61 Urban Transit

Peter G. Furth
Northeastern University

61.1 Transit Modes

The principal transit modes are bus, light rail, and metro (heavy rail).

Bus

Bus is the most common transit mode, operating in every urban area in the U.S. Nearly all transit coaches in the U.S. are powered by diesel engines, although experimentation with alternative fuels has grown since the enactment of the Clean Air Act of 1990. The standard 40-ft coach can seat 40–55 passengers, depending on seating configuration. Smaller coaches are common in settings of lower passenger demand. Large, articulated coaches, seating 60–75, are common in some cities on high-volume routes.

The bus mode uses the existing road network. Its two greatest advantages are its low capital cost and its ability to access transit demand anywhere. Sharing the road with general traffic is also the bus's main

TABLE 61.1 Priority Schemes for Bus

a. On freeways [often shared with other high-occupancy vehicles (HOVs)]
  - Median HOV roadways
  - With-flow HOV lanes
  - Contraflow HOV lanes

b. On arterials and downtown streets
  - Bus lane (curb lane)
  - Express bus lane (inside lane — no stops)
  - Contraflow lane on one-way street
  - Bus-only street
  - Exemption from turning restrictions
  - Priority merge when departing from bus stop

c. At traffic signals, toll booths, ramp meters, and other bottlenecks
  - Timing signals to favor buses' progression
  - Signal preemption
  - Queue bypass lanes

weakness: Buses suffer traffic delays, and ride quality often suffers due to poor pavement quality. Priority schemes, such as those listed in Table 61.1, can be used to reduce traffic delays. The ultimate priority scheme is a *busway,* a bus-only roadway with grade separation that gives buses a comparable level of service to that of rail lines [Bonsall, 1987]. Cities with busways include Pittsburgh, Ottawa, and Adelaide, Australia.

Light Rail

Streetcar was once the dominant transit mode, operating on tracks laid in city streets. Their replacement by buses, beginning around 1930 and largely completed by 1960, was due in part to the lower capital cost of bus systems and in part to the streetcar's inflexibility in mixed traffic. For example, to avoid being blocked by parked vehicles, tracks were usually laid in the inside lane of a multilane street, forcing passengers to board and alight in the middle of the street instead of at the curb. Few such streetcar operations remain in North America. Still surviving, and growing in number, are systems operating primarily on their own right-of-way, sometimes grade separated.

Light rail's main advantage over bus is its economy in carrying high passenger volumes, since rail cars are larger and can be joined into trains. The economy of a single operator staffing a multicar train requires self-service fare collection, usually entailing ticket-vending machines on platforms, validators (ticket-canceling machines) on platforms or on vehicles, and random fare inspection. Other advantages are that light rail produces no fumes, offers a higher-quality ride, and takes less space, which substantially lowers tunneling cost. The main disadvantage of light rail is the need for its own right-of-way, which can be compromised in small sections (e.g., a downtown transit mall) and at grade crossings, with a corresponding loss of speed. Given the right-of-way, another disadvantage of light rail versus bus is the large number of passengers who must transfer between rail and feeder bus, in contrast to a busway used as a trunk from which routes branch off, covering a wide area.

Metro

Metro, or heavy-rail systems, are high-cost, high-capacity systems operating in an exclusive, grade-separated right-of-way, often in subway, but often elevated or at grade. Floors flush with the platforms and wide doors make for rapid boarding and alighting and easy accessibility for disabled persons. Modern systems feature automatic or nearly automatic control, allowing for small headways and more reliable operation.

The distinction between light rail and heavy rail is becoming blurred in intermediate systems such as those in Lille, France, and Vancouver, Canada. They use small vehicles characteristic of light rail but have high platform loading and automatic control characteristic of metro systems.

Other Modes

Trolleybus coaches resemble diesel coaches, except that they are powered by electric motors, drawing electric current from overhead wires. They are common in Europe and operate in several North American cities. Their virtually unlimited power enables them to accelerate more quickly and climb steep hills more easily than diesel buses, and because they need no transmission, their ride is smoother. Fumes are eliminated, improving the environment and making it easier for them to operate in tunnels. The overhead wires, however, are sometimes seen as a detriment to the environment. Trolleybuses are not as flexible as diesel buses — they must be replaced by diesel buses when there is a detour, for example — but they can maneuver around a blocked lane, making them better suited to mixed traffic than light rail. The cost of the power line network generally limits them to high-volume routes, since the benefits of electrification are proportional to the number of passengers and vehicles on the route.

A more recent bus variation is the self-steering bus, guided by contact between a raised curb and small guide wheel that extends from the side of the bus. It allows a bus to operate in a smaller lane, lowering construction costs on elevated busways and in tunnels.

Commuter rail has long been used in older U.S. cities for long-distance commuting. Extensions are being built, and new systems have recently been opened in metropolitan Washington, in Los Angeles, and in south Florida. They use standard locomotives (electric or diesel) pulling passenger cars on rail lines often shared with freight or long-distance passenger traffic. More than any other mode, commuter rail relies on auto access at the suburban end of the trip and usually relies on metro for distribution at the downtown end. The design and operation of downtown terminals where many lines converge can be quite involved.

Numerous other transit modes operate in different cities. They include ferry boat, cable car, incline, and, more recently, downtown people mover.

Line Capacity

The line capacity of a transit line is the number of passengers that can be carried per hour in one direction past any point. Line capacity is often a major consideration of choice of mode for a new transit line. If there is no shortage of vehicles, the line capacity is constrained only by the headway (time interval) between vehicles:

$$\text{Line capacity}_{(\text{pass/h})} = \text{Vehicle capacity}_{(\text{pass/veh})} \times \text{Train size}_{(\text{veh})} \times 3600 / \text{Minimum headway}_{(\text{s})} \quad (61.1)$$

Vehicle capacity includes standees, according to the level of crowding deemed acceptable. Minimum headway is a function of safety and is governed primarily by vehicle interference at stations. For example, a metro with 6-car trains that can fit 240 people per car, if operating at a 4-min headway, can carry 21,600 pass/h in each direction; at a 2-min headway, line capacity is doubled. Typical line capacities for various modes are given in Table 61.2. By contrast, a freeway lane with headways of about 1.8 s and average occupancy of 1.2 persons carries only 2400 people per hour. If that lane were converted to a

TABLE 61.2 Typical Line Capacities

Mode	Train Size	Minimum Headway	Occupancy	Line Capacity
Auto on freeway	1	1.8 s	1.2	2,400/h/lane
High-occupancy freeway lane (5% buses)	1	2 s	5	9,000/h
Bus-only freeway lane	1	4 s	40	36,000/h
Bus in arterial bus lane	1	30 s	65	7,800/h
Busway	1	20 s	60	10,800/h
Light-rail exclusive way	2	1 min	150	18,000/h
Metro	6	2 min	240	43,200/h
Commuter rail	10	10 min	275	13,750/h

high-occupancy vehicle lane, with an average occupancy of 5 (95% carpools, 5% buses), then even with half the vehicular volume its passenger volume would more than double and its line capacity would almost quadruple, since the minimum headway would only barely increase.

Comparing Alternatives

Many studies have been done comparing transit alternatives. There is no clear consensus on the superiority or inferiority of any mode in a general sense. Any investment using federal funds requires an *alternative analysis* that considers a no-build alternative (including low-cost transportation systems management improvements) and at least two different modes, with variations in alignment for each. Basic analysis is done using methods described in Chapter 58. Evaluation of alternatives is done following federal guidelines considering capital cost, operating cost, expected number of new passengers, benefits to existing passengers, and financing and political considerations [Zimmerman, 1989]. As a rule of thumb, construction cost for elevated exclusive guideway is 2 to 3 times greater than for at-grade, and subway is 4 to 10 times more costly than at-grade, creating a strong incentive to utilize existing rail rights-of-way and other alignments that avoid the need for tunneling or aerial construction.

61.2 The Transit Environment

Travel Patterns and Urban Form

Between about 1870 and 1940, streetcar (first horse drawn, later electric) was the primary mode of urban travel. Consequently, urban development during this period was oriented around streetcar use — dense development along radial streetcar lines, with a heavy concentration of commercial development in the central business district (CBD). Postwar development, in contrast, has been largely auto-oriented. At first homes, later stores, and finally employers became dispersed in large numbers in the suburbs. The travel patterns of streetcar-era development — many to one or many to many along a linear corridor — lend themselves to the kind of demand concentration that transit can serve easily. The dispersed travel patterns of auto-oriented urban land use are far more difficult for transit to serve [Pushkarev and Zupan, 1977]. A transit system in an older city with a concentrated urban form faces far different problems than one serving a dispersed urban form. The former can face challenges such as how to carry the enormous demand; the main challenge in the latter is how to attract passengers.

Historically, as trip ends dispersed and income and auto ownership increased, transit ridership declined. “Captive” markets, primarily the carless poor and persons unable to drive, have shrunk. Within the “choice” market, discretionary trips (e.g., shopping trips that can be arranged to be done when an auto is available and where parking is free) were hardest hit. In cities with large downtown employment, parking is expensive, and the home-to-work commute market has held its own and in some cases grown. As a result, transit demand in large cities has become more and more peaked. The *peak-to-base ratio* (the number of buses in service during the a.m. peak period divided by the number in service during the base period) can be as great as 3:1. Passenger utilization is still more peaked because vehicles are more crowded during peak hours. This level of peaking hurts transit’s economy because fixed facilities are underutilized and because of the costs inherent in starting and stopping service.

Transit Financing

Although transit first developed as a profitable private enterprise, inflation coupled with politically mandated caps on fares and competition from autos began to cripple the industry by the end of the first World War, leading to consolidation, disinvestment (e.g., abandoning streetcar lines, not replacing aging vehicles), public subsidy, and eventually public ownership. The chief reasons for transit’s not being profitable are:

- *Low fare.* Political pressure has kept fares low for a variety of reasons, including accommodating poor riders and providing an alternative to autos which, transit proponents argue, are heavily subsidized.
- *Low demand and highly peaked demand.* These factors prevent transit from achieving economies of scale.
- *Social service.* It is politically mandated that service be offered on routes and at times of low demand.
- *High wages.* All large transit systems are unionized. When a large percentage of downtown workers use transit, the threat of a strike gives unions considerable bargaining power, which they have used effectively to negotiate high wages.
- *Restrictive work rules and various management problems.* These factors have also been blamed for transit's financial losses.

Capital costs are financed entirely by government subsidy. In the U.S. the Federal Transit Administration (FTA) usually covers 80%, with state and local government covering the remainder. For major construction projects, limited federal funds sometimes result in a smaller federal share. Public subsidies cover, as a national average, about 50% of operating costs in the U.S., although they vary a great deal from city to city. The federal contribution is small except in small cities; state and local governments cover most of the operating deficit. Instability in the source of state and local funding is a cause of much uncertainty in management in many cities.

Transit Management

Ownership of public transit agencies is exercised through a board of directors whose members usually represent the local political constituencies (cities, counties, etc.) that subsidize it. Because the board members are political appointees, many have little knowledge about managing a transit system. Because these members carry with them political views that sometimes conflict with one another, management can be politically charged, making the direction unstable. For example, urban board members may want to keep fares low, whereas suburban board members may be primarily concerned with reducing subsidies. Construction projects always entail significant political interest. Depending on how the balance of power changes, direction may change often.

The chief executive officer, usually called the *general manager,* is appointed by the board. The organization is usually divided functionally into departments such as transportation, maintenance, finance, administration, planning, real estate, and construction. Larger agencies may use a modal breakdown (bus, rail, etc.), as well, which may be under or above the functional division.

61.3 Fundamentals of Cyclic Operations

Most transit services are cyclical: vehicles leave a depot, cycle over a given route, return, and then make another cycle. Routes that go back and forth between two terminals can still be considered cyclic, and either terminal can be considered the depot. This section looks at a route during a period of the day in which ridership and running times can be treated as constant.

Fundamental Operating Parameters and Relationships

Cycle time (c) is the time a vehicle uses to perform the cycle and wait for the next cycle. It consists of running time and layover (or recovery) time. On routes with two terminals, layover time is usually distributed between the two terminals. Its primary purpose is to serve as a buffer for run-time delays, reducing the degree to which delays propagate from one cycle to the next. It also allows vehicle operators a rest. On short urban bus routes, layover is commonly 15 to 20% of running time; on longer routes or routes with less traffic congestion, 10% is typical. Layover may be further increased by schedule slack, as discussed later.

The service frequency (q) is the number of cycles per hour, the number of trips per hour passing a given point in a given direction. (*Trip* in transit terminology refers to vehicle trips, unless otherwise designated.) The reciprocal of frequency is headway (h), the time between successive trips.

