l_o^{ij} = n^{ij} l_o^{ij}$, $\qquad \forall i, j\neq$CAR

We now proceed to compute key derivatives in (A4a) and (A5a) in terms of $M_{PR}^{ij} \equiv dM^{ij}/dp^{PR}$. The travel time derivative can be written, using (2a), (3), and (A7b), as

(A8a) $$\frac{dt^{ij}}{dp^{PR}} = (t_{CAR}^{ij} / o^{iCAR}) M_{PR}^{iCAR} + (t_B^{ij} / o^{iB}) \varepsilon_V M_{PR}^{iB} + (\theta^j / V^{ij})(1-\varepsilon_V) M_{PR}^{ij}$$

where $t_{CAR}^{ij} \equiv dt^j / dV_{iCAR}$ and $t_B^{ij} \equiv dt^j / dV_{iB} = \alpha_B t_{CAR}^{ij}$. Note that $t_{CAR}^{iR} = t_B^{iR} = 0$ by our assumption that rail speeds are unaffected by road traffic. Similarly, the waiting, access, and crowding derivatives in (A4), which apply only for $j\neq CAR$, can be written using (2), (3), and (A7) as

(A8b) $$\frac{dw^{ij}}{dp^{PR}} = w_f^{ij} \frac{df^{ij}}{dM^{ij}} M_{PR}^{ij} = -w^{ij} \eta_w^{ij} \varepsilon_f^{ij} \varepsilon_V M_{PR}^{ij} / M^{ij}$$

(A8c) $$\frac{da^{ij}}{dp^{PR}} = a_D^{ij} \frac{dD^{ij}}{dM^{ij}} M_{PR}^{ij} = -a^{ij} \eta_a^{ij} \cdot (1-\varepsilon_f^{ij}) \varepsilon_V M_{PR}^{ij} / M^{ij}$$

(A8d) $$\frac{dc^{ij}}{dp^{PR}} = c_l^{ij} \frac{dl^{ij}}{dM^{ij}} M_{PR}^{ij} = c^{ij} \eta_c^{ij} (1-\varepsilon_n^{ij})(1-\varepsilon_V) M_{PR}^{ij} / M^{ij}$$

We now examine the terms in (A6) in groups. We begin by using (A8b) and (A8c) to compute *WAITACC* as given in (A4), using (10a) to simplify:

(A9) $$\begin{aligned} WAITACC &= \sum_i \sum_{j\neq CAR} \left(\rho^W \frac{dw^{ij}}{dp^{PR}} + \rho^A \frac{da^{ij}}{dp^{PR}} \right) M^{ij} = -\sum_i \sum_{j\neq CAR} \rho^W w^{ij} \eta_w^{ij} \varepsilon_V M_{PR}^{ij} \\ &= -\sum_i \sum_{j\neq CAR} MB_{scale}^{ij} M_{PR}^{ij} \end{aligned}$$

where the last equality applies definition (12b). This accounts for all the terms in (11) involving MB_{scale}. As for the other terms in the first group in (A6), we note that *TRANSITREV*, the third term in (A5a), accounts for all the terms in (11) involving p. We also see that *OPSUPPLY*, as defined by the first of the two terms involving K^{ij} in (A5a), can be written using (A7b) as

(A10) $$OPSUPPLY = \sum_i \sum_{j\neq CAR} (\varepsilon_V / o^{ij}) K^{ij} t^{ij} M_{PR}^{ij} = \sum_i \sum_{j\neq CAR} MC_{supply}^{ij} M_{PR}^{ij}$$

where the last equality uses definition (12a). Thus *OPSUPPLY* accounts for all the terms in (11) involving MC_{supply}.

We now turn to the second group of terms in (A6). The terms *USERTIM* and *OPCONG*, which are the terms in (A4a) and (A5a) involving dt^{ij}/dp^{PR}, can be combined and written, using (A8a), as

$$USERTIM + OPCONG \equiv \sum_{ij}(M^{ij}\rho^T + K^{ij}V^{ij})\frac{dt^{ij}}{dp^{PR}}$$

$$= \sum_i \sum_{j=CAR,B}(\rho^T M^{ij} + K^{ij}V^{ij})(t^{ij}_{CAR}/o^{iCAR})M^{iCAR}_{PR}$$

$$+ \sum_i \sum_{j=CAR,B}\varepsilon_V \cdot (\rho^T M^{ij} + K^{ij}V^{ij})(t^{ij}_B/o^{iB})M^{iB}_{PR}$$

$$+ \sum_i \sum_{j=B,R}(1-\varepsilon_V)(\rho^T M^{ij} + K^{ij}V^{ij})(\theta^j/V^{ij})M^{ij}_{PR}$$

where we have adopted the notational convention that K^{iCAR}=0. Using the fact that $t^{ij}_B = \alpha_B t^{ij}_{CAR}$, the definition $o^{ij}=M^{ij}/V^{ij}$, and definitions (12d), we obtain

