A Methodology for Forecasting Peak and Off-Peak Travel Volumes

MARTIN WOHL, The Urban Institute

No information can be more important to transport planners, designers, and analysts than reliable forecasts of the peak and off-peak travel volumes on transport networks. Yet, no reasonably complete and valid methodology has been proposed—much less developed and verified—that will permit the transport planner and analyst to properly differentiate travel by time of day; to determine realistically the duration and level of peaking and recognize its dependence on the transport system design and performance; and to account for shifts in trip-making from one hour to another, from one mode to another, and from car pooling to driving alone in response to changes in transport system design, in service, or in price. Further, the usual travel forecasting process treats trip generation as though trip-making were independent of transport system changes, and it treats the trip distribution, modal-split, and route assignment phases as though trip-making choices—whether to travel, final destination, mode, and route—are made sequentially and apart from the circumstances attendant with the other phases. Accordingly, the purpose of this paper is to formulate a model structure such that these phases can be treated simultaneously, that the total amount of trip-making (as well as the destination, modal, and routes choices) can be varied with the transport system and its performance and price characteristics (among other factors), that shifts from car pooling to driving alone can be represented, that shifts from one hour of travel to another can be characterized, and that the amount of travel during peak and off-peak hours (i.e., the absolute buildup or decrease in peak or off-peak flow) can be determined.

•THE TRANSPORT PLANNER and analyst have fashioned numerous models dealing completely or in part with the travel forecasting problem. The literature is vast and hardly needs repeating or a review. To my knowledge, though, none of the available models and techniques deal realistically or structurally with the matter of peaking.

More precisely, none of the models and techniques appropriately recognize that travel during peak periods is dependent both in concept and in actuality on the nature and extent of the transport system, on its performance and price characteristics during both peak and off-peak periods, and on the attendant socioeconomic conditions and preference patterns of travelers for peak and off-peak travel choices.

As we design, analyze, and evaluate marginal or major adaptations to an existing system, what is it that we need to know? In part, we want to know how usage and performance of the system and its parts will be affected. How much extra travel will there be? Which and how many people will shift from one road or mode to another? How many people will shift out of car pools either to driving alone or another mode? Obtaining data by time of day is vital. Will the peak-period volumes increase? Will people shift from other hours of the day to the peak period as capacity is added? Will the peak period shorten and by how much? Will total daily traveling increase?

Paper sponsored by Department of Urban Transportation Planning and presented at the 49th Annual Meeting.

Such questions cannot be answered by our current models and certainly not by either our simplistic peak-hour factoring techniques or our trip-purpose models developed as a proxy for time-of-day or peaking models. Nor can the analyst properly evaluate the benefits and costs stemming from one design or another without having knowledge of both peak and off-peak travel conditions.

To formulate a methodology suitable for forecasting travel volumes and performance conditions during peak and off-peak periods, distinctions must be made between the prediction process requisite for shorter time periods and that relating to longer time periods. For the former, the transport system, population, employment, and land-use patterns can be regarded as fixed. The latter considers the transport system as it affects and is affected by the land-use pattern as well as the growth and distribution of population and employment. It also will be necessary to distinguish between "demand" and "supply." Demand involves the propensity of people to travel with respect to travel service, price, and socioeconomic conditions. Supply describes the performance of the transport system with respect to the amount and composition of travel sustained by it.

LONG TERM VERSUS SHORT TERM FORECASTING

For the distinction pertaining to the time frame for our forecasting, we must ask: Are we attempting to determine the amount of travel taking place at some point in time, given the transport system, land-use, population, and employment patterns? Or are we trying to develop a more dynamic forecasting model capable of forecasting both the short- and long-range travel and land-use conditions?

For the first of these time-frame questions, we must be concerned with travel forecasting in some static or partial equilibrium sense. Specifically, as a basis for analyzing and evaluating our planning, policy, or design actions, we need to know how much travel will take place and the associated travel conditions, both hourly and daily, given the following information:

1. The socioeconomic characteristics of the people;
2. The location and character of business, industry, and residence; and
3. The physical and operating characteristics of the transportation system.

For such short-run or daily travel forecasting involving the transport system, home and business locations and the transit fleet can be regarded as fixed. By contrast, it is hardly clear that the automobile fleet or ownership should be regarded as fixed, even when forecasting trip-making and modal-split over the short run. As travelers choose among modes on a day-to-day basis, many or most of those who travel by auto, particularly those driving alone, probably made that modal choice at the time they purchased the auto and thus are not making a new decision based on the marginal daily service and price circumstances each day.

This is not a simple problem conceptually or operationally, but it is an important one. In terms of predicting the number or percentage of travelers using one mode or another, this auto distinction may not seem important because most auto travelers are car poolers and may well be viewing the travel conditions for the various modes based on the marginal day-to-day circumstances. In terms of examining traffic congestion, however, and the effects of changes in mode or capacity on its reduction, it is the number of drive-alone vehicles that is most important because these vehicles represent the great bulk of the total auto fleet during peak hours and because their drivers probably made their modal-choice decision on more than day-to-day marginal costs.