If a route is operated in isolation, one can speak of the number of vehicles (n) operating on the route. In the rail context, n is the number of trains. The fundamental relationship is

$$c = nh = n/q \tag{61.2a}$$

$$n = c/h = cq \tag{61.2b}$$

$$h = 1/q = c/n \tag{61.2c}$$

For example, to operate a route with a 4-min cycle at a 10-min headway (a frequency of 6/h) will require 4 vehicles. Or, given 6 vehicles and a 4-min cycle, the route can operate with a 6.67-min headway (a frequency of 9/h).

Fundamental Measures of Passenger Demand

An origin–destination (O–D) matrix — showing the number of passengers per hour traveling from one stop to another — is the fundamental descriptor of passenger demand on a route. (Further detail, such as fare category, usually does not matter for operations planning.) On routes operating between two terminals, it is best to divide the demand by direction, resulting in a triangular O–D matrix for each direction. The row and column totals represent the ons and offs (boardings and alightings) at each stop. The grand total is the boardings (b) on the route.

The volume profile is the passenger volume on each interstop segment on the route. On routes that empty out at a terminal, the volume profile is easily constructed stop by stop, beginning at the terminal, accumulating ons and deducting offs. On loop routes that do not empty out (e.g., a circumferential route), the volume on a segment is the sum of the demands in the O–D cells that involve travel over that segment. Once the volume on any one segment is calculated, the volume on the succeeding segments can be found by accumulating ons and deducting offs. The peak volume segment (also called *peak load point* or *peak point*) is the segment with the greatest volume; its volume is the peak volume (v*). On bidirectional routes, the direction with the greater peak volume is the peak direction, and its peak volume is the route's peak volume. Another measure of demand is passenger-miles (or passenger-km), most easily calculated by multiplying the volume on each segment by the segment length and summing over all segments:

$$\text{pass-mi (or pass-km)} = \sum \left(\text{Volume on segment } i\right)\left(\text{Length of segment } i\right) \tag{61.3}$$

The main quantifiable measures of service quality for transit passengers are travel time (in vehicle); waiting time, which is approximately half the headway when service is regular and headway is not too large; number of transfers; and level of crowding.

Basic Schedule Design

The simplest practical scheduling method involves three constraints in addition to the fundamental relationship [Eq. (61.2)]. First, the average vehicle load at the peak point, called *peak load* (l_p), must not exceed a design capacity (k), which depends on the size of the bus as well as standards of comfort and safety:

$$l_p = \frac{v^*}{q} \le k \tag{61.4}$$

Second, the headway is usually restricted to a set of acceptable values, $\{h\}$. This set is based on four considerations: (1) whole minute headways are usually required because schedules are written in whole minutes (exception: some rail systems use the half-minute or quarter-minute as the basic unit); (2) multiples of 5 min are desired for long headways; (3) headways that repeat every hour (e.g., 12, 15, 20, 30 min) are desirable; and (4) there is a maximum headway, called a *policy headway,* that may not be exceeded, usually 60 min but sometimes 30 min or smaller in peak periods. For example, one set of acceptable headways might be

$$\{h_{\mathrm{I}}\} = \{1, 2, \ldots, 20, 25, 30, 35, 40, 45, 50, 55, 60\} \tag{61.5}$$

while a more restrictive set might be

$$\{h_{\mathrm{II}}\} = \{1, 2, \ldots, 10, 12, 15, 20, 30, 60\} \tag{61.6}$$

Third, the number of vehicles must be an integer (unless the route is not operated in isolation, in which case scheduling must be done jointly for a number of routes, as discussed in Section 61.5). The result of these last two constraints is to force additional slack into the schedule, in the form of both excess capacity and excess layover.

Schedule design usually begins with a given peak volume and a given minimum cycle time ($c_{\min}$) that accounts for running time and minimum necessary layover. The schedule design procedure that follows has as its primary objective minimizing fleet size (about the same as minimizing cost); its secondary objective, for a given fleet size, is to maximize service frequency (maximize service quality). In what follows, $[\ ]^+$ means round up and $[\ ]^-$ means round down. This procedure assumes that cycle times and headways are in minutes, while frequencies and passenger volumes are hourly.

Step 1. $h_{\max} = [k/(v^*/60)]^-$ (round down to next acceptable headway)

Step 2. $n = [c_{\min}/h_{\max}]^+$

Step 3. $h = [c_{\min}/n]^+$ (round up to next acceptable headway)

Step 4. Given n and h, determine the remaining parameters (c, q, l_p) using Eqs. (61.2) and (61.4). The difference between c and $c_{\min}$, called *schedule slack,* is added to the layover.

The rounding involved in steps 1 and 2 can add substantially to operating cost. For example, consider a route for which $v^* = 260/h$, $c_{\min} = 51$ min, and $k = 50$. If one ignores rounding, the minimal service frequency is 260/50 = 5.2/h, the headway is 60/5.2 = 11.5 min, and the number of vehicles needed is 51/11.5 = 4.4. While this kind of analysis can be done in sketch planning, it does not produce a workable design. Following are two designs using the preceding procedure; their difference is that one uses set $\{h_{\mathrm{I}}\}$, which allows an 11-min headway, while the other uses $\{h_{\mathrm{II}}\}$, which does not.

Case	Set of Acceptable Headways	$h_{\max}$ Unrounded (min)	$h_{\max}$ (min)	n Unrounded	n	h (min)	c (min)	l_p
I	$\{h_1\}$	11.5	11	4.64	5	11	55	47.7
II	$\{h_{\mathrm{II}}\}$	11.5	10	5.1	6	9	54	39.0

This example demonstrates the substantial effect of rounding. In case I, rounding increased fleet requirements from the sketch planning value of 4.4 to 5. The extra resources consumed are manifest as slack in the cycle time (the final cycle time, 55 min, is 4 min greater than required) and in slack capacity (peak load, 47.7, is below the allowed capacity of 50). Case II, by not permitting an 11-min headway, requires more rounding, increasing the vehicle requirement to 6. However, the extra resources are not all wasted but are partially converted into extra service as service frequency increases, reducing passenger waiting time and crowding. Case II also illustrates the role of step 3 in achieving the secondary objective of maximizing service level. Step 3 could have been omitted, leaving $h = h_{\max} = 10$ min, and the result would have been a viable design. However, the rounding involved in calculating n (step 2) made a better value of h (9 min vs. 10 min) possible without adding a vehicle.

Finally, a schedule with a large amount of slack time in the cycle begs for opportunities to adjust the minimum cycle time, either lowering it enough to save a vehicle (e.g., by eliminating a deviation or securing traffic improvements) or lengthening it by an amount less than or equal to the slack in an effort to attract new passengers (e.g., extending the route or adding a deviation) without increasing the fleet requirement.

Example 61.1

A downtown circulator on its own right-of-way with six tops is being planned. Stops, numbered clockwise, are 0.5 mi apart, and travel time (including dwell time at stops) is 1.5 min per segment, for a 9-min overall running time. There should be little or no layover because of the nature of the service; likewise, the only restriction on headway is that it be in quarter minutes. Vehicle design capacity is 40. Two alternative configurations are to be compared. Alternative A is service in the clockwise direction only. Alternative B is service in both directions; naturally, alternative B involves a greater construction cost.

Estimated p.m. peak demand is shown in the O–D matrix in Table 61.3(a). For alternative A the volume on segment 6-1 is the sum of the cells in the O–D matrix that involve travel over that segment; those cells are shaded in Table 61.3(a). Volume on the remaining segments is found by accumulating ons and subtracting offs, resulting in the volume profile shown in Table 61.3(b). Peak volume is seen to be 1333/h. Multiplying volume by segment length and by segment travel time results in passenger-miles and passenger-minute estimates.

TABLE 61.3 Demand Analysis, One-Directional Circumferential Route

a. Origin–Destination Matrix (passengers/hr)

\ TO FROM \	1	2	3	4	5	6	TOTAL
1		150	150	300	150	150	900
2	100		50	100	50	50	350
3	100	33		100	50	50	333
4	200	67	67		67	67	467
5	100	33	33	33		50	250
6	100	33	33	33	33		233
TOTAL	600	317	333	567	350	367	2533

b. Volume Profile

SEGMENT AFTER STOP	6	1	2	3	4	5	6	TOTAL
OFF		600	317	333	567	3501	367	2533
ON		900	350	333	467	250	233	2533
VOLUME	1000	1300	1333	1333	1233	1133	1000	
MILES		0.5	0.5	0.5	0.5	0.5	0.5	
PASS-MI		650	667	667	617	567	500	3667
MIN		1.5	1.5	1.5	1.5	1.5	1.5	
PASS-MIN		1950	2000	2000	1850	1700	1500	11000

Note: Volume on first segment shown is sum of shaded cells in O–D matrix.

Following the four steps for schedule design,

$$h_{\max} = \left[\frac{40}{(1333/60)}\right]^{-} = [1.80]^{-} = 1.75 \text{ min}$$

$$n = \left[\frac{9}{1.75}\right]^{+} = [5.14]^{+} = 6$$

$$h = \left[\frac{9}{6}\right]^{+} = [1.5]^{+} = 1.5 \text{ min}$$

$$c = (1.5)(6) = 9 \text{ min, and layover} = 0$$

$$q = 60/1.5 = 40/\text{h}$$

$$l_p = 1333/40 = 33.3$$

For alternative B demand must be split between the two directions. Assuming that passengers choose the shortest direction and that those traveling three stops split themselves evenly between the two directions, O–D matrices for the two routes are shown in Table 61.4(a). Volume profiles [Table 61.4(b)] are constructed as they were for alternative A, with the shaded cells in the corresponding O–D matrix constituting load on the first segment. The peak volumes are 517/h on the clockwise route and 550/h on the counterclockwise route. Schedule design for the two routes is as follows:

Clockwise route	Counterclockwise route
$h_{\max} = \left[\frac{40}{(517/60)}\right]^{-} = [4.64]^{-} = 4.5 \text{ min}$	$h_{\max} = \left[\frac{40}{(550/60)}\right]^{-} = [4.36]^{-} = 4.25 \text{ min}$
$n = \left[\frac{9}{4.5}\right]^{+} = 2$	$n = \left[\frac{9}{4.25}\right]^{+} = [2.12]^{+} = 3$
$h = \left[\frac{9}{2}\right]^{+} = 4.5 \text{ min}$	$h = \left[\frac{9}{3}\right]^{+} = 3 \text{ min}$
$c = (2)(4.5) = 9 \text{ min, layover} = 0$	$c = (3)(3) = 9 \text{ min, layover} = 0$
$q = 60/4.5 = 13.33/\text{h}$	$q = 60/3 = 20/\text{h}$
$l_p = 517/13.33 = 38.8$	$l_p = 550/20 = 27.5$

A comparison between alternatives A and B is given in Table 61.5. Most of the table is self-explanatory. Averages in rows 3, 5, 8, and 10 are found by dividing the previous figure by total boardings (row 1). Average wait time in row 6 is taken to be half the headway; total wait (row 7) is the product of the average and the total boardings. In some alternative analyses, wait time is weighted more heavily than travel time; this measure has not been taken in this example. In comparing the two examples, one can see that alternative B requires one fewer vehicle and involves less passenger time overall. A decision between the two alternatives should consider other factors as well, including vehicle requirements in other periods, capital costs, operating statistics in other periods and in the future, and sensitivity to changes in the demand estates.

TABLE 61.4 Demand Analysis, Two Circumferential Routes

CLOCKWISE ROUTE

a. Origin–Destination Matrix (passengers / hr)

\ TO FROM /	1	2	3	4	5	6	TOTAL
1		150	150	150	0	0	450
2	0		50	100	25	0	175
3	0	0		100	50	25	176
4	100	0	0		67	67	233
5	100	17	0	0		50	167
6	100	33	17	0	0		150
TOTAL	300	200	217	350	142	142	1350

b. Volume Profile

SEGMENT AFTER STOP	6	1	2	3	4	5	6	TOTAL
OFF		300	200	217	350	142	142	1350
ON		450	175	175	233	167	150	1350
VOLUME	367	517	492	450	333	358	367	
MILES		0.5	0.5	0.5	0.5	0.5	0.5	
PASS-MI		258	246	225	167	179	183	1258
MIN		1.5	1.5	1.5	1.5	1.5	1.5	
PASS-MIN		775	737	675	500	538	550	3775

COUNTERCLOCKWISE ROUTE

a. Origin–Destination Matrix (passengers / hr)

\ TO FROM \	1	2	3	4	5	6	TOTAL
1		0	0	150	150	150	450
2	100		0	0	25	50	175
3	100	33		3	3	25	158
4	100	67	67		0	0	233
5	0	17	33	33		0	83
6	0	0	17	33	33		83
TOTAL	300	117	117	217	208	225	1183

b. Volume Profile

SEGMENT AFTER STOP	1	6	5	4	3	2	1	TOTAL
OFF		225	208	217	117	117	300	1183
ON		83	83	233	158	175	450	1183
VOLUME	550	408	283	300	342	400	550	
MILES		0.5	0.5	0.5	0.5	0.5	0.5	
PASS-MI		204	142	150	171	200	275	1142
MIN		1.5	1.5	1.5	1.5	1.5	1.5	
PASS-MIN		612	425	450	512	600	825	3425

Note: volume on first segment is sum of shaded cells in O–D matrix.

