THE DYNAMICS OF FARE AND FREQUENCY CHOICE IN URBAN TRANSIT

IAN SAVAGE
Department of Economics and the Transportation Center
Northwestern University

Correspondence Address
Ian Savage
Department of Economics
Northwestern University
2001 Sheridan Road
Evanston, Illinois 60208, USA
Ph: +1-847-491-8241
Fax: +1-847-491-7001
E-mail: ipsavage@northwestern.edu

Original Version: December 2007
This Revision: April 2008

JEL Classification: L92, R41

Abstract

This paper investigates the choice of fare and service frequency by urban mass transit agencies. A more frequent service is costly to provide but is valued by riders due to reduced waiting times at stops, and faster operating speeds on less crowding vehicles. Empirical analyses in the 1980s found that service frequencies were too high in most of the cities studied. For a given budget constraint, social welfare could be improved by reducing service frequencies and using the money saved to lower fares. The cross-sectional nature of these analyses meant that researchers were unable to address the question of when and why the oversupply occurred. This paper seeks to answer that question by conducting a time series analysis of the bus operations of the Chicago Transit Authority from 1953 to 2005. The paper finds that it has always been the case that too much service frequency was provided at too high a fare. The imbalance between fares and service frequency became larger in the 1970s when the introduction of operating subsidies coincided with an increase in the unit cost of service provision.

Keywords: transit, fares, frequency, optimality, subsidy, Chicago

INTRODUCTION

Economists have long recognized that for a given budget constraint, urban mass transportation firms can choose both the price ("fare") charged and the frequency at which the buses run ("service level"). Taking service level to be a measure of product quality, this transportation problem can be generalized to one that faces firms in many industries in both the public and private sectors (Crawford and Shum, 2006). Firms have to decide on the quality of their product, given that quality is valued by the customer but costly to provide.

The solution to this optimization problem is relatively straightforward in urban transit because there is usually monopoly provision.[1] Prompted by the widespread introduction of subsidies in the early 1970s, theoretically models were developed that define the combination of fare and service level that maximizes social welfare for an exogenously-determined budget constraint (Nash, 1978; Glaister and Collings, 1978).

An empirical literature in the early 1980s investigated whether actual fares and service levels deviated from the theoretically optimal combination. Glaister (1987), using data from 1982, found that in four of five major British cities there was too much service provided at too high a fare. Dodgson (1987) conducted a similar, but somewhat simplified, analysis for Australian cities using 1982/3 data and unambiguously found that service was overprovided. Finally, an analysis of Chicago in 1994 by Savage and Schupp (1997) found results that were in line with the situation in Australia.

The traditional folklore explanation for this phenomenon is that transit managers have a preference for attempting to maintain service output and employment in a declining market. Nash (1978) formalized this objective as "bus mile maximization." The implied reasoning for the mangers' preference is twofold. First, because transit is heavily unionized, managers have shied away from the disutility of negotiating job cuts. Second, because they partially rely on taxpayer funding, transit agencies feel obligated to provide service to all neighborhoods, including those that generate limited levels of demand. The studies cited above cannot provide any insights into the validity of these folklore explanations because they are all one-time snapshots of the situation in the various cities.

A more recent paper by Glaister (2001) re-estimated his models using data from the late 1990s. In the interim, the British bus industry went through privatization and, outside London, deregulation. In contrast to the previous findings, he now found that transit riders would prefer the provision of more service at a higher fare. Moreover, in several cities it would even make commercial sense to provide more service. Glaister does not explain the conundrum of why some transit companies are missing out on the profitable provision of additional service. One explanation for the reversal in findings is that deregulation has led to a substantial decline in unit costs, which presumably has made expansion of service less costly and hence relatively more attractive.

The British bus industry went through a revolutionary change in the late 1980s and 1990s. In most other parts of the world, there has been a more evolutionary change. This paper looks at a long time series for the bus services provided by the Chicago Transit Authority (CTA) between 1953 and 2005. The purpose is to estimate a simplified version of Savage and Schupp's model for each year, using data collected as part of another research project which looked at the historical increase in transit subsidies (Savage, 2004). The paper calculates the social welfare maximizing combination of fares and service levels in each year, and illustrates the year-to-year dynamics resulting from changes in the demand and cost functions, and political decisions on the amount of subsidy available. The paper also calculates the divergence between the optimal and actual combinations of fares and service levels, and explores how and why the magnitude of the divergence has changed over time.

THEORETICAL MODEL

For the sake of space, this paper will not repeat all of the underlying theoretical and methodological considerations that are fully covered in the theoretical literature of the 1970s and the later empirical papers. On the demand side, it suffices to say that the usual formulation is adopted whereby the generalized cost to the rider is a combination of out-of-pocket costs (the fare) and the valuation of the time taken for the trip. The time taken comprises the access and egress time associated with walking to and from the bus stop, the time taken waiting at the stop, and the in-vehicle travel time. Transit access/egress and wait times are inversely related to the density of routes and the frequency of service. More service provision means that routes will be located closer to the traveler's origin and destination and, assuming that there is some randomness in when the traveler arrives at the stop, she will have a shorter wait for the bus to arrive. In addition, in-transit time is also inversely related to service levels. Because demand has been found to be inelastic with regard to service levels, increased service levels will reduce the average number of people on each bus. Consequently, if service levels are increased, the average trip will be quicker because the bus will stop less often, and for a shorter duration, to allow fellow passengers to board and alight.

Demand, measured as the number of annual passenger trips (Q), will be expressed as a function, d(.), of the average fare paid (P), vehicle miles (VM) as a proxy for service level, and a set of exogenous demand variables (X) representing the wide variety of societal changes that have, in general, reduced transit demand over the years. These exogenous factors include the end of the six-day workweek, the rise of home-based entertainment (television), the rise of automobile ownership, the movement of population from traditional cities to the suburbs, and the outward migration of workplaces. Of course, over the long run, transit policy may have influenced some of these social changes such as the choice of residential and workplace location.

Traditionally, costs have been modeled as a function of the number of vehicle miles and/or vehicle hours operated, and the number of vehicles required for peak-period service. A large and varied literature over many decades suggests that this is an industry that displays constant returns to scale when output is measured using supply-side metrics.

Algebraically, the total cost function will be defined as a function, c(.), of vehicle miles, an exogenously determined vector of factor prices (Y), and an efficiency measure (Z). (The analysis in this paper will vary vehicle miles and peak vehicle requirement in proportion to each other, so the stylized cost function will just be expressed in terms of vehicle miles.) The efficiency measure encompasses the relative efficiency with which the transit agency combines together inputs, technological changes, and the possibility that the agency may pay inputs a price higher or lower than that which prevails in the market. An example of the latter is paying drivers and mechanics premium wages and/or benefits.

It is analytically important to recognize that while there may be constant returns to scale, there is an *increasing* marginal cost to providing higher service quality to the rider. For the proof of this consider a route on which a bus can make a round trip in one hour at a cost of $100, and passengers arrive randomly at stops. To provide a 20-minute frequency, the transit agency must deploy three buses at a cost of $300 an hour. Passengers wait on average for 10 minutes for a bus to arrive. To double the frequency to every 10 minutes requires three additional buses. The average waiting time is now only five minutes. The reduction in waiting time of five minutes has been achieved at a marginal cost of $300, or $60 for each minute of average waiting time saved. To further increase the frequency to every five minutes requires six additional buses. The average waiting time falls by two-and-a-half minutes at a marginal cost of $600, or $240 for each minute of average waiting time saved.

Denoting the politically determined level of subsidy as B, an agency with a requirement to break-even after subsidies faces a budget constraint given by:

(1) $P^*d(P, VM, X) + B = c(VM, Y, Z)$

The first term is the revenue collected from passengers, and is known as farebox revenue. This equation (less the Z variable) is the starting point for Nash's (1978) analysis. He points out that there are multiple combinations of the endogenously determined variables P and VM that satisfy equation (1). Moreover, as the equation contains squared (and perhaps even higher power) terms in both price and vehicle miles,[2] the combinations can be thought of as forming the closed boundary of a shape similar to that illustrated in Figure 1.[3]

The socially optimal combination is determined by overlaying the budget constraint with the contours of a social welfare "hill." Transit riders unambiguously prefer more service and lower fares, so welfare is increasing toward the southeast in Figure 1. There will be a tangency point, denoted by A, where welfare is maximized. When fares and service levels are at their optimal combination, denoted by P* and VM*, the literature describes them as "balanced." The empirical evidence suggests that, in actuality, transit agencies have chosen to locate to the northeast of the balance point (for example at point E).