(A11)
$$USERTIM + OPCONG = \sum_i (MC^{iCAR}_{cong}/o^{iCAR})M^{iCAR}_{PR} + \sum_i \varepsilon_V \cdot (MC^{iB}_{cong}/o^{iB})M^{iB}_{PR}$$
$$+ \sum_i \sum_{j=B,R}(1-\varepsilon_V)MC^{ij}_{dwell}M^{ij}_{PR}$$

These terms are components of sums of MC^{ij}_{ext} as defined in (12c). Next we obtain some other components of those same sums. Using the definition of ε_V and the fact that λ=u_X, the change in external costs of pollution and accidents is

(A12) $$POLLACC \equiv \frac{1}{\lambda}\sum_{ij} z^{ij}V^{ij}_M M^{ij}_{PR} = \sum_i \frac{z^{iCAR}}{u_X}\cdot\frac{1}{o^{iCAR}}\cdot M^{iCAR}_{PR} + \sum_i \sum_{j\neq CAR}(\varepsilon_V/o^{ij})\cdot\frac{z^{ij}}{u_X}\cdot M^{ij}_{PR}$$

Finally, the fuel tax revenue term in (A5) is

(A13) $$-FUELREV = \sum_I (-\tau^{iCAR}/o^{iCAR})M^{iCAR}_{PR}$$

Adding equations (A11)-(A13) and applying definitions (12c) yields

$$\sum_i MC^{iCAR}_{ext}M^{ij}_{PR} + \sum_i \sum_{j=B,R} MC^{ij}_{ext}M^{ij}_{PR}$$

which accounts for all the terms in (11) involving MC_{ext}.

Finally, we consider the last group of terms in (A6), involving crowding and the costs undertaken to avoid it. Using (A7b), A7d), (A8d), and (10b), these terms add to

$$CROWD + VEHSIZE = \sum_i \sum_{j \neq CAR} (1-\varepsilon_V)(1-\varepsilon_n^{ij})\rho^C c^{ij} \eta_c^{ij} M_{PR}^{ij} + \sum_i \sum_{j \neq CAR} t^{ij} V^{ij} k_2^{ij} n_o^{ij} o_M^{ij} M_{PR}^{ij}$$

$$= \sum_i \sum_{j \neq CAR} (1-\varepsilon_V)(1-\varepsilon_n^{ij})(t^{ij} k_2^{ij} n^{ij} / o^{ij}) M_{PR}^{ij} + \sum_i \sum_{j \neq CAR} (1-\varepsilon_V)\varepsilon_n^{ij}(t^{ij} k_2^{ij} n^{ij} / o^{ij}) M_{PR}^{ij}$$

$$= \sum_i \sum_{j \neq CAR} MC_{occ}^{ij} M_{PR}^{ij}$$

which accounts for the terms in (11) involving MC_{occ}. We have now accounted for all terms in (11), which completes the proof.

Transit agency optimization over vehicle miles of service (V).

Now consider what would happen if the agency also optimized with respect to *V*. Our model does not assume the agency actually does so, but here we use this assumption to derive a benchmark case for ε_V to use in our baseline scenario. This type of benchmark is sometimes called a "quasi-first-best" response: responding to changes according to a first-best adjustment rule even though some other factor (*e.g.* a bureaucratic bias toward offering too much service) prevents the situation from being truly first-best (Small and Verhoef 2007, p. 142).

In our case even this "quasi-first-best" value of ε_V is only approximate, as we compute it under three additional simplifying assumptions:

- Elasticities of waiting and access times (defined positively) are all equal to a common value ($\eta_w^{ij} = \eta_a^{ij} \equiv \zeta$);
- The transit agency ignores its own vehicles' contributions to congestion ($t_B^{ij} = 0$) and to other externalities (z^{iB}=z^{iR}=0);
- Dwell time for entering and exiting passengers is negligible (θ^B=θ^R=0).

The first bullet is an assumption common to the simpler models of Mohring effects—for example, that of Small (2004). A special case is when average waiting time is half the interval between vehicles, and average access time is proportional to the distance between parallel transit lines; then ζ=1.

Those assumptions enable us to derive a simple condition for maximizing (A1) with respect to the agency's choice variables, for given travel demands $\{M^{ij}\}$. In what follows, we suppress superscripts for simplicity. Maximizing with respect to *D* and *n* again yields (10). Given our first simplifying assumption, we see immediately from (10a) that average waiting cost and access cost are equated:

(A14) $\rho^W w = \rho^A a$

This result is also in Jansson (1997). Since D=V/f, it can be written as

(A15) $\rho^W \alpha_w f^{-\zeta} = \rho^A \alpha_a \cdot (V/f)^{-\zeta}$

where we have substituted in the constant-elasticity functions $w = \alpha_w f^{-\zeta}$ and $a=\alpha_a D^{-\zeta}$ describing waiting and access times, respectively. Solving (A15) for *f*, we see that it is proportional to the square root of *V*. That is, *f* is adjusted when *V* changes with elasticity ε_{fV}=½. Therefore,