TABLE 61.5 Comparison of Alternatives

		Alternative B		
	Alternative A	Clockwise	Counterclockwise	Total
1. Boardings (pass/h)	2,533	1350	1183	2,533
2. Passenger-miles (per/h)	3,667	1258	1142	2,400
3. Average trip length (mi)	1.45			0.95
4. Travel time (pass-min/h)	11,000	3775	3425	7,200
5. Average ride time (min)	4.34			2.84
6. Average wait time (min)	0.75	2.25	1.5	
7. Total wait time (pass-min/h)	1,900	3038	1775	4,812
8. Average wait time (min)	0.75			1.90
9. Total pass-min per h	12,900			12,012
10. Average travel + ride time (min)	5.09			4.74
11. Vehicles needed	6	2	3	5

61.4 Frequency Determination

Most bus, light rail, and metro routes operate at a constant headway over a time period. When demand is very low, routes follow a policy headway (H), the maximum headway allowed by system policy. When demand is very high, the headway is set so that average peak load equals or is just below the design capacity, as described in Section 61.3.

In between very low and very high demand, there is no widely accepted method for setting frequencies. An optimization framework, based on a tradeoff between operator cost and passenger waiting time, provides a rule that can be used to consistently set frequencies on routes. Let

$$(\text{OC}) = \text{operating cost per vehicle hour } (\$/\text{veh-h})$$

$$(\text{VOT}) = \text{value of passenger wait time } (\$/\text{pass-h})$$

Assuming that wait time is half the headway, the combined operator plus passenger cost per hour for route i is

$$(\text{OC})c_i\, q_i + 0.5(\text{VOT})b_i/q_i$$

Summing over all routes and then minimizing by setting to zero the derivative with respect to q_i yields the "square root rule":

$$q_i = \sqrt{\frac{0.5(\text{VOT})b_i}{(\text{OC})c_i}} \tag{61.7}$$

Incorporating policy headway and capacity constraints, the rule has two modifications:

- If the solution is below $60/H$, set $q_i = 60/H$ (use policy headway).
- If the solution is below v_i^*/k, set $q_i = v_i^*/k$ (make peak load equal capacity).

In addition, solutions must be rounded appropriately to satisfy integer constraints as described earlier.

The value of time can be explicitly specified as a matter of policy (one half the average wage is typical). Alternatively, it can be implicitly determined by a constraint on total operating cost per hour of system operation [$(\text{OC}) \sum c_i\, q_i \le \text{Budget}$], in which case the value of time will be that value for which the total operating cost when the q_i values are set by the constrained square root rule equals the budget [Furth and Wilson, 1981]. The framework can be further generalized by recognizing that demand is not fixed

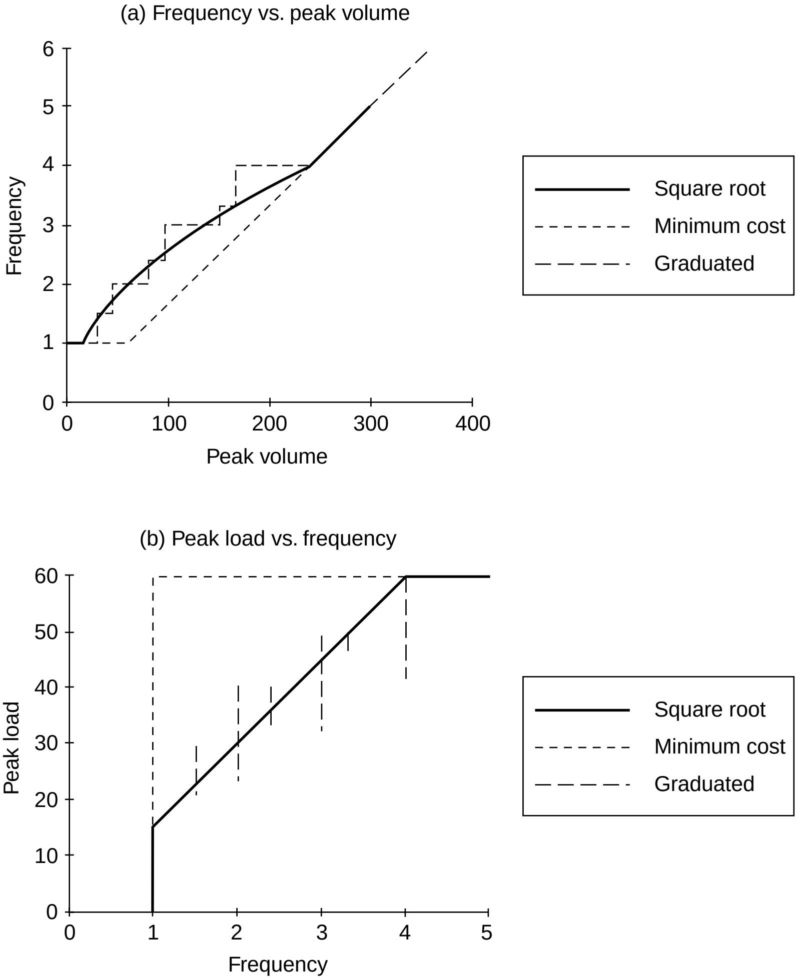


FIGURE 61.1 Frequency determination rules.

but will respond to frequency changes, so the generalized cost function should include the societal benefit of increased ridership and consumer surplus. However, nearly the same result will be reached if demand is treated as constant because the goals of minimizing waiting time for existing passengers and trying to attract new passengers are so much in harmony. If the driving limitation is number of vehicles of varying types instead of an operating cost budget, the same framework applies, with OC = 1 and Budget equal to fleet size. A useful dynamic programming solution to the latter problem using a similar optimization framework is found in Hasselstrom [1981]; unlike the calculus-based models, it yields solutions that do not need to be rounded.

Figure 61.1 depicts graphically how the constrained square root rule applies to routes with varying levels of ridership in comparison with minimum cost scheduling, using typical values for OC ($60/veh-h), c_i (1h), H (60 min), k (60), VOT ($4/hour), and the ratio v_i^*/b_i (0.5). Part (a) shows how frequency varies with ridership; part (b) shows how peak load varies with frequency.
The cures have three ranges:

1. For low-volume routes, policy headway governs; frequency is independent of passenger volume; increases in ridership are simply absorbed by increasing load.
2. For intermediate-volume routes, the square root formula applies; frequency increases with passenger volume, though less than proportionally, because part of the passenger volume increase is

absorbed by increasing the peak load. This intermediate portion is missing under minimum cost scheduling.

3. For high-volume routes, the load constraint governs; all routes are at capacity, and frequency is proportional to passenger volume.

Many transit systems claim to be minimizing operating cost subject only to policy headway and capacity constraints. If that were the case, the minimum cost curve would be followed. However, in practice it is uncommon to find a route with a policy headway whose peak load is nearly equal to capacity. This is because the scheduling rules followed by most transit agencies — whether informal or formal — resemble the square root rule in that they include a transition between policy headway and capacity constrained. An example is a graduated peak load standard used by some agencies, also illustrated in Fig. 61.1. It states that maximum peak load (design capacity) decreases from a base value of 60 to 50, 40, and finally 30 on routes whose headway is above 15, 20, and 30 min, respectively. It is somewhat "saw-toothed" due to the requirement that, beyond 12 min, only certain headways are used. The fact that it closely parallels the constrained square root rule and at the same time is more readily understood by operations planners (for example, it avoids the troublesome value-of-time parameter) makes it a useful rule.

In rail systems, frequency determination has an additional dimension — train length. Whether train length should vary between peak and off-peak involves a tradeoff in coupling costs as well as the usual operating costs and passenger convenience.

On many commuter rail routes and some express bus routes, cycle length is so long and demand so peaked that it makes little sense to speak of a constant service frequency within a time period. Scheduling in such cases is done at a more detailed level, tailoring departure time for each trip to passenger demand. An example is load-based scheduling for evening peak express service leaving a downtown terminal. One would construct a profile of cumulative passenger arrivals versus time, which may be quite irregular with numerous short peaks corresponding to common quitting times such as 4:00, 4:30, and 5:00. Departures are then scheduled whenever the cumulative arrivals since the last departure equals the desired vehicle load.

61.5 Scheduling and Routing

Desired headways and running times can vary throughout the day, usually with two peak periods when more vehicles and operators are needed than in the midday (base) period. In some cities the periods within which running time and/or headway change can be as small as 20 min. For this and other reasons, scheduling is far more complex than the fundamental case described in Section 61.3.

Given the desired timetable and running times on a network of routes, scheduling is the task of creating vehicle and operator duties to perform the specified service. A basic overview of vehicle scheduling is given in the following discussions. Because scheduling both vehicles and operators is so complex, most transit agencies use automated scheduling. Nevertheless, it is still important for route and schedule designers to understand some basic scheduling paradigms. The final discussion of Section 61.5 briefly describes route design.

Deficit Function Analysis

The number of vehicles needed to meet an arbitrary timetable (i.e., without any expectation of constant running times or headways) involving several terminals and routes can be found using *deficit functions* [Ceder and Stern, 1981]. The deficit at terminal i at time t, $d_i(t)$, is the cumulative number of departures from that terminal minus the cumulative number of arrivals at that terminal as of time t. Each departure from a terminal increases that terminal's deficit by 1; each arrival lowers it by 1. For a small network, deficit functions can be easily drawn, as in Fig. 61.2. Let d_i^* = the greatest deficit occurring at terminal i; it represents that number of vehicles that must be on hand at that terminal at the start of the day to keep that terminal from running out of vehicles. The total vehicle requirement for the network is the sum of the d_i^* values. Once the vehicle requirement is known, vehicle duties can be easily constructed by simply chaining trips together, minimizing layover. For an isolated route with trips scheduled as round trips,

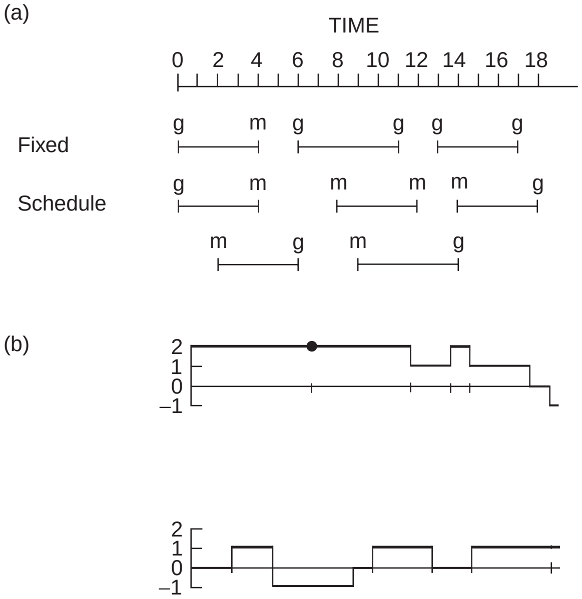


FIGURE 61.2 (a) A two-terminal fixed schedule between terminals g and m. (b) Deficit functions at g (above) and m.

another way of saying that the fleet requirement equals the peak deficit is to say that the fleet requirement equals the maximum number of departures occurring during any round-trip window.

Schedule adjustments to reduce the fleet requirement should naturally be aimed at reducing the peak deficit at the various terminals. Sometimes shifting a trip's time by a minute or two can reduce the peak deficit at one terminal without increasing it at another. Another such adjustment is to add to the schedule a deadhead (i.e., empty) trip that leaves one terminal after the time of its peak deficit and arrives at another terminal before the time of its peak deficit, reducing by one the peak deficit at the second terminal without increasing the peak deficit at the first. This kind of schedule analysis has proven particularly helpful in improving the efficiency of regional bus operations, where headways can be quite long or can vary greatly and routes frequently do not operate as simple loops.

Network Analysis

A more comprehensive and flexible framework for analyzing schedules is network analysis. Each node in the network represents a terminal and a time (either a departure or arrival time). The network has five kinds of links:

1. *Service links,* representing trips in the timetable. These links have a minimum "flow" of one.
2. *Layover links,* going from one node to another node representing the same location at a later time. These links have no minimum flow.
3. *Deadhead links,* joining a node to another node representing a different location at a later time (the time must be enough later that the connection can be made). These links have no minimum flow.
4. *Source links,* leaving a central source node, representing vehicles entering service.
5. *Sink links,* going to a central sink node, representing vehicles leaving service.