The inclusion of an efficiency variable (Z) introduces a new element to the traditional model. Presumably this variable is largely endogenous, albeit that there may be exogenous influences on a publicly owned transit agency. For example, there is evidence that in the 1970s in Chicago that there was political pressure to improve fringe benefits and introduce more lax scheduling arrangements to avoid the types of crippling strike action that bedeviled other parts of the public sector such as the schools. The model in this paper assumes that the transit agency is playing a two-stage game. In the first stage the level of tolerated (in)efficiency is chosen, and then fares and frequencies are selected in the second stage. The focus of this paper is the second stage. To the author's knowledge there has not been any theoretical literature investigating the first stage of this game, but there has been a literature discussing the empirical linkages between the availability of subsidy and cost efficiency.

DYNAMICS IN THE THEORETICAL MODEL

How changes in the various exogenous variables affect the balance point can be illustrated with the help of Figure 1. If subsidies (B) are increased, the relevant portion of the budget constraint moves downward and to the right. For a given level of vehicle miles the agency can afford to reduce the fares (as transit is generally regarded as price inelastic), or for a given level of fares the agency can provide more vehicle miles. The balance point will also move down and to the right indicating that increased subsidies should lead to lower optimal fares and increased vehicle miles.

If exogenous factors reduce demand (X), the budget constraint will move upward and to the left. That is to say, in the reverse direction of that associated with an increase in subsidy. Exogenous decreases in demand will lead to a higher balanced level of fares, and a decrease in vehicle miles.

An increase in exogenous factor prices (Y) will make the relevant part of the budget constraint steeper, which is to say it pivots upward relative to the origin. One can unambiguously conclude that the balanced level of vehicle miles will decrease, but is unclear whether the balanced level of fares will increase or decrease. This will depend on the shape of the budget constraint and the contours of the welfare hill. A similar outcome will occur if the (predetermined) level of efficiency (Z) becomes worse.

THE APPLICATION TO CHICAGO

As discussed in Savage (2004), the CTA presents a unique opportunity for time-series analysis. It was taken into public ownership in 1947. While regional mechanisms for transportation financing emerged in the 1970s, the CTA continued to have its own corporate governance, and responsibility for service planning and pricing. It has also remained untouched by the global trend to privatization and contracting, and directly operates its own regular route bus and rail services. In 2005 it provided service with a peak requirement of 1,700 buses, and 1,000 railcars on seven "elevated" rail routes, in the City of Chicago and the older inner suburbs.

While Savage and Schupp analyzed both the bus and elevated rail services, this paper is solely concerned with the bus system. This is because changes in service output of the rail system have primarily been associated with new line construction (in the late 1960s, early 1980s, and mid-1990s). In contrast, changes in the bus system have been more gradual and reflective of social and land use changes. The paper uses the shorthand term "bus" to mean the service on the surface city streets. This service has been provided by a combination of streetcars (which were eliminated by 1958), trolleybuses (which replaced streetcars on many routes but were themselves eliminated by 1973) and motor buses. The analysis starts in 1953 following the acquisition by the CTA of the Chicago Motor Coach Company in October 1952.

An unusual aspect of the situation in Chicago means that the number of vehicle miles operated is a good proxy for service frequency. The CTA did not expand its service area to cover the new suburbs that emerged after the Second World War. These suburbs were served by private and municipal systems that ultimately became part of a separate publicly-owned suburban bus company in the 1970s and 1980s.[4] In addition, Chicago is built on a grid system of streets, with the major streets a standard distance apart. Because service has always been based on these major streets, there is an unusually strong relationship between vehicle miles and the service frequency received by riders.[5]

THE TRANSIT MARKET IN CHICAGO

An attractive feature of undertaking a time-series analysis of the dynamics of the transit market in Chicago is that there has been considerable variation in the endogenous and exogenous variables over the past fifty-three years.

The trends in the two main endogenous variables are shown in Figure 2. Vehicle miles, which are plotted on the right-hand axis, have declined almost continually, and are now 45% below their 1953 value. There were small service increases in the mid-1960s and after 1999. In contrast, real average fare (calculated as farebox revenue divided by ridership, and expressed in 2005 dollars using the Consumer Price Index), that is plotted on the left-hand axis, has varied considerably. It increased in the 1950s and 1960s, and reached a record high level in the late 1960s. Real fares fell considerably in the 1970s as the nominal fare was held almost constant during an era of high inflation. Real fares started to rise again in the 1980s, but there was another ten-year freeze in nominal fares between 1993 and 2003.

Figure 3 shows a graph of ridership on the surface system. Ridership is measured as annual unlinked trips (which means that a journey that requires a transfer to another bus or to the elevated rail system is counted as two trips). Actual ridership, shown as the solid line, fell by two-thirds between 1953 and 2005. Of course, the rise in real fares and the decline in service levels have been partly to blame. A counterfactual estimate of demand based on the holding fares and vehicle miles at their 1953 values can be found by applying elasticities calculated in Savage (2004). This is shown as the dashed line. The continual downward slope of this dashed line represents the relentless erosion of demand due to the exogenous conditions that were described previously. Albeit that there has been some leveling off in the downward trend in recent years due to a modest repopulating and gentrification of the inner city.

Figure 4 shows trends in real unit costs, defined as the surface system operating cost plus the annualized purchase cost of vehicles divided by vehicle miles and expressed in 2005 prices. Real unit costs have increased by 120% from \$4.82 a mile in 1953 to \$10.84 a mile in 2005. Part of this increase is due to exogenous factor price pressures. Adopting the same methodology as Savage (2004), the exogenous effects can be estimated by constructing a cost index which is 90% based of real wages in the national economy, and 10% on real changes in the fuel and power component of the national producer price index. Applying this index to the 1953 unit costs produces the dashed line. This indicates that real unit costs would have increased by more than 80% simply due to exogenous market conditions. The gray line with the triangle markers measures the difference between the two lines, and represents endogenous efficiency decisions. Prior to 1966 the CTA was achieving efficiency gains. However, inefficiencies were introduced between 1967 and 1980, which coincidentally was the period of the rapid rise in subsidies. There were some efficiency gains immediately after 1980, and this was followed by twenty years of approximately constant efficiency. Inefficiencies appear to have increased since 2002. It is worth noting that if the level of efficiency achieved in 1966 had been sustained to the present, unit costs would be 35% lower.

Finally, Figure 5 shows trends in surface system real farebox revenue and real total cost. Total costs were constant until the mid-1960s, increased between 1965 and 1980, declined in the following twenty years, and have increased considerably since 2000. Revenue was also reasonably constant until the mid-1960s and has declined almost continuously since. The difference between the two lines is the operating profit or loss. Prior to 1965 the CTA was making a small operating profit (of less than 6% of revenue). It had to do so as it was required to pay principal and interest on bonds that were issued in the 1940s and early 1950s to allow the CTA to purchase the assets of its predecessor companies. Between 1965 and 1980, subsidies increased rapidly. (The operating loss will be indicative of subsidy to the bus system, as the CTA is required to present an overall balanced budget each year, after subsidies, when combining together its bus, elevated rail, and subcontracted paratransit service to the elderly and physically challenged.)

Following a regional transit-funding crisis in 1980, a new funding system emerged in 1983. The amount of subsidy received by the CTA became truly exogenous as it was tied to a specific local sales tax levy combined with a set ratio of matching revenues from State government. However, for a period in the early to mid-1990s and in the years between 2002 and 2007 these specific subsidies were insufficient to balance the budget. One-off grants, the reassigning of capital grants for operating purposes, and underfunding of the necessary contributions to the employees' pension and retirees' healthcare funds covered the shortfall. In early 2008, after a protracted political battle, the State legislature voted to increase the sales tax levy.