(A16) $\varepsilon_{fM} = \varepsilon_{fV}\varepsilon_V = ½\varepsilon_V$

We now consider maximizing with respect to *V*. Given our second assumption, *V* affects (A1) only through the terms involving waiting time *w*, access *a*, crowding *c*, and operator cost *OC*, the latter entering through budget constraint (9). The first-order condition is therefore

$$0 = \frac{\partial \tilde{U}}{\partial V} = \tilde{U}_w \frac{\partial w}{\partial V} + \tilde{U}_a \frac{\partial a}{\partial V} + \tilde{U}_c \frac{\partial c}{\partial V} + \tilde{U}_{TAX} \frac{\partial OC}{\partial V}$$

$$= -\lambda M \rho^W w_f f_V - \lambda M \rho^A a_D D_V - \lambda M \rho^C c_l l_o \cdot (do/dV) - \lambda Kt - \lambda tVk_2 n_o \cdot (do/dV)$$

where the last equality uses the definitions in (6b) and (8) and the result λ=u^X. Dividing by λ and using (A7), (2a), and (10), this implies

$$0 = \rho^W w\zeta\varepsilon_f M/V + \rho^A a\zeta \cdot (1-\varepsilon_f)o + \rho^C c\eta_c \cdot (1-\varepsilon_n)o - k_1 t - k_2 tn \cdot (1-\varepsilon_n)$$

$$= \rho^W w\zeta M/V - k_1 t$$

or

(A17) $\frac{wM}{V} = \frac{k_1 t}{\rho^W \zeta}$

Under our second assumption, the right-hand side of (A17) is a constant as far as the agency is concerned. On the left-hand side, $w = \alpha_w f^{-\zeta}$. Therefore,

(A18) $f^{-\zeta} MV^{-1} = constant$

Now let *M* change parametrically, with all the service variables *f*, *n*, and *V* changing in response. Differentiating the logarithm of (A18) with respect to log(*M*) yields

(A19) $-\zeta\varepsilon_{fM} + 1 - \varepsilon_V = 0$

Substituting (A16) into (A19) and solving yields ε_V=2/(2+ζ). For the common case ζ=1, this yields ε_V=2/3, as in Small (2004) and a special case of Nash (1988).

The intuition for this result is somewhat subtle. If ζ is near zero, wait and access costs are relatively unaffected by vehicle miles of service, so vehicles are operated only as necessary to handle the passenger loads; thus increased passenger loads require a proportional increase in vehicle miles, *i.e.* ε_V=1. If ζ is large, the operator accounts for the substantial effects on user costs by running extra vehicles for passengers' convenience even when *M* is small; in that case, when *M* increases, the operator can absorb some of the increase through higher occupancy, thereby reaping more of the advantages of scale; this means choosing a smaller value of ε_V. We take ζ=1 as our base case (ε_V=2/3) and consider sensitivity $\zeta\in[0,2]$ by treating ε_V=1 and ε_V=½.

Appendix B. Assessment of Parameter Values

Here we describe our methodology for estimating parameter values along with data sources; Table 2, which is discussed in the text, summarizes our key estimates. For some parameters, breakdowns by mode or time of day are unavailable from statistical sources; in these cases we use various estimation procedures or our own judgment. U.K. monetary numbers are converted to U.S. dollars using the average 1998–2003 exchange rate of £1.0 = US$1.6.

System operating characteristics. Basic data are compiled from the operating agencies and various national statistics.[1] For London, we allocate total passenger miles across time of day using the observed fraction of passenger trips occurring at peak period, 0.62 for rail and 0.48 for bus, and an assumed average trip length in the peak equal to 1.6 times that in the off-peak.[2] Passenger miles per hour are then computed assuming that the peak period covers 6 hours per workday (30 hours per week) and the off-peak covers 10 hours every day (70 hours per week). We assume peak shares are each 0.05 higher for Washington (which has a high proportion of government employment) and 0.05 lower for Los Angeles (which has a smaller discrepancy between peak and off-peak vehicles per hour).[3] To obtain vehicle miles offered by time period, we assume that observed total vehicle miles are allocated across the available 100 weekly hours in proportion to passenger miles per hour to the power ε_V=0.67, our baseline assumption as discussed in Appendix A.[4]

For Washington and Los Angeles, automobile vehicle miles by time of day are from David Schrank and Tim Lomax (2003), and occupancy is from the 2001 National Household Travel Survey on average occupancy per trip in large metropolitan areas. For London, auto

[1] For the United States, see the National Transit Database (FTA 2003), and for the United Kingdom, see Transport for London (TfL 2003, Tables 1.1,1.2, 3.6), TfL (2004a, b), and U.K. DfT (2003, Tables 5.3, 5.16). Rail data encompass subways and light rail but not commuter rail.

[2] For the entire United Kingdom, commuting trips have length around twice that of trips for education, shopping, or other personal business (U.K. DfT 2003, Table 10); however, we expect a smaller discrepancy for transit trips because of the high fixed time cost of accessing transit.

[3] The Washington adjustments are in line with unpublished statistics we obtained from transit agency representatives; the Los Angeles transit authority has no such data on trips by time of day.