An example network is shown in Fig. 61.3. The fleet requirement for the schedule is the minimum flow that must begin at the source node, filter through the network satisfying flow conservation at every node (inflow = outflow) and meeting the flow requirements of the service links, and return to the sink node. Mathematical programming solutions to this problem, including a linear programming approach, are

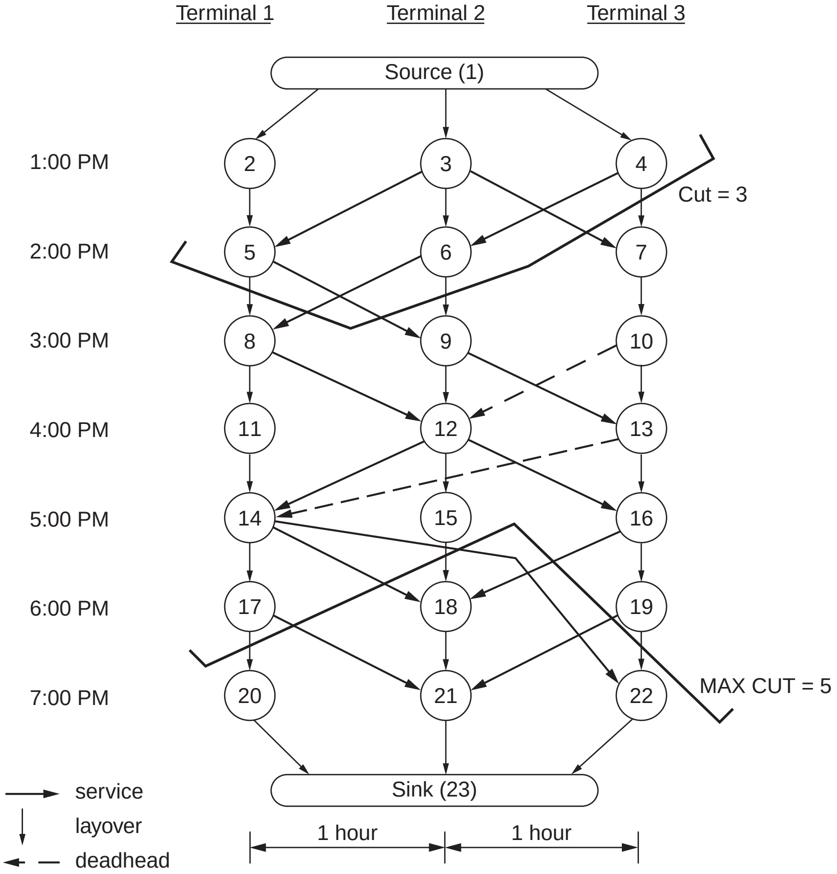


FIGURE 61.3 Network analysis.

well known. One intuitive way of looking for a solution is to examine cuts that divide the network. A valid cut (1) divides the network into two parts, an early part containing the source node and a late part containing the sink node; (2) intersects no link more than once; and (3) intersects links in such a way that the beginning of every intersected link lies in the early part of the network, and the end of every intersected link lies in the late part. The minimum number of vehicles needed is the greatest number of service links that can be intersected by a valid cut. In Fig. 61.3, two valid cuts are shown, one intersecting 3 service links, the other intersecting 5 service links. The reader can verify that the fleet requirement is indeed 5.

A refinement is to seek the minimum cost solution [Scott, 1984]. Costs are assessed to deadhead and layover links reflecting labor and vehicle costs. A flow of one is required on service arcs, so no cost needs to be applied. The cost of a source link should reflect the cost of deadheading from the garage plus a substantial penalty for increasing the peak vehicle requirement. Sink links are assessed the cost of deadheading to the garage. The problem is now a "transshipment problem," which has several well-known solution algorithms, including the *network simplex algorithm.* To apply transshipment model algorithms, it is necessary to consider the source and sink nodes as being connected to every other node.

Automated Scheduling

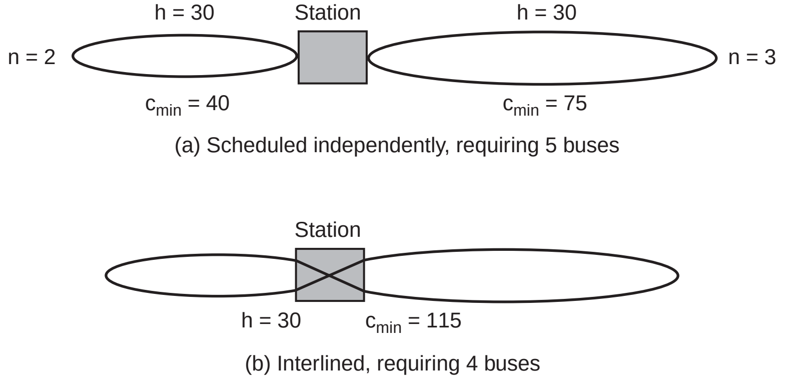
Vehicles can easily operate for 16 or 20 hours without a break, but operators' schedules are constrained by a variety of work rules, such as minimum number of paid hours (typically 8), maximum number of hours (typically 8.25 to 9), restrictions on the number of part-time operators, paid breaks after 5 hours of uninterrupted service, maximum spread (for a "split shift," i.e., a shift with an unpaid break in the middle, where spread is the amount of time between beginning and end of the workday), pay premiums for time after 8 hours and for spread exceeding a certain amount, and the requirement that every shift end where it began. Most automated scheduling packages first do run cutting — using network-optimization procedures described earlier to create efficient vehicle schedules — and then operator scheduling, using other optimization methods to split the vehicle schedules into pieces of about 4 hours and then match them into legal, efficient operator schedules.

There are several scheduling packages on the market that compete with one another and with manual scheduling. Although the cost of these packages can be high, benchmark tests have shown their ability to reduce operator labor costs by up to 3%, well justifying the investment [Blais et al., 1990]. Scheduling software does a great deal of valuable bookkeeping and usually includes graphical interfaces that enable schedulers to easily make manual schedule adjustments. The software can also be integrated with other information systems, such as timetable publishing, payroll, work force management, and data collection, adding to their value.

Interlining and Through-Routing

Interlining means scheduling a vehicle to switch between routes. It can be done on an *ad hoc* basis, but it can also be done systematically in a scheme that can be called *cyclical interlining*. Cyclical interlining means that two or more routes with a common headway and a common terminal are scheduled jointly in such a way that each vehicle does a round trip on one route followed by a round trip on the other. Figure 61.4 illustrates such a situation. In part (a), the two routes are scheduled independently and require a total of 5 vehicles. In part (b), they are interlined, with an aggregate cycle consisting of the two route cycles back to back, and require only 4 buses. Cyclical interlining can save a vehicle whenever the combined schedule slack of the two routes equals or exceeds their common headway. Interlining can also be done with more than two routes. If *k* routes are interlined, all having a common headway and common terminus, it is theoretically possible to save as many as *k* – 1 vehicles if each route has much schedule slack. In practice it is uncommon for groups of more than three routes to be cyclically interlined. Cyclically interlined routes can maintain separate names for the public, or the route combination can have a single name; either way, to the scheduler they are a single unit. While the primary motivation for cyclical interlining is to reduce vehicle requirements by eliminating schedule slack, interlining also benefits passengers by eliminating the need to transfer between the routes that are interlined.

When two routes are interlined to save a vehicle, some freedom in choosing departure times is lost. For example, in Fig. 61.4(a), it is possible for the routes to have simultaneous departures. However, in Fig. 61.4(b), once interlined and operated with 4 vehicles, they cannot have simultaneous departures. If there is a route 1 departure at 10:00, the same vehicle will depart on route 2 immediately after completing



(a) Scheduled independently, requiring 5 buses

(b) Interlined, requiring 4 buses

FIGURE 61.4 Cyclical interlining routing.

the route 1 cycle, that is, at 10:40, or up to 5 min later, since the interlined cycle has 5 min slack. Working backward with a 30-min headway, it is apparent that previous route 2 departures must lie in the time windows at 10:10–10:15, 9:40–9:45, etc. If the route 2 departures are scheduled outside this 5-min window, there will be no vehicle savings from interlining.

Through-routing is the practice of joining two routes that go to (roughly) opposite sides of the central business district (CBD) or a major terminal. In small cities, where it is common for all the radial routes to share a single, central CBD terminal, through-routing is the same as cyclical interlining. In cities with large CBDs, good CBD distribution can be attained only if routes extend more than halfway into the CBD, resulting in several CBD terminals and considerable cross-CBD traffic. In such a case, joining two opposite radial routes is more than just interlining, since it also means sharing the cross-CBD link. The benefits of through-routing include the following:

1. Sharing the cross-CBD reduces the combined cycle time, potentially reducing vehicle requirements. Combining the cycles also tends to reduce the loss due to slack (as with cyclical interlining).
2. More extensive CBD distribution for passenger convenience occurs, especially for reaching destinations such as universities and hospitals that are often located on the fringe of the CBD.
3. A smaller contribution to traffic congestion in the CBD occurs.
4. Rail lines are nearly always through-routed to avoid the need for duplicate trackage, tunneling, etc.
5. Time-consuming (and, for rail systems, expensive) turnarounds in the CBD are eliminated.
6. Layovers in the CBD, where space is at a premium, are likewise eliminated.

Through-routing has two disadvantages, however. First, eliminating CBD layovers means that departures from the CBD are less prone to be on time. This can cause severe crowding in the p.m. peak, when evenly spaced departures from the CBD are essential to keeping loads balanced. The longer routes are harder to control in other ways, as well. For this reason, some large cities avoid through-routing. Second, routes that are through-routed must have a common headway and common vehicle and train sizes at all times. They cannot easily be uncoupled when demand and, consequently, desired headways on the two sides of the CBD differ. If, during a certain period of the day or at some time in the future, the east side of a through-route requires an 8-min headway (7.5 trips/hour) while the west side requires a 6-min headway (10 trips/hour), the combined route must be operated with the smaller headway, implying that 2.5 trips/hour on the east side will be made unnecessarily. When designing through-routed rail lines, one design objective is to balance the demands on the two sides of the CBD by varying line length and availability of parking.

Pulse (Timed Transfer) Systems

Perhaps the most-disliked aspect of the passenger journey is the transfer from one route to another. Transfers cannot be eliminated; there is simply not enough demand to afford direct service for every desired trip. Pulse scheduling is a concept aimed at making transfers less onerous [Vuchic, 1981]. Routes are concentrated at a single terminal (in more complex versions of pulse scheduling, there are several major pulse points) and all have simultaneous departures at a common headway (usually every hour or half-hour). Vehicles from all the routes arrive shortly before a pulse, then depart together, leaving a small time window for passengers to transfer between routes, similar to operations at an airline hub. Route cycle times must be a multiple of the common headway, so routes should be laid out so that their running time is just under a multiple of the common headway to avoid excess schedule slack (which cannot be reduced by interlining due to the need for simultaneous departures). It is possible to compromise the ideal plan to accommodate routes of greater and less demand; for instance, high-demand routes could pulse every 15 min, while low-demand routes could pulse every 30 min.

The main advantages of a pulse system are as follows:

1. The waiting time when transferring is small.
2. Compared to transferring on a street corner, the transfer environment is vastly improved. There is increased passenger security (due to not being alone), there is less anxiety about having missed a connection, and amenities can include concession stands and protected waiting areas.

3. The schedule is easy for passengers to remember.
4. The centralized terminal can increase transit's visibility and improve its image to its patrons and the community.

The chief disadvantages of a pulse system are as follows:

1. There is a cost for building and maintaining the transfer center.
2. There is increased need for space at the transfer center, which is usually located in the CBD or another major activity center where space is at a premium. In a pulse system, each route needs its own berth; if departures were staggered, routes could share berths.
3. The need for cycle time to be a multiple of the common headway can lead to considerable schedule slack, increasing vehicle and operator requirements. Schedule slack can be reduced by making routes more circuitous in an effort to reach more passengers, but the benefits are usually small. Interlining cannot save buses because departures must be simultaneous.

Pulse systems are most effective in small cities where most routes operate at long (30- or 60-min) headways; in larger-sized cities during evening hours; or in suburban areas, sometimes at rail stations, to facilitate bus-to-bus transfers. When demand is great enough that headways become smaller than 20 or 30 min, the transferring benefit of pulse scheduling becomes smaller and the operating cost impact becomes too large to make pulse scheduling practical. Timed transfers can be also arranged for a pair or small subset of routes that share a common headway.