METHODOLOGY AND DATA

This paper estimates a simplified version of Savage and Schupp's model for each year from 1953 to 2005. There are two principal simplifications. First, the paper only considers the bus system, whereas Savage and Schupp analyzed both the bus and rail systems. Second, the paper does not disaggregate by time of day and day of week. Much of the necessary historical data was collected as part of a separate study of transit subsidies (Savage, 2004). While this paper builds off of the detailed calculations for 1994 made by Savage and Schupp, all financial data have been expressed in 2005 prices. For the sake of space, the following paragraphs will not repeat all of the discussion of data sources and data manipulations that were in the two earlier papers, and will just highlight the main points, and any specific calculations that were necessary for this paper.

Demand Data

As already discussed, ridership is measured as annual unlinked passenger trips, and average fare is measured as total bus farebox revenue per unlinked trip. Information on farebox revenue disaggregated into bus and rail modes has only been available since 2002. Since then the average revenue per trip on the bus system averages 90% of the average for the bus and rail system combined. The analysis assumes that this ratio holds for all years. While the CTA has a flat fare system that does not differentiate by mode or distance traveled, the bus system may have a higher proportion of riders such as school children or seniors who qualify for discounted fares.

In common with the 1994 analysis, a nonlinear waiting time function is used that is derived from a classic paper by Seddon and Day (1974). The function relates average waiting time to the scheduled average interval between buses, known as headway. In 1994 the average headway was 9.798 minutes when averaged across all routes and time periods (excluding the "owl" overnight hours). The estimated headway in other years is calculated by multiplying the 1994 headway by the ratio of the vehicle miles in a given year to the vehicle miles in 1994. There is an implicit assumption that the change in vehicle miles over the years has been manifested in higher or lower frequencies rather than in a change in the density of routes. As already discussed, the grid pattern of streets in Chicago makes this more likely than in many other cities.

In-transit time is endogenous. While it forms part of generalized cost which determines ridership, the level of ridership affects in-transit time as a more crowded bus has to stop more often and has longer dwell times to allow passengers to board and alight (see equations 13 and 14 in Savage and Schupp). In 1994 an average bus passenger's trip was 2.37 miles long, and he or she would be delayed as a result of 10.81 other passengers boarded the vehicle. The generally accepted average boarding times for the type of vehicles used by the CTA is 2.5 seconds (0.042 minutes). Therefore, based on known average speed, in-transit time (T) in minutes is given by the equation:

$$(2) \qquad T = 13.212 + 0.042\,((2.37 * \text{Passenger trips})/\text{Vehicle Miles})$$

If there were no passengers on board, the bus would take 13.212 minutes to travel 2.37 miles. In specifying the demand function, some manipulation and collecting of terms is necessary to take account of the endogenous nature of in-transit time.

A standard approach has been used to valuing time. In-transit time is valued at half the average wage rate, and waiting time at twice this amount. Real wages have increased over the past fifty years and the real values of time have been similarly adjusted using data on real wages.

Of course, it is rather heroic to assume that the socioeconomic characteristics of riders have remained constant for fifty years and hence that the value of time of transit riders has remained a constant proportion of the real average wage rate in the economy. In defense of this assumption, it should be pointed out that even today the CTA has a very diverse ridership, with plenty of middle and upper income riders especially on the busy services along the lakefront. One would imagine that those riders who abandoned transit for the automobile had the highest value of time and were the least price sensitive. This would imply that transit riders today would prefer less frequent service at lower fares compared with their counterparts in the 1950s.

The demand function in any given year is assumed to be locally linear around the actual observed generalized cost and demand. This is a less egregious assumption that it might seem because the optimization will generally change demand by less than 5%. The point generalized cost elasticity at the actual observed level of ridership is obtained by transforming a known price elasticity (see equation 12 in Savage and Schupp). The same price elasticity will be assumed at the point of actual observed ridership in every year. This elasticity is -0.457, which was estimated by Savage (2004) for the Chicago bus system in a time-series (1948-1997) analysis that also included vehicle miles as an explanatory variable.

Bus Operating Costs

Costs are the combination of operating costs and the capital costs of vehicles. Operating cost data disaggregated into bus and rail components are available since 1982. An appendix to Savage (2004) discusses how total CTA operating costs can be disaggregated by mode for earlier years. The annualized cost of vehicles, which would normally appear in the capital budget, is also included, as this will vary as service levels are optimized. The purchase costs are annualized over a 12-year life, and are assumed (as was the case in 1994) to be equivalent to 7.31% of operating costs.

In the 1994 analysis, costs were divided by line item into (1) costs that vary with the number of vehicle hours or miles operated, (2) cost that vary with the number of vehicles owned, and (3) costs that are invariant with the level of service provided (See table 4 in Savage and Schupp). For the bus system, about 16% of costs fall into the third category. Because this analysis does not disaggregate by time of day or day of week, any changes in vehicle miles will also require a proportionate change in the vehicle requirement. Consequently, items (1) and (2) can be amalgamated and expressed in terms of an average unit cost per vehicle mile. This will also be taken to be the marginal cost of a vehicle mile.

Transit Finances

The estimate of B in equation (1) is total bus system cost, as defined in the previous paragraph, less total bus farebox revenue. Trends in this variable were illustrated in Figure 5.

Road Congestion

The 1994 analysis also attempted to quantify the benefits of reduced congestion for road users resulting from improvements in transit services. Benefits were assumed to only accrue in peak periods when congestion is the most severe. Reductions in road traffic were associated with the mode switching of a subset of the new transit riders who were formerly auto drivers or taxi passengers. The calculations in 1994 were problematic. There was comparatively little information of the previous mode choice of new transit riders. Moreover, there was little to no quantitative information available on the level of congestion actually experienced on the city's arterial and local streets. Given the rather heroic assumptions and calculations required in the 1994 analysis, this paper does not attempt to extrapolate these benefits to other years.

The 1994 analysis did have some startling findings. In the weekday peak period, reducing fares on the buses was found to generate additional congestion reduction benefits equivalent to 9% of the benefits accruing directly to bus riders in the peak. However, improving bus service by operating more vehicle miles actually had a negative net effect on traffic congestion! A marginal bus mile was estimated to remove just 1.24 auto-miles from the road. Unfortunately, in standard traffic flow models, a bus that stops in the roadway rather than in bus bays is counted as the equivalent of 4.37 cars. Overall, a third of the benefits to bus riders in the weekday peak periods from increased frequencies were offset by the worse road congestion that the additional buses caused.

Therefore, in interpreting the findings of this paper, the reader should remember that incorporating road congestion would reinforce the argument that society would be better off if less transit service was provided at a lower fare.

Other Benefits of Transit

Neither the 1994 analysis nor this paper quantifies the wider benefits of transit. It is frequently argued that transit serves a social role by providing the ability for persons of modest means and/or those who live in socially-segregated areas to access jobs and hence share in the economic vitality of the city (O'Regan and Quigley, 1999). Transit can also affect land use patterns, and bring about agglomeration economies. Chicago is an example of what Thomson (1977) termed a "weak centered city." These cities have to struggle to maintain a downtown, and prevent all economic activity from moving to the suburbs. A subsidized radial transit system is part of the cost of supporting an active and viable downtown.

Compared with out-of-pocket costs and travel time savings, it is difficult to assign monetary values to social and land use benefits. However, it is reasonable to assume that the magnitude of these latter benefits will monotonically increase with transit system ridership. Therefore balancing fares and service levels so that demand and rider benefits are maximized should also be associated with maximizing these other benefits within a given budget constraint.

Optimization Process

The optimization process was conducted in Excel, and involved moving along the budget constraint in each year to find the combination of fare and vehicle miles that maximized transit rider social welfare.[6] Computational simplification was possible because the analysis treats riders as a single market, and does not disaggregate by rider type or by time of day or day of week. Consequently, the constrained combination of fare and vehicle miles that maximizes rider welfare will occur at the point where ridership is maximized.