[4] Total vehicle miles for U.S. rail systems were obtained by multiplying vehicle-car miles by average cars per train. For peak periods the latter is calculated by the ratio of rail cars to trains; off-peak train length is assumed to be slightly lower based on common observation. For London rail, vehicle-miles (i.e. train-miles) is measured directly (in train-km) as 65.4 million train-km from London Underground Limitied (LUL 2003), p. 2; the same figure appears rounded off as 65 train-km in TfL (2003), Table 1.1.

passenger miles by time of day are from TfL (2003, Table 3.6), including car/motorcycle and taxi. Auto vehicle-miles are from TfL (2003, Tables 1.2, 31.); to allocate them across time periods we use information about person trips by time period and overall average trip distance, along with an assumed ratio of peak to off-peak trip distance of 1.3. Auto occupancy for London is then calculated as the ratio of passenger-miles to vehicle-miles.

Operating costs and fares. We assume that vehicle capital costs are proportional to capacity n, whereas other operating costs are independent of n. Thus in aggregate, vehicle capital costs constitute $k_1 tV$ and other operating costs $k_2 ntV$, using (8). Operating costs, aside from vehicle capital costs, are taken from the operating statistics of the transit agencies. For rail, we assume that 10 percent of these are the fixed cost of maintaining stations (F^{iR} in (8a)). When expressed per vehicle hour of service, we assume that the rest of these costs are 25 percent greater during peak than off-peak periods because of difficulties in scheduling labor for split shifts; hence we obtain k_1^{ij} in (8b).

As for vehicle capital costs, we estimate them ourselves by annualizing the purchase cost of a rail or bus car, assuming lifetimes of 25 and 12 years, respectively, and a real interest rate of 7 percent.[5] (One advantage of this procedure is that we need not rely on agency data for vehicle purchase costs, which may be distorted by various financing mechanisms such as tax-free bonds.) We allocate vehicle capital costs entirely to the peak period, on the assumption that any increase in vehicle miles in that period requires purchasing more vehicles, whereas an increase in the off-peak period does not; hence we obtain $k_2^{Pj} n^{Pj}$ and k_2^{Oj} = 0 in (8b). Vehicle capital costs are 27 to 52 percent of other peak variable operating costs. Thus our assumption that they are the portion of costs that is proportional to n leads to results consistent with several other studies of size-related costs, as reviewed by Small (2004, 156 and note 13).

Fares were obtained by dividing agency passenger fare revenue by passenger mileage (for Washington rail, peak fares were higher than off-peak in 2002, but the discrepancy was modest and we ignore it).

[5] We use U.S. nationwide figures for all vehicle prices (from APTA 2002, Table 60) except for Los Angeles rail, for which figures were available from www.mta.net/press/pressroom/facts_glance (where necessary, figures are updated to 2002 using the CPI for Transportation Equipment). The vehicle lifetimes chosen are common in the transit cost literature, and the interest rate is that recommended for cost-benefit analysis by U.S. OMB (1992).

Wait costs. Based on evidence summarized in Small and Verhoef (2007, 53), we assume the value of in-vehicle time ρ^T in the U.S. is half the gross wage rate; in the UK we assume the value of time is only 40 percent of the gross wage rate, due to higher labor taxes there which reduce the net wage rate relative to the gross wage. In both nations we assume the value of waiting time at transit stops, ρ^W, is $1.8\rho^T$, the midpoint of the range suggested by Small and Verhoef. The median gross wage rate is measured at \$16.93, \$14.19, and \$18.83 for Washington, Los Angeles, and London, respectively, and then expressed per minute.[6] We assume that people of different wage rates have sorted themselves into different modes and time periods initially, as follows: auto and rail travelers each have wages that are 15% above the area's median gross wage rate for peak periods, and 15 percent below for off-peak (since peak travelers are more likely to be higher-wage workers); and bus travelers have wage rates that are 80 percent of these amounts.

We obtain initial wait times and the wait time elasticity as follows. Let H be average minutes between transit vehicles at a given stop, or headway (the inverse of frequency). When H is small, it is reasonable to assume that travelers arrive randomly at a stop and incur expected wait time $H/2$. When headways are larger, at least some travelers will use transit timetables, which, following Peter Tisato (1998), we assume involves three time costs. The first two are planning and precautionary time required because the exact vehicle arrival time is uncertain; we assume these are 1 and 5 minutes, respectively, and each is valued at rate ρ^W. The third is the expected cost of early arrival at the destination, assuming the traveler chooses a transit vehicle arriving prior to the desired time to ensure against late arrival. This is $\rho^E H/2$ where ρ^E is the per minute cost of early arrival, assumed conservatively to be $0.2\rho^W$; that is, a minute of early arrival is equivalent to 0.2 minutes of extra planning or precautionary time.[7] All these costs are therefore accounted for by setting wait time per trip, $w \cdot d$, to $6+0.2H/2$ for those using a timetable.