Operating Strategies for High-Demand Corridors

Corridors with high passenger demand create opportunities for differentiating service to better serve certain markets and to better match capacity to demand. Along low-demand routes there is little alternative to the standard local bus route that satisfies the fundamental needs of access and an acceptably small headway. As demand grows, one has the choice of simply increasing the frequency of the local route or employing other routing and scheduling strategies for bus service. The strategies described in this section apply equally to rail service except to the extent that they rely on overtaking.

Express Service

A common strategy is to supplement local bus service with express service. Passengers obviously benefit from the reduced travel time, and operating cost can likewise be reduced due to higher speed if the load on the express route is sufficient. Express service in many cities attracts auto users via park-and-ride lots. The speed attainable depends, of course, on the available roadways. Priority treatments on highways, arterials, and in the downtown can make a great difference [Levinson et al., 1975]. Express service in a high-volume corridor must offer a sufficient travel-time savings over the local route and a small-enough headway to "compete" successfully with the local route and thus capture its own market. Then the express and local services can be scheduled independently.

Zonal Express Service

When the demand for express service is great, rather than to simply increase the frequency of the express route, it is often more cost effective to divide the corridor into zones and provide an express route for each zone [Turnquist, 1979]. Passengers in the outer zones will enjoy a further travel-time savings (but a longer headway), and operating cost will be reduced as fewer vehicles have to cover the full length of the corridor. The same approach has been successfully applied to commuter rail [Salzborn, 1969].

Alternating Deadheading

During peak periods when there is a strong direction imbalance in travel demand, some routes, particularly express routes, deadhead (return empty) in the reverse direction. If offering reverse-direction service is desirable but demand is still far below peak direction demand, and there exists a high-speed path for reverse-direction deadheading, buses can sometimes be saved by having only a fraction of the trips return in service, with the remainder deadheading [Furth, 1985]. Because the deadheading trips must be coordinated with those returning in service (since they both will continue on the same peak direction route), a systematic coordination "mode" will be needed. The most effective mode is often 1:1, meaning that for every trip returning in service, one deadheads. As a rule of thumb, a bus can be saved if the time saved by deadheading equals two headways. The extent of alternating deadheading is limited by capacity and the policy headway constraints.

Restricted Zonal Service

Zonal design can be applied to radial local service as well as to express if the peak volume is large and the volume profile shows a steady increase from the outer end of the route and to just outside the downtown [Furth, 1986]. The corridor is divided into zones, each large enough to support a route with an acceptably small headway. Buses never leave the main route of the corridor (unless deadheading) but employ boarding and alighting restrictions between their zone and the downtown terminal. Inbound, buses let passengers board in their zone only but let them alight anywhere. Outbound, alighting is restricted to the route's zone, while boarding is unrestricted. With this strategy, direct service is still offered between every pair of stops in the corridor, but there is only one zonal route that a passenger can take between any given origin and destination. Therefore, the corridor O–D matrix can simply be split into the markets served by the different routes, and each route can be scheduled independently. In general, the advantage of restricted zonal service is that it allows the passenger-carrying capacity to increase along the route as the volume profile increases, reducing unused capacity in the outer zones and thereby saving vehicles. However, once an inbound bus enters the portion of its route in which boarding is restricted, alighting passengers cannot be replaced, resulting in some unused capacity in the inner zones. For this reason, restricted zonal service is effective only when the proportion of outer zone riders alighting before the downtown is small. Another disadvantage of the strategy is that with more routes, there is more unproductive slack due to rounding. Moreover, there can be problems in passenger understanding of and acceptance of the boarding and alighting restrictions, although the strategy has been used successfully in some cities for years.

Short-Turning

Short-turning, like restricted zonal service, means some buses traverse only the inner portion of the route, allowing provided capacity to more closely match demand [Furth, 1988]. Unlike restricted service, there are no boarding or alighting restrictions. In a two-route system (three-route systems are uncommon for practical reasons), passengers with either an origin or destination in the outer zone must use a bus serving the full route, while those whose trip lies entirely within the inner zone can use buses on either the full or short-turning route. Efficient operation demands that most of these "choice" passengers use the short-turning route. Unless a reduced fare can be offered on the short-turning route, the way to effect this choice is to coordinate the scheduling of the routes, having a short-turning bus lead a full-route bus by a small time interval. For example, full-route buses might pass the turnback point at 7:00, 7:10, 7:20, etc., while short-turning buses leave the turnback point at 6:58, 7:08, 7:18, etc. In this example, the headway module is 10 min, divided into a "leader's headway" of 8 min and a "follower's headway" of 2 min. Each short-turning bus will therefore carry 8 minutes' worth of the choice market, while each full-route bus carries only 2 minutes' worth of the choice market. This is an example of 1:1 schedule coordination. Other coordination modes are also possible. For example, if full-route buses pass the turnback point at 7:00, 7:10, 7:20, etc., 1:2 coordination might have short-turning buses depart at 7:04, 7:08, 7:14, 7:18, 7:24, 7:28, etc. Then each full-route bus will get two minutes' worth of the choice market, while each short-turning bus gets four minutes' worth. The challenge of design is to choose the turnback point(s) and headway module and then split the headway module in such a way that the resulting split in the choice market gives each route a peak volume near, but not exceeding, design capacity.

Other strategies for high-volume corridors include *limited stop* and *skip-stop* service [Furth, 1985].

Route Design

For the most part, route design is done manually, using standard evaluation methods to choose between alternative routings. Standards have been developed in different agencies regarding route length, circuity, stop spacing, route spacing, and, of course, expected demand.

Efforts to automate route (and, for that matter, entire network) design have resulted in some useful models that are available as software packages [Hasselstrom, 1981; Babin et al., 1982; Chapleau, 1986]. Some of these packages have the capacity to select route alignments, but their main contribution is in evaluating alternative networks. The investment in the software and in the network coding has limited their use to major investment analyses (e.g., design of a new rail line) and to systems with a large and dynamic ridership.

In newly developing areas, successful route layout can be very difficult if developers ignore transit and pedestrian access. Whenever possible, transit agencies and political authorities should try to influence developers to facilitate transit use by such measures as providing through streets that transit routes can use; providing walkways and pedestrian bridges for direct, easy pedestrian access to through streets; siting commercial buildings to allow for easy pedestrian access to through streets; and clustering high-density uses near potential transit routes.

61.6 Patronage Prediction and Pricing

Evaluation of proposed service or fare changes requires prediction of ridership and revenue impacts. The first requisite for such evaluation is methods for measuring current ridership, discussed in Sections 61.8 and 61.9. Nearly all patronage-forecasting methods assume a knowledge of current ridership, and before/after studies, a valuable analysis technique, require little more than accurate measurement of actual ridership.

The amount of effort spent on predicting impacts of service changes should be proportionate to the cost of making the service change. For example, consider implementation of a new Saturday bus service requiring one bus for 8 hours a day. It is obviously not worth investing 100 hours in analysis to predict how successfully the service will be when operating it on a trial basis for 6 months would cost only 200 bus-hrs. For low-cost changes, a low-cost prediction method to screen out changes that are unlikely to succeed coupled with before/after evaluation is appropriate. On the other hand, for very large investments, full-scale modeling and demand-forecasting techniques, discussed in Chapter 58, are appropriate. For short-range transit planning, prediction methods can be divided into two groups: predicting ridership changes in response to service changes and predicting ridership on a new service.

Predicting Changes

The most commonly used method to predict ridership changes, particularly in response to fare changes, is elasticities. Fare elasticity is the relative change in demand divided by the relative change in fare, or

$$\varepsilon = \frac{\Delta\text{Demand}/\text{Demand}}{\Delta\text{Price}/\text{Price}} \tag{61.8}$$

Most fare elasticities measured from before/after studies lie between –0.10 and –0.70. The industry rule of thumb for many years was $\varepsilon = -0.30$; more recently, experience indicates that elasticity is closer to –0.2. Factors that lead to smaller elasticity (i.e., elasticity closer to 0) include high transit dependency, a predominance of work and school trips, and a low current fare. Opposite factors, such as a predominance of discretionary trips, lead to greater elasticity.

Predictions using fare elasticity can be made directly from Eq. (61.8), in which demand and price are entered at their base levels. For example, using the old rule-of-thumb elasticity, a fare increase from \$0.75 to \$1.00 will cause a relative ridership change of

$$\frac{\Delta\text{Demand}}{\text{Demand}} = -0.3\frac{\$0.25}{\$0.75} = -0.10$$

that is, a 10% drop. Of course, this simplistic example assumes that everyone pays full fare. Some market segments — those making transfers, those using passes, and those getting discounts, for example — may experience different fare changes and may have different elasticities. It is therefore preferable to make predictions by market segment or at least to base the prediction on average price rather than nominal fare.

Equation (61.8) represents one type of elasticity, the *shrinkage ratio.* There are some inherent inconsistencies in this form — for example, if fare in the previous example returns to \$0.75, predicted demand will not return to its original value. The *log-arc* elasticity is a theoretically consistent form. Log-arc elasticities are estimated from before/after data using the following equation:

$$\varepsilon = \frac{\ln\left(\text{New demand}/\text{Old demand}\right)}{\ln\left(\text{New price}/\text{Old price}\right)} \tag{61.9}$$

Use of the log-arc elasticity for making predictions is illustrated for the previous example:

$$\frac{\text{New demand}}{\text{Old demand}} = \left(\frac{\text{New price}}{\text{Old price}}\right)^{\varepsilon} = \left(\frac{\$1.00}{\$0.75}\right)^{-0.3} = 0.917 \tag{61.10}$$

implying an 8.3% drop in demand. The log-arc forms give different answers than the shrinkage ratio form, and the differences can be very large for large changes. There is a third type of elasticity as well — linear-arc elasticity — for which predictions are not materially different from predictions made using log-arc elasticity.

Although the log-arc elasticity is consistent with itself, it is not entirely consistent with reality — for example, it predicts infinite demand when price goes to zero. In reality, elasticity changes as price and demand change — demand becomes more elastic (elasticity increases in magnitude) as price becomes greater and as the transit mode share becomes smaller. One way of facing this reality is to carefully select an elasticity from a catalog of elasticities [Mayworm et al., 1980; Charles River Associates and Levinson, 1988], trying to best match it to current circumstances.

Another way to face the reality of varying elasticity is to use an incremental demand model that does not assume constant elasticity. The incremental logit method, an abbreviated form of the logit model described in Chapter 58, is such a method. It looks at transit demand as a share of the wider market that could use the transit service in question. The prediction formula is

$$\frac{\text{New demand}}{\text{Old demand}} = \frac{e^{(\text{coef})(\Delta\ \text{price})}}{(\text{shr})e^{(\text{coef})(\Delta\ \text{price})} + (1 - \text{shr})} \tag{61.11}$$

where shr = current transit share and coef = logit model coefficient for price. Logit model coefficients are not as widely catalogued as elasticities. A typical value is –0.6/\$, so if transit's share of all trips that could use transit is 20%, the previous example leads to the prediction

$$\frac{\text{New demand}}{\text{Old demand}} = \frac{e^{-0.6(0.25)}}{(0.2)e^{-0.6(0.25)} + 0.8} = 0.89$$

implying a drop in demand of 11%.

The incremental logit method implies a point elasticity of

$$\varepsilon = (1 - \text{shr})(\text{coef})(\text{Current price}) \tag{61.12}$$

which for the example comes out to be $\varepsilon = (1 - 0.2)(-0.6)(0.75) = -0.36$. As Eq. (61.12) indicates, it is inherent in the incremental logit method for elasticity to change with both price and transit share. Equation (61.12) can also be used to determine a coefficient that is consistent with a given elasticity. The main drawbacks of the incremental logit model are the need to specify a coefficient and to estimate the transit share.

Changes in service attributes such as headway can be evaluated in a similar manner to fare changes, using either elasticities or incremental logit. Elasticities with respect to attributes other than price are less well studied and should be applied with care. The incremental logit model is suitable when the service change can be expressed as a change in a passenger's utility. For example, headway changes affect passenger waiting time, which is a part of most logit utility functions, with a typical coefficient of about –0.08/min. A change in headway from 18 to 12 min implies a drop in average waiting time of about 3 min, so the prediction will be

$$\frac{\text{New demand}}{\text{Old demand}} = \frac{e^{-0.08(-3)}}{(\text{shr})e^{-0.08(-3)} + (1-\text{shr})}$$

which yields an increase of 21% if the transit share is 20% (i.e., an increase from a share of 20% to a share of 24.2%).