FINDINGS

The actual and the balanced levels of fares and service levels in each year are shown in Table 1. The table also shows the actual ridership and the estimated ridership at the balance point. To illustrate the results, Figure 6 plots the trajectory of the fare/vehicle mile combinations over time. Fares are plotted on the vertical axis and vehicle miles on the horizontal axis. The actual combination is shown as the thin line with the circles, and the balanced combination as the thick line with the squares. One can think of the thick line as representing how point A in Figure 1 has moved over the years.

Because Figure 6 can be rather confusing, the ratio of balanced to actual values for fares, vehicle miles and ridership for each year is shown in Figure 7. A ratio below unity indicates that the balanced value for the variable is lower than the actually observed value. It is immediately clear from Figure 7 that even in 1953 there was "too much service at too high a price." Therefore one can conclude that the oversupply of service found by researchers in the 1980s and 1990s was, at least in the case of Chicago, not a recent phenomenon.

In analyzing what happened over the following years, it is probably beneficial to look at four eras, rather trying to analyze changes from year to year. The four eras are 1953-1965, 1966-1980, 1981-1999, and 2000 to the present.

The Commercial Era (1953-1965)

The era was marked by a substantial exogenous decrease in demand, and exogenous pressures on costs as real wages increased in the economy in general. Because the CTA was constrained to make a small operating profit to service its bonds, the exogenous factor price increases were counteracted by efficiency gains. The decline in demand moved the budget constraint upward and to the left. The modeling predicts that the balance point in 1965 has 28% higher fares and 9% lower vehicle miles than the balance point in 1953. The actual changes in these variables were actually quite similar (fares increased by 30% and vehicle miles were cut by 9%). However, because there was already too much service at too high a price in 1953, the imbalance was perpetuated throughout the period. Inspection of the relevant part of Figure 7 shows that the ratio of the balanced to actual values of fares and service levels was remarkably constant.

The Introduction of Subsidies (1966-1980)

The CTA faced substantial exogenous pressures on both demand and cost in the late 1960s and 1970s. There was a continued exogenous decrease in demand, exacerbated by the social turmoil of the time and the movement of population to the suburbs. Real factor prices increased due to exogenous Vietnam-era pressures in the labor market and the energy crises of the 1970s. In addition, the cost efficiencies achieved in the 1950s and 1960s were reversed. Subsidies were introduced, and grew rapidly during the 1970s.

In terms of the theoretical model, the increased subsidies move the budget constraint downward and to the right. The movement was partly offset by the exogenous declines in demand, which work in the opposite direction. The increased unit costs pivot the budget constraint upward relative to the origin. Theoretically, the increased subsidies should move the balance point so that more service and lower fares result, while the increase in unit costs is predicted to result in less service provision and an indeterminate change in fares. The empirical modeling finds that the negative effect due to increased unit costs overwhelmed the positive effect of increased subsidies on the balanced level of service. The balanced level of fares is predicted to fall by a substantial amount as a result of the massive influx of subsidies.

The balance point in 1980 has vehicle miles that are 27% below the balance point in 1965, and fares that are 48% lower. In actuality, vehicle miles fell by slightly less (25%) and fares declined by considerably less (31%). Therefore, while actual vehicle miles deviated somewhat more from the balance point in 1980 compared with 1965, the disparity for fares was very much more. This is because the increase in unit costs makes the budget constraint steeper. Providing additional vehicle miles above and beyond the balance point necessitates a much higher level of fares to pay for them.

A Constant Budget Constraint (1981-1999)

The rapid increase in subsidies in the 1970s eventually overwhelmed the taxing abilities of the regional funding authority. The funding system was reconstituted in 1983 and resulted in a consistent and dedicated stream of sales tax revenues that was truly exogenous in magnitude. Not surprisingly, this led to a period of cost containment. After some of the more egregious cost inefficiencies were eliminated in the early 1980s, unit costs were remarkably constant over the following fifteen years. The only major dynamic at work in the theoretical model in the 1980s and 1990s was the continued exogenous decrease in demand. The budget constraint moved upward and to the left. Consequently, the balance point in 1999 has 43% higher fares and 25% fewer vehicle miles compared with the balance point in 1980. In actually vehicle miles declined by somewhat more (27%) and fares increased by less (33%). As a result, the CTA was moving (marginally) back toward balance, although to get back to balance in 1999 would still have required service levels to be trimmed by a further 15% and the money used to reduce fares by an additional 30%.

Toward Doomsday (2000-2005)

Between 2004 and 2008 there was a lengthy political debate on transit funding that echoed the crises of the early 1970s and early 1980s. Ultimately, the sales tax levy was increased to provide more subsidies, but this only occurred after the CTA had repeatedly threatened "doomsday" service cuts and fare increases.

The current crisis had it origins in dramatic changes in the variables in the theoretical model in the years after 1999. On the positive side, the longstanding exogenous erosion of demand seemed to have lessened. However, unit costs increased starting in 2002 and, after two decades of stability, the deficit in bus operations climbed by 63% between 1999 and 2005. The larger deficit resulted from the increased unit costs, a fare freeze that had been in place since 1993, and an increase in vehicle miles that partially reversed the service cutbacks of the mid-1990s. As discussed earlier, the increased operating deficit required extraordinary funding sources to supplement the regular sources of subsidy.

In modeling terms, the situation was reminiscent of the 1970s. Theoretically, the balance point in 2005 should have had 62% lower fares and 5% fewer vehicle miles compared with the balance point in 1999. In actuality, vehicle miles increased by 9%, and fares fell by only 7%. The CTA was moving even further away from the balance point.

CONCLUSIONS

Our analysis finds that for the CTA there has "always" been an imbalance whereby too much service is provided at too high a price. This conclusion would be strengthened if the effects of transit on highway congestion were incorporated into the model. This is because evidence suggests that adding bus mileage exacerbates congestion rather than improving it.

Consequently, the findings of this paper do not answer the question of why the imbalance occurred in the first place, and whether there was ever a time in Chicago transit history when fares and service levels were in balance. The traditional arguments that transit managers are reluctant to shrink service in the face to declining demand were as true in 1953 as they were in 1973 or 1993. Ridership had peaked on the combined system of the CTA predecessor companies in 1927. While there was a resurgence during the motoring restrictions of the Second World War, the end of these restrictions and other social changes led to a 40% decline in ridership between 1946 and 1953 (Young, 1998, Table A4).

An undergraduate audience member at the Massachusetts Institute of Technology had an interesting alternative explanation for the findings in this paper. He posited that the local politicians, transit managers, and the "political elite" who are vocal in public policy are busy people with higher than average incomes. Consequently, they have lower than average price sensitive and a greater than average value of time. The observed fare and output decisions may be consistent with maximizing the welfare of these powerful groups when they ride the system, but inconsistent with the preferences of the vast majority of transit riders who would prefer less service at a lower price.

The paper finds that the magnitude of the disparity between actual and balanced values became larger in time periods characterized by unit costs increases. A counterfactual analysis can address the question of what would have happened to the disparity if cost inefficiencies had not set in after the mid-1960s. This question is relevant to proponents of privatization and competitive contracting of the type that has emerged in London and elsewhere around the world. The question was partial answered by Savage and Schupp who modeled the effects of 10%, 20% and 30% cost reductions in 1994. Their results, shown in Table 2, indicate the ratio of the benefits to the increase in subsidy necessary to decrease fares by 10% or increase service levels by 10%.[7] If fares and service levels are in balance, the benefit-cost ratio of a marginal increase in subsidy will be identical irrespective of whether fares were reduced or service was increased.

In interpreting this table it is worth bearing in mind that had the level of efficiency achieved by the CTA in 1966 been sustained to 1994 (subject to exogenous changes in factor prices), unit costs would have been 29.75% lower. The peak period is very unfavorable to service level expansions because additional service requires more vehicles, and the added road congestion reduces the benefits. In the weekday off-peak and on the weekends, cost reduction in the low 20% range would bring fares and service levels into balance. Therefore, privatization has dual benefits. The direct benefit is the removal of cost inefficiencies. The indirect benefit comes from moving transit service toward the balance point. This provides an explanation for why Glaister's findings for British cities in the late 1990s differed markedly from his findings in the same cities prior to deregulation and privatization.