We therefore assume that when $H<15$, all users arrive randomly, so the average wait time per trip is $w \cdot d=H/2$; whereas when $H>60$, everyone uses a schedule, so $w \cdot d=6+H/10$. In the first case, $\eta_w^{ij} \equiv (dw/dH) \cdot (H/w)=1$, whereas in the latter case, $\eta_w^{ij}=1/[1+(60/\mathrm{H})]$. For simplicity, we

[6] Wages are from U.S. Bureau of Labor Statistics (BLS 2004), TfL (2003, p. 49) and U.K. Office for National Statistics (ONS 2004).

[7] Richard Arnott, André de Palma, and Robin Lindsey (1993) and others use the value of ρ^E/ρ^T estimated by Small (1982) for work trips, which is 0.61. The ratio is likely to be much lower for non-work trips, which less often have a fixed schedule. So to be conservative we use half the work-trip value, or $\rho^E=0.305\rho^T \approx 0.2\rho^W$.

assume that the elasticity (which is what enters our calculations directly) falls gradually as a mixture of the elasticities applying to these two regimes:

$$\text{(B1)} \qquad \eta_w^{ij} = \begin{cases} 1 & \text{if } \; 0 < H \leq 15 \\ (1-\lambda) + \dfrac{\lambda}{1+(60/H)} & \text{if } \; 15 < H < 60 \\ \dfrac{1}{1+(60/H)} & \text{if } \; H \geq 60 \end{cases}$$

where λ=(H-15)/45. Substituting this value for λ, the middle part of (B1) can be written as η_w^{ij}=(4/3)–(5/3)/[1+(60/H)], from which we can see that η_w^{ij} is monotonically declining in H from a value of 1.0 at H=15 to 0.5 at H=60.[8] (It then rises gradually for H>60, but this regime is inapplicable here because in our simulations, headways never rise above 26 minutes.)

Ideally we would compute initial headways from data on car-miles, directional route-miles, and assumed duration of each time period. However some data on route-miles are unavailable, and even if they are the resulting average headways are not realistic because individual routes are heterogeneous. Therefore we use these data as guidance, but ultimately assume initial headways based on our judgment. For the US, we assume rail headways of 5 and 10 minutes (peak and off-peak), and bus headways of 12 and 25 minutes. For London, we assume smaller headways of 3 and 8 minutes for rail, 6 and 15 for bus.

Marginal benefits from scale economies and marginal cost from occupancy. These are easily computed from (12b), using above values for parameters ε_V, ρ^W, w^{ij}, η_w^{ij}, and $k_2^{ij} n^{ij}$.

Marginal congestion costs. For automobiles, MC_{cong}^{iCAR} is obtained directly from equation (12d). The travel-time functions are assumed to follow the commonly used BPR-type functional form (US Bureau of Public Roads,) in which travel delay (dropping the i superscript for simplicity) is proportional to $(V/C)^\gamma$ where $V=V^{CAR}+\alpha_B V^B$ is total traffic (in passenger-car equivalents),

[8] In the optimal subsidy and 50% subsidy calculations, where we have non-marginal price changes, we also must compute wd itself in order to obtain the full prices for use in (14). In this case we integrate (B1) to obtain ln(wd) = 5.6 + (4/3)ln$H$ – (5/3)ln(H+60). In our calculations, H>15 occurs in only two cases (US off-peak bus), when H≈25. At H=25 this formula implies wd=12.03, which suggests that about seven-eighths of travelers arrive randomly (thus having wait time H/2=12.5), while one-eighth of them use a schedule (wait time 6+H/10=8.5); thus the penetration rate of schedules is less than λ, which is 0.22 at this headway.

C=capacity, and γ is an exponent that we take to be 3.7 based on some aggregate relationships from the U.S.[9] We also assume that bus speed is a fixed proportion of car speed. It is easy to show that with these assumptions, marginal delays to cars and buses, respectively, due to an auto vehicle-mile are:

$$t_{CAR}^{CAR} \equiv dt^{CAR} / dV^{CAR} = \gamma \cdot (\text{average delay/mile})/V$$

$$t_{CAR}^{B} \equiv dt^{B} / dV^{CAR} = \left(t^{B} / t^{CAR}\right) t_{CAR}^{CAR}$$

Average delay/mile for the U.S. cities is obtained from total person hours of delay from Schrank and Lomax (2003), allocating 85 percent of it to the peak period, and dividing by passenger miles; this yields an average peak delay (in min/pass-mile) of 0.33 minutes per passenger mile for Washington and 0.49 for Los Angeles). Our data provide direct estimates of average traffic speeds in Greater London during the peak and daytime off-peak periods; we add 10 percent to the latter to account for evenings and nights (bus speeds are 10.3 and 12.3 mph in peak and off-peak period). Average delay is then inferred assuming a free-flow speed of 30 miles per hour, with a result of 1.91 min/mile peak and 1.08 min/mile off-peak. We assume the passenger-car equivalent for buses, α_B in (12d), is 4.0 for the U.S. cities and 5.0 for London, where buses are larger and cars are smaller.[10] Based on agency data, we find the ratio of auto to bus speed to be 2.8 in Washington, 2.7 in Los Angeles, and 1.6 in London. In London, about one-fourth of the marginal congestion cost MC_{cong}^{iCAR} turns out to be attributed to the effect of congestion on bus passengers and operators, despite that only about one-eighth of passengers using roadways do so on buses; the difference is due to the considerable adverse effects of congestion on agency costs.