Revenue Forecasting and Pricing

Revenue is simply ridership times average fare. Because different markets pay different fares and are affected differently by fare and service changes, revenue forecasting is best done by market segment. Market segmentation can be done along various lines, depending on the purpose. For example, the market can be segmented by type of fare paid (cash, pass, bulk purchase); by service (bus, metro, both bus and metro); by time of day (peak, off-peak, weekend); and by location (city, suburb, suburb to city). The matter is further complicated by the fact that some markets are fluid. For example, changes in pricing can make patrons switch from using a pass to paying cash, and so on.

Manipulating Eq. (61.9) leads to

$$\frac{\text{New revenue}}{\text{Old revenue}} = \frac{\text{New demand} \times \text{New price}}{\text{Old demand} \times \text{Old price}} = \left(\frac{\text{New price}}{\text{Old price}}\right)^{1+\varepsilon} \tag{61.13}$$

A desire is often expressed to increase revenue by *lowering* fares, the idea being that so many new riders will be attracted that revenues from them will more than make up the loss from current riders. As demonstrated by Eq. (61.13), this will not happen unless $|\varepsilon| > 1$, a situation almost never seen in transit. With typical low elasticities (around –0.2), fare increases will lead to revenue increases, although they are not as large (relatively) as the fare increases themselves due to the loss in ridership.

Political and fiscal realities often lead planners to look for ways to increase revenue with the smallest possible attendant loss in ridership. The solution, in a theoretical sense, is to raise fares in the least-elastic markets while holding steady or even lowering fares in the most-elastic markets. This approach, called *price discrimination,* is widely practiced by the airlines. In transit, it has been the basis for peak/off-peak price differentials, because peak-period demand is less elastic, and for deep discounting for occasional riders, who are considered to be a relatively elastic market (i.e., willing to use transit more if offered a discount).

An example given in Table 61.6 illustrates this phenomenon. A uniform fare increase from $1.00 to $1.20 raises revenue by $13.6 million, but at a ridership loss of 5.3%. A targeted fare increase, raising the fare to $1.30 for the less-elastic market (for argument's sake, peak-period travelers) while lowering the fare to $0.90 for the more-elastic market, yields the same revenue increase with a ridership loss of only 1.5%.

Because transit services are subsidized, pricing determines in part how subsidies are distributed. There has been strong criticism that some pricing policies subsidize high-income users more than low-income users [Cervero, 1981]. Pricing efficiency has been hindered by practical difficulties with distance and time-based pricing. Some of these difficulties may be eliminated and new opportunities opened by advanced information technology that is beginning to appear in the industry.

TABLE 61.6 Varying Ridership Impacts of Fare Increases

		Current			Uniform Fare Increase			Targeted Fare Increase		
	Elasticity	Fare ($)	Riders (Million)	Revenue ($Million)	Fare ($)	Riders (Million)	Revenue ($Million)	Fare ($)	Riders (Million)	Revenue ($Million)
Market 1	–0.20	1.00	65.7	65.7	1.20	63.3	76.0	1.30	62.3	81.0
Market 2	–0.50	1.00	34.3	34.3	1.20	31.3	37.6	0.90	36.2	32.5
Total			100.0	100.0		94.7	113.6		98.5	113.6

Prediction for Service to New Markets

There is no well-defined method that is widely used for predicting patronage on a service to a market not currently served by transit. A common approach is to compare the new service to similar services in similar areas, with subjective adjustments made to account for the extent of dissimilarity. For example, a city might find that existing park-and-ride (P&R) express lots attract 12 riders per 1000 inhabitants living within a 3 mi × 3 mi square centered on the lot. This factor could be used for predicting ridership in a new market, with adjustments for population density, service frequency, income, distance from the city center, and so on.

Because predictions will necessarily be imprecise, it is important to monitor ridership and evaluate the new services at regular intervals, retaining services that meet established criteria for new services. Such criteria might take a form such as "at least 10 passengers per service hour after 3 months," "at least 15 passengers per service hour after 6 months," and so on.

61.7 Operating Cost Models

Decisions regarding transit service often require estimates of the operating cost. The most simplistic way of estimating operating cost is based on a single factor, such as vehicle-miles. A transit agency can simply divide its annual operating costs by the annual vehicle-miles, resulting in a figure such as $4.20/veh-mi, and multiply this by the change in vehicle-miles involved in a proposed change in service.

A more accurate model is based on several factors, such as vehicle-miles, vehicle-hours, peak vehicles (number of vehicles needed in the peak), trackage, and revenue. For bus service, a commonly used model has this form:

$$\Delta\text{Operating cost} = b_1\left(\Delta\text{vehicle-miles}\right) + b_2\left(\Delta\text{vehicle-hours}\right) + b_3\left(\Delta\text{peak vehicles}\right)$$

where Δ means "change in" and b_1, b_2, and b_3 are coefficients estimated from accounting data. To estimate the coefficients, historical operating costs are allocated to one or more factors based on the most likely causal relationship. For example, costs for fuel, tires, and maintenance are allocated to factor 1, vehicle-miles. Labor costs for bus operators and supervisors are allocated to factor 2, vehicle-hours. Costs for insurance, space, and various overhead costs are usually allocated to factor 3, peak vehicles. Some costs may be allocated to more than one factor; for example, the cost of marketing may be allocated 50% to factor 1 and 50% to factor 3. The total cost allocated to each of various factors is then divided by the corresponding factor total (e.g., for factor 1, total vehicle-miles) to determine the coefficients b_1, b_2, and b_3 [Booz, Allen & Hamilton, 1981; Cherwony and Mundle, 1980].

It is generally conceded that peak-period service is more expensive to operate than off-peak service. A schedule change that will require any additional vehicles operating during the peak can be substantially more costly than a schedule change (involving the same change in vehicle-miles and vehicle-hours) that would not require an extra vehicle during the peak but would involve keeping a vehicle busy for longer during the off-peak. This phenomenon is partly reflected in cost models that include peak vehicles as a cost factor. However, the typical models do not reflect the additional labor and maintenance costs associated with peaking (due to spread penalties, pull-outs, etc.). Efforts to better account for these factors have led to peak/off-peak cost models, which are theoretically appealing but not widely accepted because they lack an objective basis for allocating costs between peak and off-peak.

Most cost models are *fully allocated* models, in which the full annual operating cost is allocated to the various factors when estimating the coefficients. The result is estimates of *average* cost per factor unit instead of *marginal* cost, which is more appropriate for a model used to estimate the cost of service changes. Marginal cost coefficients can be estimated by eliminating from the accounting fixed costs and by using more complex models for various cost components. However, there is no widely accepted agreement on which costs are fixed since, in the long run, virtually all costs are variable, so it is common to use simple average costs. The use of complex cost models for estimating labor costs of a schedule change has, for the most part, been made obsolete by the availability of scheduling software described in Section 61.5, with which a detailed optimized schedule complete with actual operator costs and vehicle mileage can easily be obtained for any proposed service change.

61.8 Monitoring Operations, Ridership, and Service Quality

Effective transit management requires ongoing monitoring of what is actually happening: how close operations are to what was planned, how many riders there are, and the quality of the service delivered.

Operations Monitoring

Virtually all systems have supervisors or inspectors to oversee vehicle operators, see that service is operating as planned, authorize adjustments in response to disruptions, and maintain logs. In most transit systems, vehicles are equipped with two-way radios that can be used for obtaining information from operators as well as for sending messages. For many operations, these measures provide adequate monitoring.

In rail systems, varying degrees of electronic monitoring are used. At the extreme end are automated systems that have constant communication between the vehicles and the central computer and therefore constant monitoring of every vehicle's location, speed, and other attributes. Other systems maintain some degree of manual control but still have constant communication with the central computer used to tell the operator desired speed or desired dwell time at a station. Either way, constant communication means that the location of every vehicle can be displayed on an electronic map, and various statistics such as actual headway at key points can be constantly monitored.

Systems lacking the facility for constant communication can use detectors located at key points along the track to monitor movements. If the detection system uses radio technology to identify a vehicle number (i.e., by sending a signal to a passing vehicle which is returned by an on-vehicle transponder), it is called an *automatic vehicle identification (AVI) system.* If it merely notes the presence of a passing vehicle, it is just a throughput detector. While either type of detector can monitor headway at detection points, AVI is vital for identifying service on various branches of a route, out-of-service vehicles, vehicles that have turned back, and individual vehicle running times.

Bus systems in mixed traffic do not afford the possibility of constant communication. Real-time monitoring can be done with a radio-based automatic vehicle location or automatic vehicle monitoring (AVL or AVM) system. A central radio tower polls every vehicle in turn once every polling cycle (typically one to three minutes) by sending out a signal to the effect of "bus number *xyz,* please respond." The bus radio then responds by sending back a stream of information that typically includes identifiers, location information, and alarm status (on or off) for various mechanical (e.g., oil pressure) and security alarms. In most existing systems location is ascertained using signposts located along routes that emit a weak radio signal that the bus radio receives as it passes the signpost. When polled, the bus sends back to the central tower the identification number of the signpost most recently passed, and the number of odometer "clicks" since passing the signpost (each click on a digital odometer is typically one axle revolution). Newer systems are being developed that rely on satellite-based systems rather than signposts for location information. AVL systems permit a central computer to display approximate location of voltages and to calculate statistics such as schedule deviation. This kind of information can be used to radio instructions to an operator (e.g., slow down) or to suggest a service change such as a turnback or placing a standby vehicle into service. It can be used to display real-time information to waiting passengers concerning vehicle arrival time. It is also valuable for locating vehicles when the operator activates a silent alarm, which can be of great importance in enhancing the security of operators who may have to drive near-empty buses at night in dangerous or isolated areas.

AVL systems are spreading slowly in the transit community because of their high costs and because of questions about how valuable location information is in practice. Besides cost, another limiting factor is the number of radio channels available. The following formula shows the relationship between the various parameters of an AVL system:

$$\text{Polling cycle} = \frac{\text{No. of vehicles}}{\text{No. of channels}} \times \text{Poll length} \tag{61.14}$$

For example, if there is only one channel and each pool takes 2 s, then monitoring 400 buses will require a polling cycle of 800 s (more than 13 min) — an unacceptably long time. In order to poll 400 vehicles every 120 s, 7 channels will be needed. In a large city, obtaining permission from the appropriate authorities to use 7 radio channels for an AVL system, in addition to the channels needed for voice communication, can be difficult. It is obviously desirable to reduce the poll length, which depends primarily on how many bits of information are sent during a poll.

Passenger Counting

Transit systems do not issue point-to-point tickets to all passengers as airlines do, so they must rely on counts and samples. Common types of counts and samples are:

- *Farebox or driver counts.* In some cities, it is policy that every passenger is counted. The current generation of electronic fareboxes makes it possible to count passengers by numerous fare categories by vehicle, with a separate count for each trip.
- *Revenue.* Closely related to ridership, revenue is always counted systemwide, sometimes by farebox or turnstile, sometimes by time of day or route.
- *Ride checks.* An on-board checker records ons and offs by stop, as well as time at key points. Usually checkers are full-time employees of the transit system; sometimes temporary help is used. Hand-held electronic units with stored stop lists make collection and processing easier.
- *Point checks.* A wayside checker records the load of passing vehicles and the time. Accuracy can be questionable, and tinted windows and security considerations limit their use.
- *"No questions asked" surveys.* These involve distributing to each boarding passenger a card coded by origin stop (or segment) and collecting the cards as passengers alight, filing them by destination stop so that both boarding stop and alighting stop are known for each passenger [Stopher et al., 1985]. Response rate is usually over 90%, so the resulting O–D matrix is quite reliable, provided the sample size is large enough.
- *Passenger surveys.* These surveys can request a variety of information used for planning and marketing, such as trip purpose, questions about travel habits (do you have a car? how frequently do you use transit?), trip origin and destination, transfers made, and customer satisfaction. Response rate can vary widely, sometimes as low as 20%. When the response rate is low, non-response bias becomes an issue that puts into question the validity of the expanded results. However, this is the only practical method to obtain much of this information and to learn about *linked* (or *revenue*) *trips* — passenger trips from their initial origin to final destination, including accessing the transit system and transfers between routes.

Other data sources include turnstile counts, ticket and pass sales, and transfer counts and surveys.

A few transit systems use automatic passenger counters, which detect and count ons and offs by stop. Detection is based on either infrared beams across the doorway or instrumented treadle mats. Stop location has to be inferred from odometer and clock readings that are automatically recorded.

Most of the counts listed above are samples — they do not count every passenger, every day. Section 61.9 deals with making estimates from samples and determining their statistical validity and, conversely, determining the sample size needed to ensure a statistically valid estimate.

Service Standards

Effective management is assisted by adopting a set of measures of productivity, efficiency, and service quality that are regularly monitored. For most measures a minimally acceptable level is established based on management goals; this level is called a *performance standard* or a *service standard.* By comparing measures of performance with service standards, performance can be evaluated and, hopefully, improved [Wilson et al., 1984].