Finally, despite the seemingly large disparity between actual and balanced fares and service levels, the consequent effect on the level of ridership, and by implication consumer welfare, is surprisingly small. Exclude the years of unusual financial difficulties (1994-5 and 2003-5), balancing increases demand by less than 5%. In 38 of the 53 years the increase is 3% or less.

ACKNOWLEDGMENT

I would like to thank David M. Levinson (University of Minnesota) for asking a question at a presentation of my 2004 paper that sparked my interest in undertaking this analysis.

ENDNOTES

1 There is a separate literature, spurred by deregulation in Britain in 1986, on frequency setting in a competitive environment. See, for example, Foster and Golay (1986).

2 The first term in equation (1) indicates that there will be (at least) a squared term in price. There will also be (at least) a squared term in vehicle miles because average waiting time, which

enters the demand function, is calculated by dividing a measure of time and space by vehicle miles. (Think of a route that is 5 miles long. In a given hour the frequency of service is 60 minutes multiplied by 5 miles divided by the number of vehicle miles operated on the route in that hour.)

[3] Only part of this boundary, the lowest fare consistent with a given level of vehicle miles, is relevant to the analysis. Therefore for the remainder of the paper, the phrase "budget constraint" should be taken to mean the segment that is to the south and to the east.

[4] The only major exception was the partial replacement service for a defunct private system in the adjacent suburb of Evanston in 1973. However, the amount of mileage was small, representing a fraction of a percent of CTA output.

[5] There is a remarkable resemblance between the current CTA route diagram, and the 1946 map produced by its predecessor the Chicago Rapid Transit Company (reproduced on page 17 of Ovenden (2007)).

[6] The welfare of taxpayers can be ignored as, by definition, the level of subsidy is held constant along the budget constraint.

[7] The shadow value of a dollar of tax revenues used to fund the additional subsidies was $1.26, so additional subsidies would only be justified if the ratio of benefits per dollar of subsidy exceeds 1.26.

REFERENCES

Crawford, Greg and Matthew Shum (2006). The welfare effects of endogenous quality choice: The case of cable television. Mimeo, University of Arizona.

Dodgson, John S. (1987). Benefits of changes in urban public transport subsidies in the major Australian cities. In Stephen Glaister (ed.), *Transport Subsidy*. Newbury, UK: Policy Journals, pp 52-62.

Foster, Christopher and Jeanne Golay (1986). Some curious old practices and their relevance to equilibrium in bus competition. *Journal of Transport Economics and Policy* 20(2): 191-216.

Glaister, Stephen (1987). Allocation of urban public transport subsidy. In Stephen Glaister (ed.), *Transport Subsidy*. Newbury, UK: Policy Journals, pp 27-39.

Glaister, Stephen (2001). The economic assessment of local transport subsidies in large cities. In Tony Grayling (ed.), *Any More Fares? Delivering Better Bus Services.* London: Institute for Public Policy Research, pp 55-76.

Glaister, Stephen and John J. Collings (1978). Maximization of passenger miles in theory and practice. *Journal of Transport Economics and Policy* 12(3): 304-321.

Nash, Christopher A. (1978). Management objectives, fares and service in bus transport. *Journal of Transport Economics and Policy* 12(1): 70-85.

O'Regan, Katherine M., and John M. Quigley (1999). Accessibility and economic opportunity. In José A. Gómez-Ibáñez, William B. Tye and Clifford Winston (eds.), *Essays in Transportation Economics and Policy: A Handbook in Honor of John R. Meyer*. Washington D.C.: Brookings Institution, pp. 437-466.

Ovenden, Mark (2007). *Transit Maps of the World*. London: Penguin.

Savage, Ian (2004). Management objectives and the causes of mass transit deficits. *Transportation Research A* 38(3): 181-199.

Savage, Ian and August Schupp (1997). Evaluating transit subsidies in Chicago. *Journal of Public Transportation* 1(2): 93-117.

Seddon, P.A., and M.P. Day (1974). Bus passenger waiting times in Greater Manchester. *Traffic Engineering and Control* 15(9): 442-445.

Thomson, J. Michael (1977). *Great Cities and their Traffic*. Harmondsworth, UK: Penguin.

Young, David M. (1998). *Chicago Transit: An Illustrated History*. DeKalb, Ill.:Northern Illinois University Press.

TABLE 1: Actual and balanced fares and service levels

	Average Fare			Vehicle Miles (000s)			Ridership (000s)		
Year	Actual	Balance	Ratio	Actual	Balance	Ratio	Actual	Balance	Ratio
1953	$0.73	$0.65	0.88	122,363	105,700	0.86	919,715	929,517	1.01
1954	$0.76	$0.67	0.87	120,937	103,300	0.85	847,895	858,221	1.01
1955	$0.79	$0.69	0.88	119,402	103,000	0.86	816,966	825,834	1.01
1956	$0.77	$0.70	0.90	118,244	104,600	0.88	808,998	815,216	1.01
1957	$0.83	$0.72	0.87	117,843	100,800	0.86	751,656	760,812	1.01
1958	$0.89	$0.76	0.85	113,617	94,600	0.83	681,963	693,239	1.02
1959	$0.89	$0.78	0.88	109,920	94,800	0.86	692,295	699,913	1.01
1960	$0.91	$0.80	0.88	109,546	94,000	0.86	674,931	682,861	1.01
1961	$0.95	$0.82	0.86	107,536	91,100	0.85	630,222	638,910	1.01
1962	$0.99	$0.86	0.87	106,190	90,800	0.86	625,718	633,310	1.01
1963	$0.98	$0.85	0.87	105,832	90,300	0.85	607,749	615,428	1.01
1964	$0.96	$0.84	0.87	108,584	92,700	0.85	602,540	610,340	1.01
1965	$0.95	$0.83	0.87	111,092	95,400	0.86	618,681	626,250	1.01
1966	$0.93	$0.82	0.89	112,273	98,500	0.88	649,524	655,609	1.01
1967	$0.92	$0.82	0.89	107,074	94,800	0.89	625,920	631,185	1.01
1968	$1.02	$0.88	0.87	103,792	89,400	0.86	564,019	571,022	1.01
1969	$1.21	$1.00	0.83	102,192	83,500	0.82	529,698	540,569	1.02
1970	$1.20	$0.98	0.82	98,314	80,400	0.82	503,342	514,164	1.02
1971	$1.22	$0.98	0.80	95,199	77,400	0.81	492,680	504,213	1.02
1972	$1.17	$0.89	0.76	95,154	75,800	0.80	488,936	503,986	1.03
1973	$1.09	$0.85	0.78	90,702	74,600	0.82	482,397	494,150	1.02
1974	$0.94	$0.73	0.78	88,178	74,400	0.84	511,351	522,299	1.02
1975	$0.84	$0.62	0.75	88,484	74,300	0.84	502,957	515,713	1.03
1976	$0.79	$0.61	0.77	87,468	75,200	0.86	523,876	534,311	1.02
1977	$0.79	$0.55	0.70	86,332	71,900	0.83	535,416	551,697	1.03
1978	$0.73	$0.52	0.72	83,815	71,100	0.85	545,875	560,229	1.03
1979	$0.67	$0.52	0.78	80,021	70,800	0.88	560,905	569,813	1.02
1980	$0.66	$0.43	0.66	83,383	70,000	0.84	537,693	555,659	1.03
1981	$0.80	$0.55	0.68	81,449	67,500	0.83	492,578	508,711	1.03
1982	$0.79	$0.64	0.81	75,884	67,400	0.89	467,110	473,317	1.01
1983	$0.76	$0.61	0.80	75,505	67,000	0.89	473,433	480,075	1.01
1984	$0.73	$0.64	0.87	72,277	66,900	0.93	482,237	485,113	1.01
1985	$0.71	$0.60	0.85	72,183	66,400	0.92	486,515	490,103	1.01
1986	$0.80	$0.69	0.87	72,326	66,400	0.92	466,396	469,644	1.01
1987	$0.81	$0.66	0.82	72,408	64,900	0.90	438,676	443,732	1.01
1988	$0.84	$0.66	0.79	74,154	64,900	0.88	430,089	437,155	1.02
1989	$0.80	$0.64	0.80	72,799	64,300	0.88	420,573	426,792	1.01