Pollution and accident externalities. We start with nationwide average values from the assessment by Parry and Small (2005) of U.S. and U.K. automobile externalities: namely, 2.0 cents per vehicle mile for local pollution; 6 cents per gallon of gasoline for global warming; and 3.0 and 2.4 cents per vehicle mile for accidents in the United States and the United Kingdom, respectively. To account for greater population exposure in urban areas, we double the local

[9] Small (1992, 70–71) found that total delay is well approximated by a power function of traffic volume, with power 4.1 in Toronto and 3.3 in Boston; we average to set γ=3.7.

[10] U.S. Federal Highway Administration (FHWA 1997, Table V-23) gives the passenger-car equivalent as only 2.0; however, this is only for federal urban highways where buses stop very infrequently, and it excludes mileage on city and suburban streets.

pollution figure for Washington and London, and we triple it for Los Angeles, whose topography causes pollutants to disperse especially slowly. We do not adjust external accident costs because the evidence suggests that, despite higher traffic densities in urban areas, external accident risks are not necessarily higher, given the counteracting effect of slower-moving traffic (Lindberg 2001, 406–407).

Also from Parry and Small (2005), we assume fuel taxes of 40 cents per gallon for the U.S. cities[11] and 280 cents per gallon in London. We use their nationwide average fuel efficiencies of 20 and 30 miles per gallon for the off-peak period (on the assumption that most travel nationwide is in conditions similar to off-peak travel in these very large metropolitan areas) but reduce them by 25 percent in the peak period to adjust for the effect of congestion on fuel economy.

For bus, accidents costs per vehicle mile are taken to be the same as for auto because buses move more slowly and are driven by professionals, offsetting their much greater weight, but pollution is taken to be triple that for automobiles.[12] Bus global-warming costs are computed assuming fuel efficiency of 5 miles per gallon. When expressed per passenger mile, all three of these external costs are very small for bus (one cent per passenger-mile), and are taken to be zero for rail.

Our estimates of external costs omit road wear, which is negligible for autos but perhaps not for buses due to their weight and small number of axles over which it is distributed. Still, we think these are small enough to ignore. Buses probably cause marginal road damage similar to that of lighter single-unit trucks, which has been estimated at around 3 cents per vehicle-mile on urban interstate roads (U.S. FHWA 2000, Table 13), an amount that would have negligible effect on our results. The costs might be considerably higher on city streets because of their thinner pavements, but we would not expect them to dominate the congestion costs and scale economies that we find.

Dwell times. For bus, we adopt the midrange values for typical boarding and alighting times from Transportation Research Board (TRB 2000, Exhibit 27-9), assuming two doors for

[11] The federal tax was 18.4 cents per gallon; state-level taxes in California, the District of Columbia, Virginia, and Maryland were approximately 20 cents per gallon (U.S. DOC 2003, Table 730).

[12] These assumptions are consistent with estimates of relative external costs per vehicle mile for heavy trucks and autos in U.S. FHWA (1997, Table 13); separate estimates for bus are not available.

alighting and boarding. We assume cash payment for the U.S. cities and prepayment (which allows rear-door boarding) for London. This yields values of 4.275 seconds for the U.S. cities and 3.375 sec for London (for comparison, Kenneth J. Dueker et al. 2004 estimate 5.18 sec in Portland). For rail, we use the estimate by Kraus (1991, 256) from observations in Boston, which is 1.0/N_T sec where N_T is the number of cars per train. In each case we divide by trip length to specify parameter θ^{ij}. The marginal cost of increased dwell time is then calculated from (12d), using parameters already described.

Generalized price of travel. The components of q^{ij} are given by (10c); besides parameter values already described, we need the time per mile of transit vehicles t^{ij} and access and crowding elasticities η_a^{ij} and η_c^{ij}. (This is in fact the only place where we need an empirical estimate of η_c^{ij}.)

To calculate t^{ij}, we divide total vehicle miles by vehicle hours to give average speeds, over the day, of 23 and 11 miles per hour for Washington rail and bus, and 23 and 12 miles per hour for Los Angeles rail and bus. For London, we have a direct estimate of average speeds from the agency: namely 20 miles per hour for rail, 11 for bus. For all three cities we assume the ratio of peak to off-peak speed is 1.0 for rail, while for bus it is the same as that for autos: approximately 0.86 for Washington and London, 0.79 for Los Angeles.

The access-time elasticity η_a^{ij} depends on route density in a manner similar to how the wait-time elasticity depends on service frequency. It is one if people live at uniformly distributed locations and walk to the nearest transit stop, and smaller if people living farther away choose a faster access mode with a fixed cost (e.g., park and ride). The less dense the transit network, the more important these other access modes, so the lower the elasticity. We assume other access modes have minor importance in London but more in Washington and more still in Los Angeles, and so choose η_a^{ij} = 0.8, 0.65 and 0.5 for these cities, respectively.