Service standards at the system level are effective for monitoring the performance of an overall operating strategy. More helpful, however, are standards at the route level or route/direction/period (R/D/P) level, which can be useful for monitoring individual services, evaluating service changes, and suggesting service improvements.

Some performance measures apply equally to transit and other service industries, such as absenteeism rate and measures of performance that relate to finance or data-processing functions. Likewise, service standards in the vehicle maintenance function are very helpful. This section concentrates on measures that are particular to the transportation function of transit.

One common group of service standards is productivity and economic standards. Productivity standards usually take the form of a ratio of a measure of utilization to a measure of input, such as passengers per vehicle-hour, passengers per vehicle-mile, and passenger-miles per vehicle-mile (average payload). Routes that fail to meet standards may become candidates for elimination or remedial action with more careful monitoring. Comparing the performance of different routes and observing a route's performance over time can suggest improvements or be helpful in evaluating the effect of service or policy changes. Economic performance measures include revenue per vehicle-hour or per vehicle-mile, revenue per passenger (average fare), cost per vehicle-hour or per vehicle-mile, cost per passenger, and the ratio of revenue to operating cost, known as the *recovery ratio.* Some of these measures may be unavailable at the route level because of the difficulty in estimating route level revenue — for example, due to a high level of monthly pass usage — or due to lack of a reliable route-costing model. Some systems have politically mandated recovery ratios; if they fail to meet the standard, they must either reduce operating costs or raise fares.

Efficiency standards are ratios of input to output, such as vehicle-hours per pay hour. Indicators of negative output likewise belong in this category, such as percent missed trips and accidents or breakdowns per vehicle-mile. Because efficiency measures are more under the control of the transit agency than are productivity measures, the agency can be held more accountable for them. In contrast, low productivity could be caused in part by bad management and in part by erosion in the demand for transit due to increased auto ownership or economic recession.

Service quality measures relate to the value or quality of the service as perceived by the passenger. They include relatively static measures that relate to the route network or the schedule. Examples are percentage of dwellings or jobs within 0.5 mi of a transit station; number of departures in the a.m. peak (e.g., on a commuter rail or ferry route); and vehicle-mi of service offered at night. Other measures require monitoring of actual service. The most direct measures of service quality are measures of crowding, on-time performance or waiting time, and travel time. Because the first two of these are strongly affected by randomness in operations, they will be elaborated upon in the following paragraphs. Measures of safety (injuries per 100,000 mi) and passenger satisfaction (complaints or survey results) can be important, as well.

For services with a headway of 10 min or more, for which passengers tend to consult the timetable, on-time performance is measured against scheduled departure times. It is common to measure the percentage of trips that are early, on time, and late. “On time” is usually defined to be 0 to 5 min late. Early trips are especially cause for concern, since they are inexcusable and can leave a passenger stranded for a full headway.

For more frequent services, on-time performance is measured based on headway, since passengers are not usually aiming for a specific trip but hope to enjoy a short wait. The average wait during a given headway, assuming that passengers arrive at random, is half the headway. However, if headways are not regular — that is, if some headways are large while others are small — the average wait is not half the average headway; it is larger than that because, even if there are equal numbers of long headways and short headways, if passengers arrive at random, more will arrive during long headways than short, and thus more will experience a long wait. In fact, assuming that passengers arrive at random and can board the first vehicle that comes by, the average wait is given by the formula

$$\text{Average wait} = \frac{\text{Average headway}}{2}\left(1+v_h^2\right) \tag{61.15}$$

where v_h = coefficient of variation of headway. A reduction in average wait can therefore be accomplished either by reducing the average headway or by reducing headway irregularity. The former remedy implies providing additional service, since average headway is the inverse of service frequency, while the latter requires only better control. For example, with perfectly regular headways, $v_h = 0$, so the average wait is half the average headway. But if vehicles come in bunches of two, with headways alternating between 0 and twice the average headway, $v_h = 1$ and the average wait equals the average headway. Likewise, if headways are so random that they follow the exponential probability distribution, $v_h = 1$ and the average wait equals the average headway. A good measure of service quality with respect to on-time performance, then, is average experienced wait, which can be divided into the scheduled wait time (half the scheduled headway) and the balance (increase due to missed trips and headway irregularities). Sometimes the balance is negative because extra service has been provided.

Another measure of on-time performance for frequent services is the percentage of passengers who wait on scheduled headway or less [Wilson et al., 1992]. Assuming that passengers arrive at a steady rate during the period in question, if the scheduled headway is, say, 5 min, and all of the actual headways are 5 min or less, than 100% of the passengers wait less than a scheduled headway. Now suppose that, over the course of an hour, there is a 6-min headway, a 7-min headway, three 4-min headways, and seven 5-min headways. Then 3 minutes’ worth of passengers (those arriving during the first minute of the 6-min headway and those arriving during the first two minutes of the 7-min headway) have to wait more than 5 min, so only 57/60, or 95% of the passengers, wait less than a scheduled headway. A drawback of this measure is that it tends to be worse for lower headway services, when, in fact, having to wait more than one headway when the headway is very small (say, 3 min) is not as serious as extra waiting time when the headway is longer.

Crowding, like average waiting time, depends on both the frequency and the regularity of service as well as on the passenger demand. However, because load variations are primarily due to headway variations, it is not usually the practice to measure variations in load (the airlines are an exception in this regard). Measuring average load will be sufficient to see that enough overall capacity is provided, and measuring headway variation will indicate whether there is sufficient control to keep headways regular and loads balanced.

However, to understand the passenger’s experience, it is helpful to know that the average experienced load at a point is given by

$$\text{Average experienced load} = \left(\text{Average load}\right)\times\left(1+v_{\text{load}}^2\right) \tag{61.16}$$

where v_{load} = coefficient of variation of load at that point. That average experienced load is greater than average load is clear if one considers two trips, one with a large load and the other with a small load. More passengers are on the first one, so more than half the passengers experience the larger load. Therefore, the average experienced load is greater than the average load of the two trips.

Common practice is to measure load at the peak point of the route, although that may not be the peak point for each individual trip. Moreover, load measured at a single point fails to distinguish between crowding that lasts for a long time and crowding that occurs on a brief segment only; on the other hand, load averaged over every route segment tends to hide crowding where it does occur. To further complicate things, the disutility to passengers due to crowding is highly nonlinear. As long as the load is less than the number of seats, there is very little disutility from increasing load. Once there are standees, the marginal effect of additional passengers becomes more and more severe as the crowding increases. No satisfactory measure of crowdedness has been accepted that reflects the passenger's viewpoint of all of these aspects of crowding.

Data Collection Program Design

Every transit agency has a data collection program, although with some it is more formalized than with others. This program consists of a set of data collection activities that are performed regularly; a system of recording, processing, storing, and reporting the data; and a set of measures that are calculated and standards they are compared against [Furth et al., 1985].

The design of the data collection program should first pay attention to data needs, which arise from (1) primary needs of various departments, such as scheduling, planning, budgeting, and marketing; (2) external reporting requirements; and (3) service standards. Second, methods of data collection should be determined. Where automated systems, such as electronic farebox systems, have been already installed, attention should be given to making full use of the data. Where automated systems have not been installed, a full range of methods, from manual to fully automated, can be considered. In many cases the most economical solution is technology-enhanced manual techniques, such as using handheld devices. Third, for items requiring sampling, a sampling strategy and sample sizes must be determined to meet statistical accuracy requirements. Fourth, economies arising from overlapping needs should be identified. For example, a point checker at a single terminal can gather data to meet several needs, such as estimating average load and on-time performance on all the routes that pass that point. By stationing checkers simultaneously at the opposite ends of some of those routes, the same checker's data become useful for estimating running time. Finally, a schedule of data collection activities can be developed that meets the sample size requirements efficiently. It must be coupled with a system for obtaining counts that are not sampled (e.g., revenue counts or counts from turnstiles or the farebox system).

The plan for gathering the data must then be completed with a plan for processing, reporting, and storing the data. Modern database and other data-processing software can greatly expedite this task. Efforts should be made to avoid "information overload" by reporting only what will actually be used. Regular communication between the users of the data and those responsible for gathering and processing the data is absolutely vital to keep the data collection program responsive, to correct errors, and to ensure the best use of the information. If data they are gathering is not being used, data quality will eventually deteriorate. On the other hand, if they are given rapid feedback concerning the value of the data, quality will improve.

61.9 Ridership Estimation and Sampling

Ridership and Passenger-Miles Estimation

Knowing the current ridership is important for planning and scheduling service, estimating transit's benefits, monitoring service effectiveness, and meeting reporting requirements of funding agencies. All operators receiving federal operating assistance (which amounts to nearly all transit operators) must, at a minimum, meet the uniform reporting requirements of Section 15 of the Federal Transit Act (formerly called the Urban Mass Transportation Act of 1964), as amended. Although Section 15 deals mostly with accounting information, it also requires annual estimates of boardings and of passenger-miles for each mode with an accuracy of ±10% precision at the 95% confidence level.

In some transit systems all passengers are counted as a matter of course. However, in many systems — including most large bus systems, barrier-free rail systems, and elsewhere — ridership must be estimated by sampling. In almost all systems, passenger-miles estimates must be made based on sampling. Estimates of other demand measures, including peak load and O–D matrices, also depend on sampling.

Direct Estimation with Simple Random Sampling

To estimate mean boardings per trip, mean load at a given point, or passenger-miles per trip, the item of interest can be measured for a sample of *n* trips and the mean calculated as

$$\bar{y} = \frac{1}{n} \sum y_i \qquad (61.17)$$

where y_i = value trip for *i*, $\bar{y}$ is the sample mean, and *n* is the sample size. Simple random sampling is most easily accomplished by constructing a sampling frame (a list of all the trips) and using a random number sequence to select trips from that frame. "Trip" in this context can mean a one-way trip or a round trip; sampling by round trip is usually more efficient when checkers are involved since they almost always have to return to their starting point. If round trips are the basic sampling unit, the sampling frame is a list of round trips. One-way trips without a natural pair can either stand on their own or be linked to another round trip.

The sample variance is

$$s^2 = \frac{1}{n-1} \sum \left(y_i - \bar{y}\right)^2 \qquad (61.18)$$

The (absolute) tolerance and the (relative) precision of the estimate are

$$\text{Precision} = \frac{ts}{\bar{y}\sqrt{n}}, \quad \text{Tolerance} = \frac{ts}{\sqrt{n}} \qquad (61.19)$$

where *t* is the ordinate of the *t*-distribution for the desired confidence level with $n - 1$ degrees of freedom (*t* values are tabulated in most statistics texts). At the 95% confidence level, with a sample size greater than 30, $t \approx 2.0$. For example, if $\bar{y} = 40$, $s = 24$, and $n = 64$, the tolerance at the 95% confidence level is 6 and the precision is 0.15, or ±15%. This means that one can be 95% confident that the true mean lies within 15% of the estimated mean, that is, in the interval [40 ± 6]. To expand the result to a system total such as total annual boardings, simply multiply the mean by the number of trips in the sampling frame. The tolerance expands likewise, whereas the precision remains unchanged.

To find the sample size necessary to achieve a given precision, these formulas can be reversed. Of course, an estimate of the coefficient of variation ($s/\bar{y}$) will be needed. It is best obtained from historical data. If historical data are unavailable, default coefficients of variation for various measures of interest are found in Furth et al. [1985].

For example, suppose we wish to estimate annual passenger-miles to a precision of ±10% at the 95% confidence level. A sample of round trips will be selected for which on/off counts at every stop will be done, and passenger-miles will be calculated using Eq. (61.3). How large a sample will be needed? The necessary sample size, reversing Eq. (61.17) and using $t + 2$, is

$$n = 2.0^2 \frac{\left(s/\bar{y}\right)^2}{\left(0.10\right)^2}$$

Assuming that $(s/\bar{y})$ was estimated from previous data on round trips to be 1.05, $n = 440$ round trips will be needed. They should be spread throughout the entire year, doing 8 one week and 9 the next, with random sampling within the week. To randomly select 8 or 9 round trips within a week, determine N_{wk} = the number of round trips operated in that week (usually, it is 5 times the number of round trips on the weekday schedule plus the number of weekend round trips) and assign to each trip a sequence number from 1 to N_{wk}. Then select 8 or 9 random numbers between 1 and N_{wk}.

Quarterly estimates, based on only a quarter of the data, will have a precision twice as large as annual estimates, since precision is inversely proportional to the square root of sample size. To get an annual precision of ±1%, 100 times more data would be needed than to achieve a precision of ±10%.