1990	$0.80	$0.63	0.78	72,522	63,600	0.88	421,184	428,256	1.02
1991	$0.80	$0.56	0.70	71,737	60,700	0.85	392,088	403,134	1.03
1992	$0.90	$0.55	0.62	70,803	57,100	0.81	370,335	386,370	1.04
1993	$0.97	$0.56	0.57	69,970	54,700	0.78	326,656	344,111	1.05
1994	$0.95	$0.47	0.49	72,686	55,100	0.76	331,521	354,610	1.07
1995	$0.93	$0.43	0.46	70,681	53,400	0.76	306,076	328,619	1.07
1996	$0.94	$0.56	0.60	67,073	53,700	0.80	302,115	316,051	1.05
1997	$0.94	$0.53	0.56	64,933	51,200	0.79	287,628	302,907	1.05
1998	$0.93	$0.59	0.63	60,889	49,900	0.82	290,531	301,501	1.04
1999	$0.88	$0.62	0.71	61,271	52,300	0.85	299,058	306,620	1.03
2000	$0.84	$0.57	0.68	61,869	52,400	0.85	302,090	310,785	1.03
2001	$0.82	$0.50	0.61	63,758	52,700	0.83	301,691	313,367	1.04
2002	$0.79	$0.41	0.52	65,901	52,900	0.80	303,295	319,608	1.05
2003	$0.83	$0.32	0.39	66,378	50,500	0.76	291,804	315,441	1.08
2004	$0.84	$0.27	0.32	66,572	49,500	0.74	294,031	321,709	1.09
2005	$0.82	$0.24	0.29	66,812	49,500	0.74	303,244	333,397	1.10

TABLE 2: Benefit-Cost Ratio from Increasing Subsidies to Improve Service Levels by 10% or Reduce Fares by 10% in 1994

	Monday-Friday		Saturday	Sunday
	Peak	Off Peak		
Fares decreased by 10%	1.39	1.77	1.77	1.80
Vehicle Miles increased by 10%	0.21	1.11	1.24	1.16
- with a 10% unit cost reduction	0.23	1.32	1.47	1.37
- with a 20% unit cost reduction	0.27	1.61	1.82	1.68
- with a 30% unit cost reduction	0.31	2.07	2.39	2.17

Source: Savage and Schupp (1997), Tables 6 and 7.

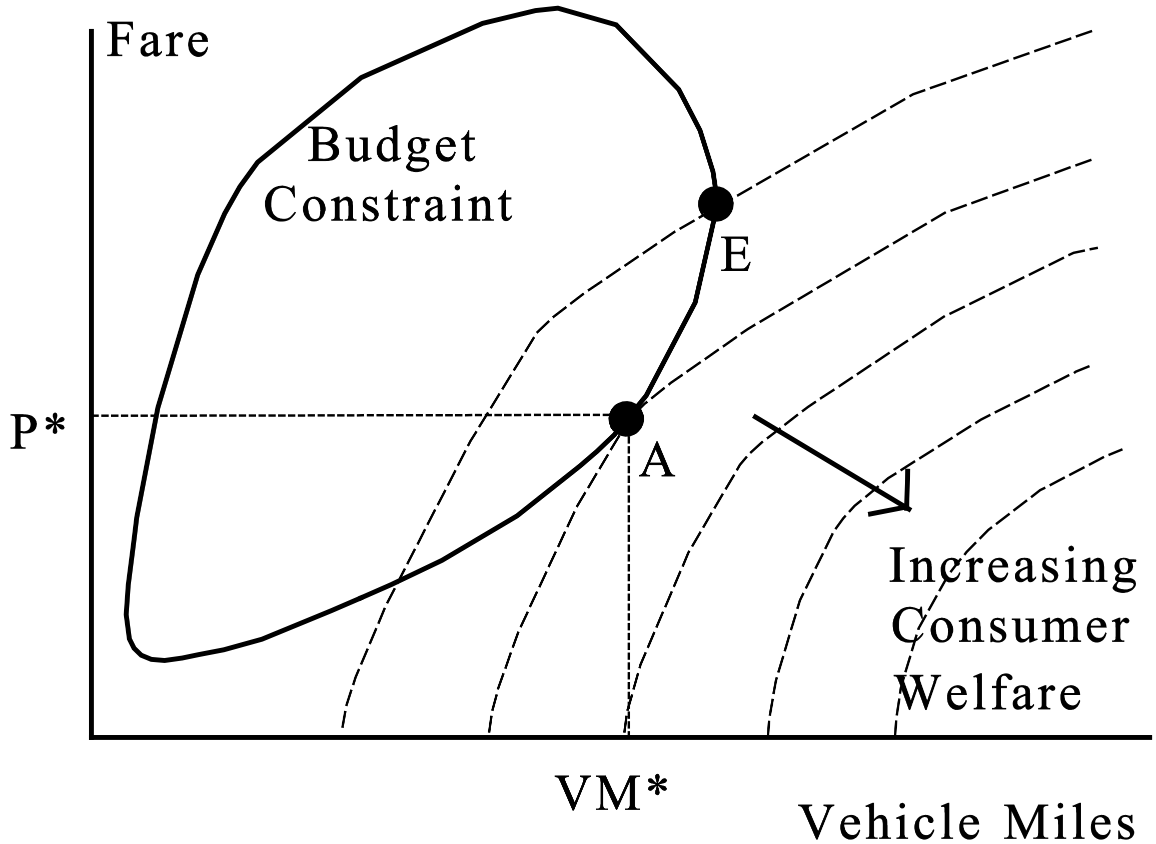

FIGURE 1: Balancing Fares and Vehicle Miles for a Given Budget Constraint

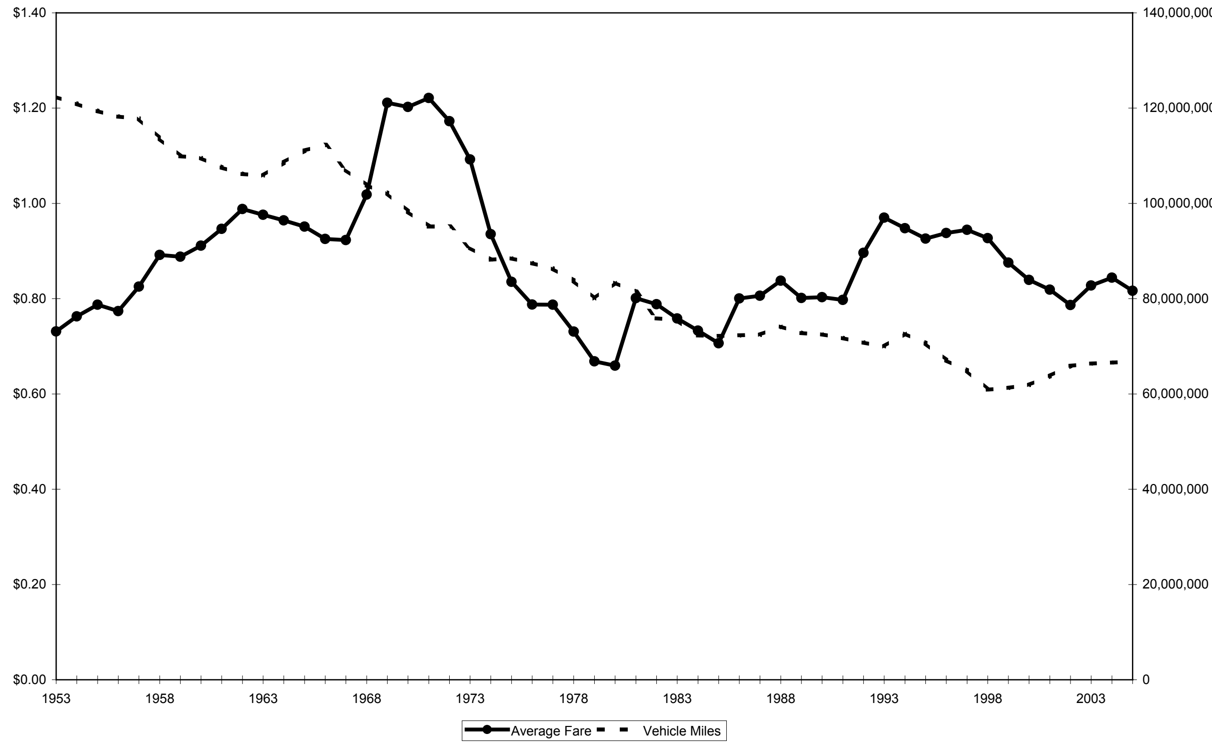

FIGURE 2: Trends in Real Average Fare (left axis) and Vehicle Miles (right axis)