There is little empirical basis for gauging η_c^{ij}, which is positive only for peak service; we set it to 1.5 in the baseline, though our results are not sensitive to different assumptions (because crowding costs are relatively small).

Own-price travel demand elasticities. Our model calls for elasticities of each mode's passenger demand with respect to its own generalized price q^{ij}, denoted as η_q^{ij}. However, most empirical evidence is based on elasticities with respect to fare p^{ij}, which we denote as η_p^{ij}. We first review the evidence on η_p^{ij}, then describe how we convert to η_q^{ij}.

Based on Armando M. Lago, Patrick D. Mayworm, and J. Matthew McEnroe (1981), Phil Goodwin (1992), and Richard H. Pratt et al. (2000), we assume that the own-fare demand elasticity, averaged over peak and off-peak time periods, is -0.5 for bus and -0.3 for rail,[13] and that in each case the elasticity in the off-peak period is twice that in the peak. Given that about 70 percent of passenger mileage occurs during the peak period, the values just stated imply own-fare elasticities η_p^{ij} of approximately -0.40 and -0.8 for peak and off-peak bus, and -0.24 and -0.48 for peak and off-peak rail, respectively. To convert these to generalized-price elasticities η_q^{ij}, we assume that the empirical measurement of η_p^{ij} incorporates the effects of p^{ij} on w^{ij} in (10c), as discussed in the derivation of (14c); that is, we assume

$$\text{(B2)} \qquad \eta_p^{ij} = \frac{p^{ij}}{M^{ij}} \frac{dM^{ij}}{dq^{PR}} \frac{dq^{ij}}{dp^{ij}} = \eta_q^{ij} \frac{p^{ij}}{q^{ij}} \frac{dq^{ij}}{dp^{ij}}$$

where the ratio and the derivative on the right-hand side are both obtained from (10c). Thus we simply invert equation (B2) to obtain our estimates of η_q^{ij}, which we assume to be constants.

Modal diversion ratios, m_{kl}^{ij}. Pratt et al. (2000, 12–41 ff.) provide several estimates for U.S. cities of the proportion of incremental transit trips that are diverted to or from other modes following a change in transit price; typical numbers, averaged across time of day, are about 65 percent and 80 percent for Atlanta and Los Angeles, respectively. Nearly all of these shifts are to or from cars. We assume that Washington is like Atlanta, and that peak values m_{PCAR}^{Pj} are 0.05 higher, and off-peak values m_{OCAR}^{Oj} 0.05 lower, than these average values.

Now consider the cross-elasticities between bus and rail transit. The few studies available typically find them to be about half the direct elasticities in cities with good coverage by both

[13] A recent review of mostly U.K. studies by Neil Paulley et al. (2006) produces somewhat larger long-run elasticities, which they suggest is because elasticities have risen in magnitude and are higher in the United Kingdom than in other nations. Many of the studies relied upon by Paulley et al. are unpublished, and we do not feel the evidence is strong enough to apply these higher elasticities to our U.K. simulations.

bus and rail transit systems, such as London and Chicago (Christopher L. Gilbert and Hossein Jalilian 1991, Table 3b; Antti Talvitie 1973). Assuming equal travel volume by mode, this would imply $m_{iR}^{iB} = m_{iB}^{iR} \approx 0.5$ for i=P,O. However, we expect the substitutability between modes to decrease as one expands beyond the city to the metropolitan area, and to decrease more for cities with less and less well developed rail networks. We also expect them to have declined considerably from the 1970s or 1980s to the year 2000 because of increasing competition from the automobile. Finally, in the newer U.S. transit systems the bus lines are typically reconfigured to serve as feeders to the rail system, with competitive routes discontinued. Therefore, we assume the cross-mode diversion ratios to be just 10 percent for Washington ($m_{iR}^{iB} = m_{iB}^{iR} = 0.1$) and 5 percent for Los Angeles ($m_{iR}^{iB} = m_{iB}^{iR} = 0.05$).

For London, we expect less diversion to automobile and more to the other transit mode because of the smaller initial share of automobiles and travelers' more extensive transit choices. We therefore set London's diversion ratios to be like those for Washington, except 0.20 smaller for auto in the same time period, and 0.20 larger for other transit in the same time period.

Little information is available about shifts of transit riders across time periods. We assume that in each case, 10 percent of the change in transit ridership represents such shifts, and that the shifts occur entirely to the same transit mode.

Those assumptions lead to the values shown in Table 2. The fraction of extra transit trips from increased travel demand is a residual, equal to between zero and 20 percent. The review by Pratt et al. (2000) suggests that 10 percent and 26 percent of new transit trips in Los Angeles and Atlanta, respectively, represented some combination of changes in walking, trip frequency, and destination during the 1990s. Given the likely further decline in this fraction due to metropolitan decentralization, this evidence is roughly consistent with our assumed values.