Using Conversion Factors

When the variable to be estimated is closely related to another variable whose total is known, sampling can be done to estimate the ratio, or conversion factor, between the item of interest and the related or *auxiliary item* [Furth and McCollom, 1987]. For example, if the number of total boardings is known from electronic farebox counts, passenger-miles can be estimated by first estimating average passenger-trip length (APL) and then expanding it by total boardings. APL is the ratio of passenger-miles (the item of interest) to boardings (the auxiliary item), which can be estimated from on/off counts made on a sample of trips. If the number of total boardings is not known but total revenue is, a passenger-miles-to-revenue ratio can be estimated by measuring passenger-miles and cash revenue on a sample of trips. This technique can be far less costly than simple direct estimation.

Again, a simple random sample of n trips (either one-way or round trips can be the basic unit) is needed; for each sampled trip, both the item of interest (y) and the auxiliary item (x) are measured. The estimate of the ratio r is

$$r = \frac{\sum y_i}{\sum x_i} \tag{61.20}$$

and the estimated total is found by simple expansion:

$$Y_{\text{total}} = rX_{\text{total}} \tag{61.21}$$

The relative variance per sampled trip is

$$u_r^2 = \frac{1}{\bar{y}^2(n-1)} \sum (y_i - rx_i)^2 \tag{61.22}$$

and the precision of both the ratio and the expanded total is

$$\text{Precision} = \frac{tu_r}{\sqrt{n}} \tag{61.23}$$

This last formula can be inverted to determine necessary sample size, using historical data to determine u_r. For example, suppose electronic fareboxes count all boardings, and a sample of ride checks on round trips will be made to estimate the ratio of passenger-miles to boardings. From a historical sample of round trips, u_r is estimated [Eq. (61.22)] to be 0.6. The necessary sample size to achieve 10% precision at the 95% confidence level would be

$$n = \left(\frac{tu_r}{\text{Precision}}\right)^2 = \left(\frac{(2.0)(0.6)}{0.1}\right)^2 = 144 \tag{61.24}$$

The resulting sampling plan would probably call for 3 round trips per week.

Other Sampling Techniques

It is possible to meet Section 15 sampling requirements by following one of two specified sampling plans, using direct estimation (FTA Circular 2710.1) or ratio-to-revenue estimation (Circular 2710.4), avoiding the need to do statistical analysis. These plans require annually about 550 and 208 individually sampled one-way trips, respectively.

More advanced sampling techniques — including stratified sampling, cluster sampling, and multistage sampling — can be effective in improving precision or reducing necessary sampling size. For example, in estimating passenger-miles using a ratio to boardings, stratifying routes by length can be extremely effective. Cluster sampling can reduce costs by allowing for samples to be taken on efficient clusters of trips. On a system with only one or two lines (e.g., a light-rail system) and a high degree of precision specified, the sample size may be so great that every trip in the weekday schedule can be sampled at least once. Using two-stage sampling, the effect of variance between different scheduled trips can be eliminated due to the finite population correction, and only the variance between days for each scheduled trip will affect the precision. The same applies to a bus route that gets a so-called "100% ride check" in which every trip in the schedule is checked, usually on the same day. The result is still a sample, albeit one in which the trip-to-trip variation has been eliminated and only the day-to-day variation remains.

Estimating a Route-Level Origin–Destination Matrix

An O–D matrix for a route/direction/period (R/D/P) is a valuable input for designing service changes along a route such as short-turns, express service, or route restructuring. There are two general methods for estimating an R/D/P level O–D matrix: direct estimation and updating. Direct estimation is best done using the "no questions asked" survey on a sample of trips (see Section 61.8) and simply expanding the results. This type of survey has a response rate near 100% and does not suffer from biases caused by low response rate in questionnaire-type surveys.

The simplest updating method relies on a sample of ride checks to obtain on and off totals at each stop for the R/D/P, which serve as row and column totals for the O–D matrix. Updating begins with a seed matrix, which can be an old O–D matrix, if available; a small-sample O–D matrix obtained through a questionnaire-type survey; or a matrix of propensities, as used in a gravity model [Ben-Akiva et al., 1985]. Each row is balanced (factored proportionately to match its target row total); then each column is balanced likewise. The process is repeated iteratively until no more adjustments are needed. This method, known as either *biproportional method* or *iterative proportional fit,* is wholly equivalent to the doubly constrained gravity model.

References

Babin, A., Florian, M., James-Lefebvre, L., and Spiess, H. 1982. EMME/2: Interactive graphic method for road and transit planning. *Transportation Research Record.* 866:1–9.

Ben-Akiva, M., Macke, P., and Hsu, P. S. 1985. Alternative methods to estimate route level trip tables and expand on-board surveys. *Transportation Research Record.* 1037:1–11.

Blais, J. Y., Lamont, J., and Rousseau, J. M. 1990. The HASTUS vehicle and manpower scheduling system at the Societe de Transport de la Communaute Urbaine de Montreal. *Interfaces.* 20(1):26–42.

Bonsall, J. A. 1987. *Transitways — The Ottawa Experience.* OC Transpo, Ottawa.

Booz, Allen & Hamilton. 1981. *Bus Route Costing Procedures, Interim Report No. 2: Proposed Method.* Report no. UMTA-IT-09-9014-81-1. Urban Mass Transportation Administration.

Ceder, A. and Stern, H. 1981. Deficit function bus scheduling with deadheading trip insertions for fleet size reduction. *Transportation Science* 15:338–363.

Cervero, R. 1981. Efficiency and equity impacts of current transit fare policies. *Transportation Research Record.* 790:7–15.

Chapleau, R. 1986. *Transit Network Analysis and Evaluation With a Total Different Approach Using MADITUC.* Ecole Polytechnique, Montreal.

Charles River Associates, Inc., and H. S. Levinson. *Characteristics of Urban Transportation Demand — An Update* (rev. ed.) 1988. UMTA Report no. DOT-T-88-18. U.S. Department of Transportation, Washington, D.C.

Cherwony, W. and Mundle, S. R. 1980. Transit cost allocation model development. *Transportation Engineering Journal of ASCE.* 106 (TE1):31–42.

Furth, P. G. Alternating deadheading in bus route operations. 1985. *Transportation Science* 19:13–28.

Furth, P. G. Zonal route design for transit corridors. 1986. *Transportation Science.* 20:1–12.

Furth, P. G. Short-turning on transit routes. 1988. *Transportation Research Record.* 1108:42–52.

Furth, P. G., Attanucci, J. P., Burns, I., and Wilson, N. H. 1985. *Transit Data Collection Design Manual.* Report DOT-I-85-38. U.S. Department of Transportation, Washington, D.C.

Furth, P. G. and Day, F. B. Transit routing and scheduling strategies for heavy demand corridors. 1985. *Transportation Research Record.* 1011:23–26.

Furth, P. G., Killough, K. L., and Ruprecht, G. F. Cluster sampling techniques for estimating transit system patronage. 1988. *Transportation Research Record.* 1165:105–114.

Furth, P. G. and McCollom, B. Using conversion factors to lower transit data collection costs. 1987. *Transportation Research Record.* 1144:1–6.

Furth, P. G. and Wilson, N. H. 1981. Setting frequencies on bus routes: Theory and practice. *Transportation Research Record.* 818:1–7.

Gomez-Ibanez, J. A. and Meyer, J. R. 1993. *Going Private: The International Experience with Transit Privatization.* Brookings Institute, Washington, D.C.

Hasselstrom, D. 1981. *Public Transportation Planning: A Mathematical Programming Approach.* PhD Thesis. Department of Business Administration. University of Gothenburg, Sweden.

Lampkin, W. and Saalmans, P. D. 1967. The design of routes, service frequencies and schedules for a municipal bus undertaking: A case study. *Operations Research Quarterly.* 18(4):375–397.

Levinson, H. S., Adams, C. L., and Hoey, W. F. 1975. *Bus Use of Highways: Planning and Design Guidelines.* NCHRP Report 155. Transportation Research Board, Washington, D.C.

Mayworm, P., Lago, A. M., and McEnroe, J. M. 1980. *Patronage Impacts on Changes in Transit Fares and Services.* Report no. 1205-UT. U.S. Department of Transportation, Washington, D.C.

Metropolitan Transit Authority of Harris County (METRO). 1984. *Bus Service Evaluation Methods: A Review.* Report no. DOT-I-84-49. U.S. Department of Transportation, Washington, D.C.

Pickrell, D. H. 1983. *The Causes of Rising Transit Operating Deficits.* Report no. DOT-I-83-47. U.S. Department of Transportation, Washington, D.C.

Pickrell, D. H. 1989. *Urban Rail Transit Projects: Forecast Versus Actual Ridership and Costs.* Prepared by the Transportation Systems Center for UMTA, U.S. Government Printing Office, Washington, D.C.

Pushkarev, B. S. and Zupan, J. M. 1977. *Public Transportation and Land Use Policy.* Indiana University Press, Bloomington.

Salzborn, F. J. 1969. Timetables for a suburban rail transit system. *Transportation Science.* 3:297–316.

Scott, D. 1984. *A Method for Scheduling Urban Transit Vehicles Which Takes Account of Operation Labor Cost.* Publication #365. Centre de Recherche sur les Transports, Univ. de Montreal.

Stopher, P. R., Shillito, L., Grober, D. T., and Stopher, H. M. 1985. On-board bus surveys: No questions asked. *Transportation Research Record.* 1085:50–57.

TRB. 1980. *Bus Route and Schedule Planning Guidelines,* NCHRP Synthesis of Highway Practice 69. Transportation Research Board, Washington, D.C.

Turnquist, M. 1979. Zone scheduling of urban bus routes. *Transportation Engineering Journal* 105(1):1–12.

Vuchic, V. R. 1981. *Timed Transfer System Planning, Design, and Operation.* Prepared for UMTA University Research and Training Program. Report no. PA-11-0021. Department of Civil and Urban Engineering, University of Pennsylvania, Philadelphia.

Wilson, N. H., Bauer, A., Gonzalez, S., and Shriver, J. 1984. Short range transit planning: Current practice and a proposed framework. Report DOT-I-84-44. U.S. Department of Transportation, Washington, D.C.

Wilson, N. H., Nelson, D., Palmere, A., Grayson, T., and Cederquist, C. 1992. Service quality monitoring for high-frequency transit lines. *Transportation Research Record*. 1349:1–11.

Zimmerman, S. L. 1989. UMTA and major investments: Evaluation process and results. *Transportation Research Record*. 1209:32–36.

Further Information

Valuable comprehensive texts are:

Canadian Transit Handbook (3rd edition), Canadian Urban Transit Association, 1993.

Gray, G. E. and L. A. Hoel (eds.), *Public Transportation* (2nd edition), Prentice Hall, 1992.

Vuchic, V. R., *Urban Public Transportation Systems and Technology,* Prentice Hall, 1981.

For a more thorough treatment of transit management, its political environment, and terminology see

Altshuler, A., *The Urban Transportation System: Politics and Policy Innovation,* MIT Press, 1979.

Fielding, G. J., *Managing Public Transportation Strategically: A Comprehensive Approach to Strengthening Service and Monitoring Performance,* Jossey-Bass, 1987.

Smerk, G. M., *Mass Transit Management: A Handbook for Small Cities; Part 1: Goals, Support and Finance; Part 2: Management and Control; Part 3: Operations; Part 4: Marketing* (3rd ed., rev.), prepared by Indiana University Institute for Urban Transportation for UMTA, Report no. DOT-T-88-12, 1988.

White, P. R., *Public Transport, Its Planning, Management and Operation.* London: Hutchinson, 1986.

Gray, B. H. (ed.), *Urban Public Transportation Glossary,* Transportation Research Board, 1989.

More detail on advanced technology and software can be found in

Odoni, A. R., J. M. Rousseau, and N. H. Wilson, Models in urban and air transportation, in *Handbook on Operations Research and the Public Sector,* Elsevier, 1994.

Paixao, J. and J. R. Daduna (eds.), *Computer-Aided Transit Scheduling,* Springer-Verlag, 1994.

Rousseau, J. M. (ed.), *Computer Scheduling of Public Transport 2,* Elsevier Science Publishers B.V., North-Holland, 1985.

Wren, A. (ed.), *Computer Scheduling of Public Transport Urban Passenger Vehicle and Crew Scheduling,* Elsevier Science Publishers B.V., North-Holland, 1981.

U.S. Department of Transportation, *Advanced Public Transportation Systems: The State of the Art,* Report DOT-VNTSC-UMTA-91-2 and updated annually, U.S. Department of Transportation.

Many helpful conferences are held and reports published by the following organizations:

Federal Transit Administration, U.S. Department of Transportation
American Public Transportation Association
Canadian Urban Transit Association
Transportation Research Board