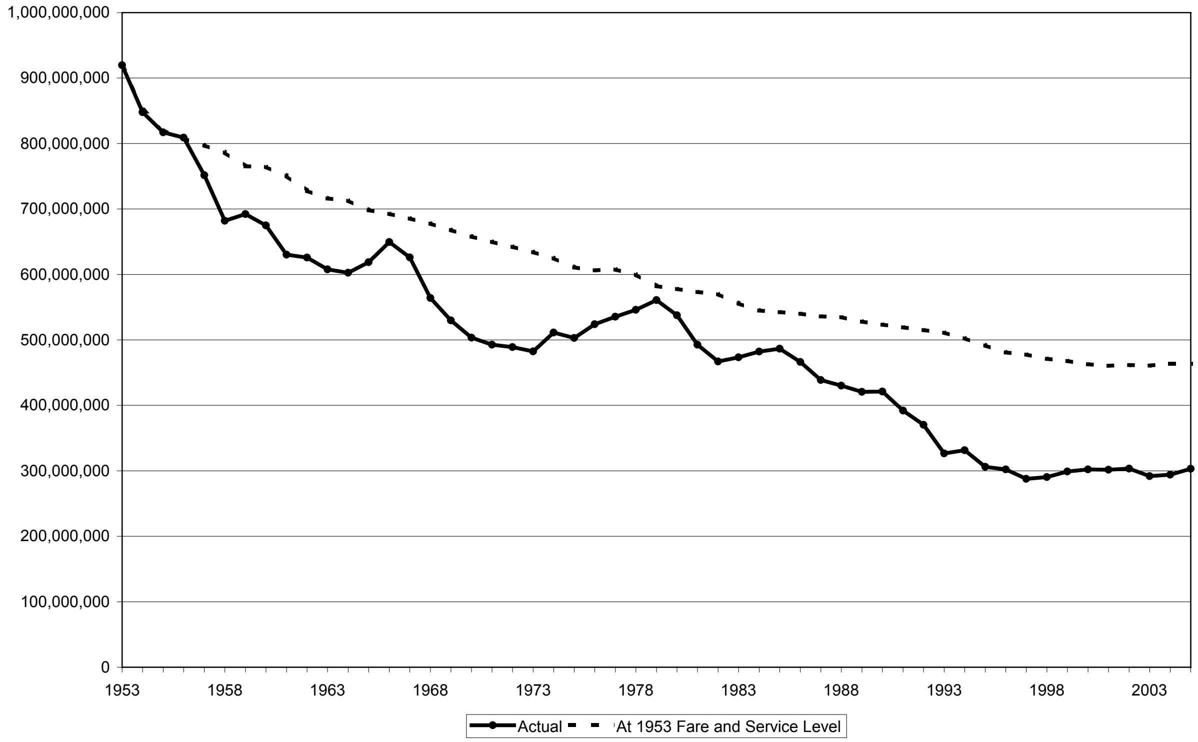

FIGURE 3: Trends in Surface System Unlinked Passenger Trips

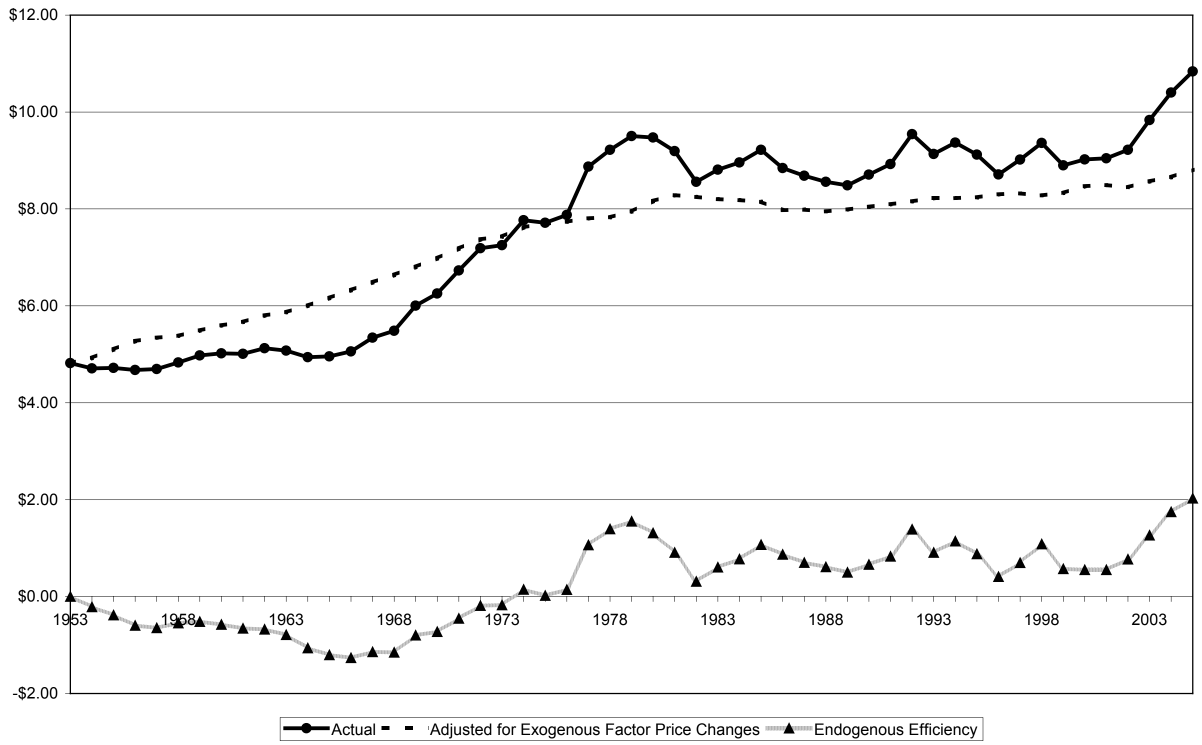

FIGURE 4: Analysis of Real Unit Costs

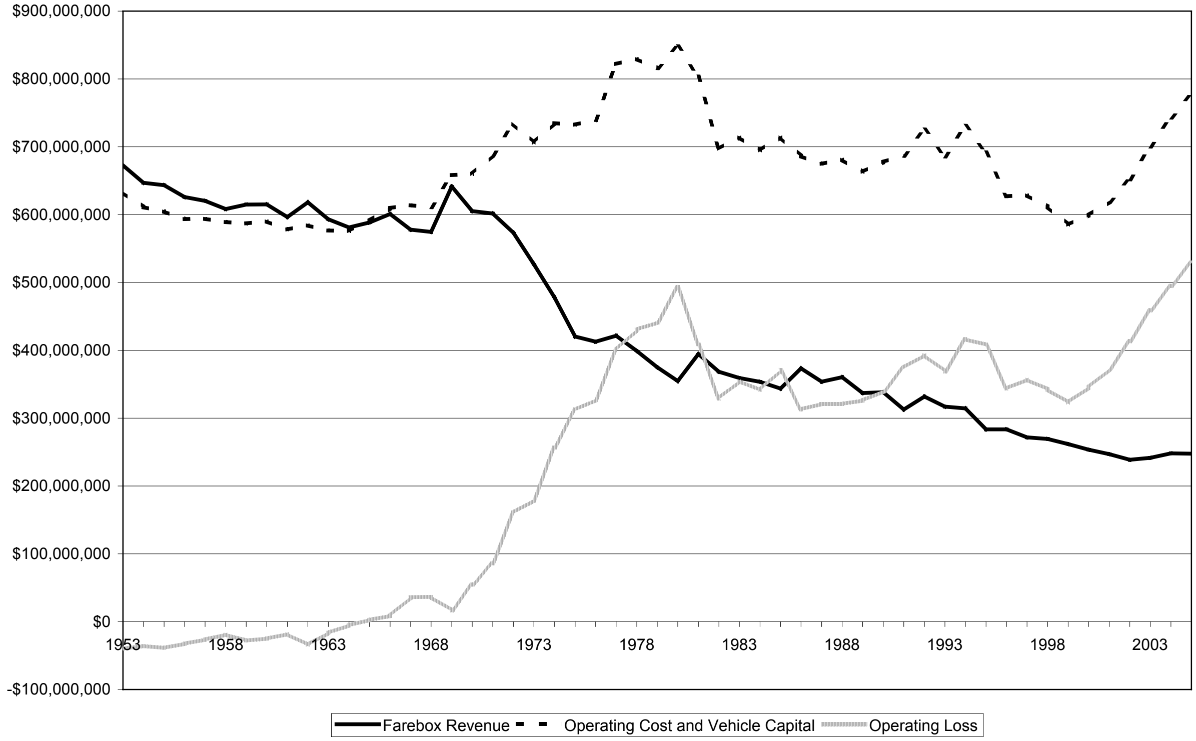

FIGURE 5: Surface System Finances in 2005 Prices