Additional References for Appendices

APTA. 2002. *Public Transportation Fact Book*. American Public Transportation Association, Washington, DC.

Arnott, Richard, André de Palma, and Robin Lindsey. 1993. "A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand." *American Economic Review* 83: 161–79.

Dueker, Kenneth J., Thomas J. Kimpel, and James G. Strathman. 2004. "Determinants of Bus Dwell Time." *Journal of Public Transportation* 7: 21–40.

Gilbert, Christopher L., and Hossein Jalilian. 1991. "The Demand for Travel and for Travelcards on London Regional Transport." *Journal of Transport Economics and Policy* 25: 3–29.

Goodwin, Phil B. 1992. "A Review of New Demand Elasticities with Special Reference to Short and Long Run Effects of Price Changes." *Journal of Transport Economics and Policy*: 26: 155–69.

Lago, Armando M., Patrick D. Mayworm, and J. Matthew McEnroe. 1981. "Further Evidence on Aggregate and Disaggregate Transit Fare Elasticities." *Transportation Research Record* 799: 42–47.

Lindberg, Gunnar. 2001. "Traffic Insurance and Accident Externality Charges." *Journal of Transport Economics and Policy* 35: 399–416.

LUL. 2003. *Environmental Report 2003*. London: London Underground Ltd. http://www.tfl.gov.uk/assets/downloads/environmental-report-2003.pdf.

Paulley, Neil, Richard Balcombe, Roger Mackett, Helena Titheridge, John Preston, Mark Wardman, Jeremy Shires, and Peter White. 2006. "The Demand for Public Transport: The Effects of Fares, Quality of Service, Income and Car Ownership." *Transport Reviews* 13: 295–306.

Pratt, Richard H., Texas Transportation Institute, Cambridge Systematics, Parsons Brinkerhoff Quade & Douglas, SG Associates, and McCollom Management Consulting. 2000. *Traveler Response to Transportation System Changes: Interim Handbook.* Transportation Research Board, Transit Cooperative Research Program Web Document 12 (March).

Schrank, David, and Tim Lomax. 2003. *The 2003 Urban Mobility Report*, Texas Transportation Institute, Texas A&M University System, College Station, Texas.

Small, Kenneth A. 1982. "The Scheduling of Consumer Activities: Work Trips." *American Economic Review* 72: 467-479.

Small, Kenneth A. 1992. *Urban Transportation Economics*. Fundamentals of Pure and Applied Economics, Volume 51, Harwood Academic Press, Chur, Switzerland.

Talvitie, Antti. 1973. "A Direct Demand Model for Downtown Worktrips." *Transportation* 2: 121–52.

TfL. 2003. *London Travel Report 2003*. London: Transport for London. www.tfl.gov.uk/tfl/reports_library_stats.shtml, accessed June 11, 2004.

TfL. 2004a. *About London Buses*. www.tfl.gov.uk/buses/cib_about.shtml, accessed July 30, 2004.

TfL. 2004b. *Amazing Facts*. http://tube.tfl.gov.uk/content/faq/facts.asp, accessed July 30, 2004.

Tisato, Peter. 1998. "Optimal Bus Subsidy and Cross Subsidy with a Logit Choice Model." *Journal of Transport Economics and Policy* 32: 331–50.

TRB. 2000. *Highway Capacity Manual*. Transportation Research Board, National Research Council, Washington, DC.

U.K. DfT. 2003. *Transport Statistics Great Britain*. Department for Transport, London. http://www.dft.gov.uk/pgr/statistics/datatablespublications/public/annualbulletins/abulletinofpublictransportst5264.

U.K. ONS. 2004. "Hours Worked: Growth Levels Off." In: *Labour Market in 2002*, Office for National Statistics. www.statistics.gov.uk/cci/nugget.asp?id=696, accessed June 11, 2004.

U.S. BLS. 2004. "2002 Metropolitan Area Occupational Employment and Wage Estimates." Bureau of Labor Statistics, Washington. http://www.bls.gov/oes/oes_2002.htm.

U.S. DOC. 2003. *Statistical Abstract of the United States 2003*. U.S. Department of Commerce, Census Bureau, Washington.

U.S. FHWA. 1997. *1997 Federal Highway Cost Allocation Study*. U.S. Federal Highway Administration, Department of Transportation, Washington. http://www.fhwa.dot.gov/policy/otps/costallocation.htm.

U.S. FHWA. 2000. *Addendum to the 1997 Federal Highway Cost Allocation Study Final Report*. U.S. Federal Highway Administration, Department of Transportation, Washington, D.C. http://www.fhwa.dot.gov/policy/otps/costallocation.htm.

U.S. FTA. 2003. *National Transit Databases, 2003 Database*. Federal Transit Administration, Department of Transportation, Washington. www.ntdprogram.com.

U.S. OMB. 1992. *Guidelines and Discount Rates for Benefit-Cost Analysis of Federal Programs*, Circular No. A-94, Revised, Section 8. U.S. Office of Management and Budget, Washington, October.