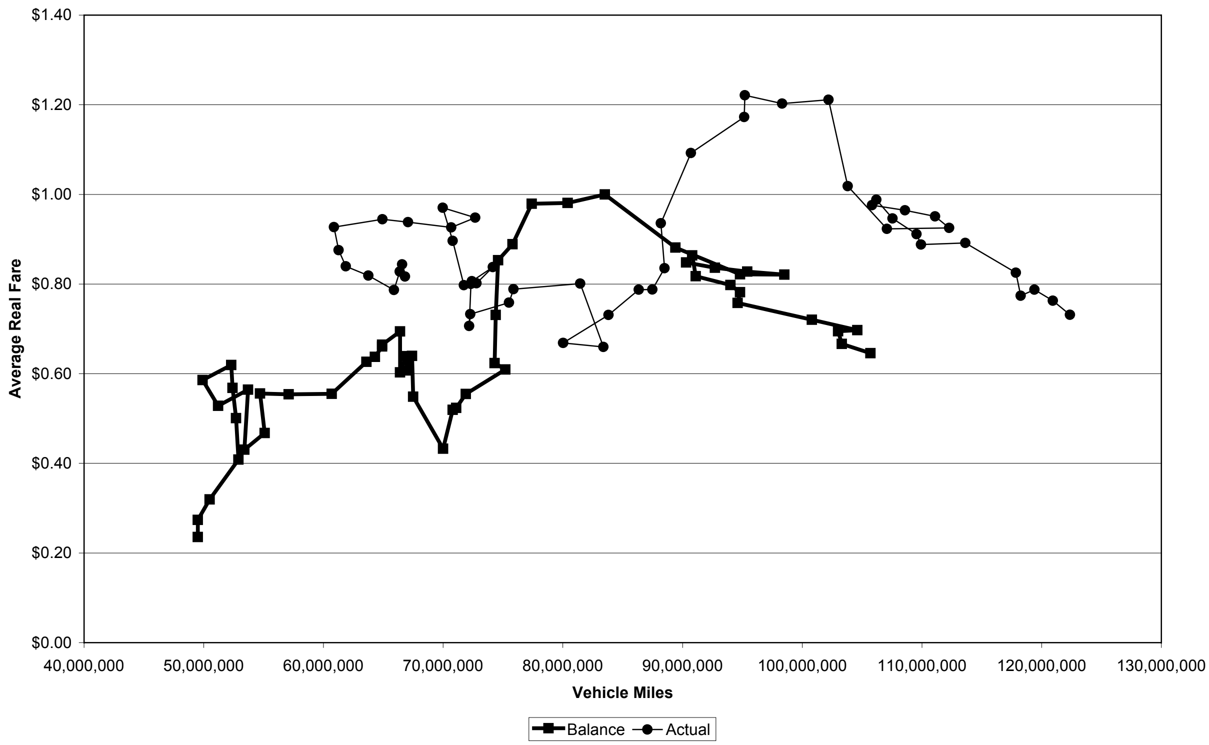

FIGURE 6: Balanced and Actual Combinations of Fare and Vehicle Miles 1953-2005

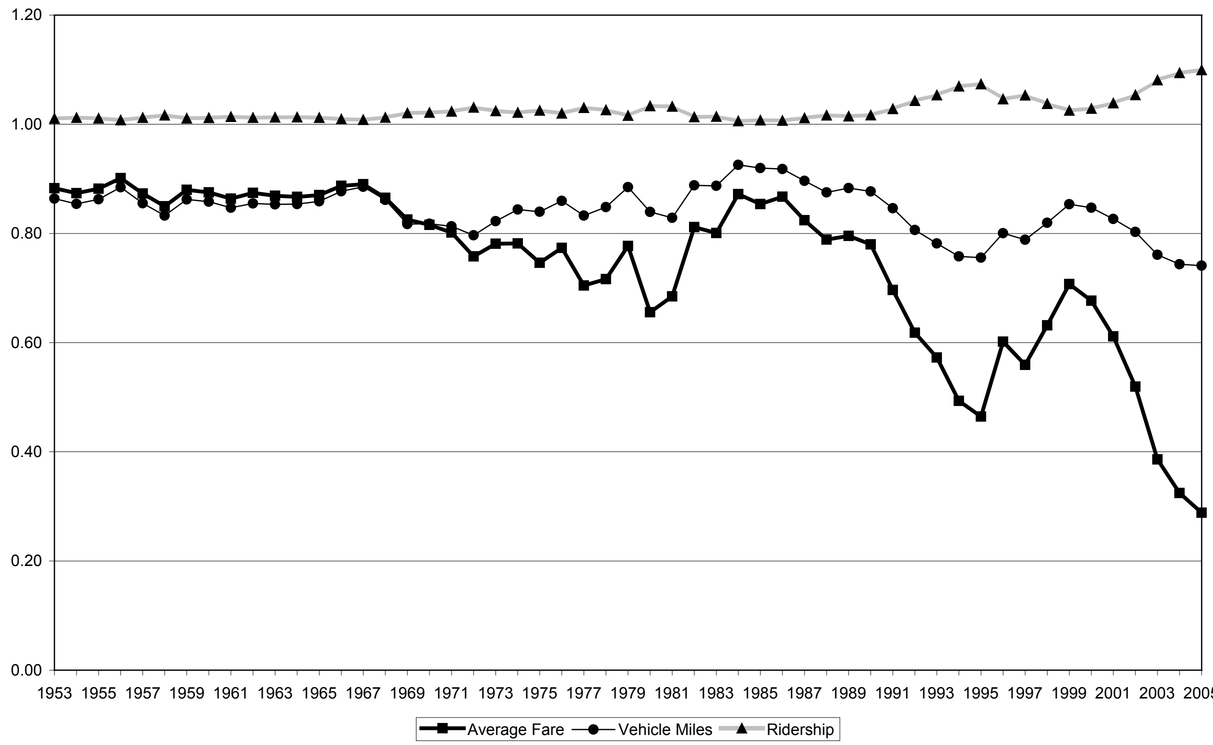

FIGURE 7: Ratio of Balanced to Actual Average Fare, Vehicle Miles and Ridership