Also, and to cast this matter in a slightly different fashion, consider the urban dweller who is examining the tradeoffs associated with a suburban versus central-city dwelling unit. Although the living space, privacy, school conditions, type of neighbor, and housing cost are probably the most important factors he considers, no doubt he also takes into account the available modal choices with respect to travel time, convenience, and cost. For the latter, he probably thinks about the total auto ownership and operating costs because a second car often will be required. In short, he buys the second car based on the day-to-day travel time and convenience expectations and on the long-range travel cost factors. If these hypotheses are correct, our modal-split models must indicate these short- and long-range considerations.

Once the analyst has developed the capability of forecasting daily travel volumes and performance conditions, he can turn to the more formidable problem of long-range forecasting. He can ask a host of location-living-transportation behavioral questions such as: How does the buildup of congestion and the attendant costs, taken with other factors of production and preferences (with respect to patterns of living, quality of life indices, work/employment/shopping/business locations, and so forth), influence changes in employer's plant or business growth and location, and in home or work location? How do these shifts then change the performance of the transport system, which in circular fashion then influences other locational shifts or growth patterns? How will native preferences about transportation services and living patterns (to take but two aspects of importance) change over time, either in response to income changes or in response to shifts in society's scale of values and mores? Because shifts in location and growth stem partially from expectations about the daily travel conditions at different points in time, a relation exists between long- and short-range forecasting. Essentially, the long-run changes in growth, location, and transport service are the result of the accumulated short-run or daily circumstances and conditions which occur over the longer time period.

The interrelationship and distinction between short- and long-range travel forecasting can be expressed in a number of ways, one of which is shown in Figure 1. This flow-chart representation for the general equilibrium or long term forecasting problem is particularly weak in at least one respect. Even though locational shifts and land-use growth do occur incrementally from year to year (or whatever time lag seems appropriate for modeling of this sort), one should not infer that the yearly shifts or growth result simply from the present-day equilibrium flows, prices, performance levels, costs, and so forth. Rather, it seems likely that dwellers and businessmen, in shifting to new home or work locations and making modal choices (decisions which are partially interdependent with the former), are responding both to the present-day transport, land-use, and socioeconomic conditions and to those that are expected for all (foreseeable) future years. As a consequence, the time-lag type of procedure for linking the static

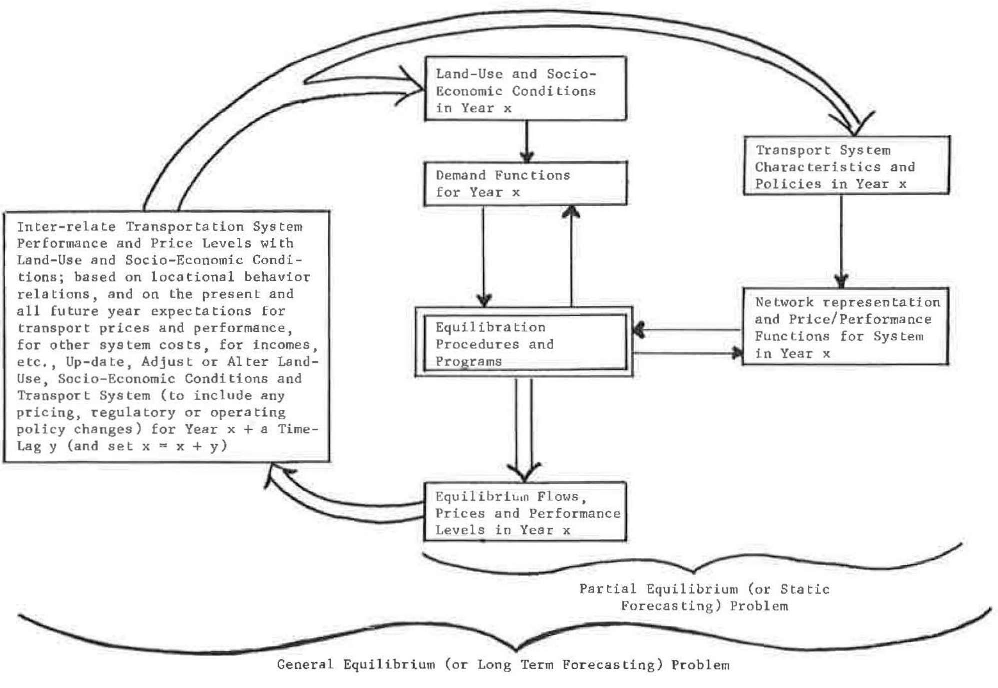


Figure 1. Schematic diagram for travel forecasting process.

and long term forecasting models is considerably more complicated than is illustrated here; the adjustments from year x to year x + 1 and some shortrun decisions rely not simply on the year x equilibrium conditions but on those for years x, x + 1, x + 2, ... N (assuming an N-year planning horizon).

THE ROLES OF DEMAND AND SUPPLY IN THE TRAVEL FORECASTING PROCESS

The distinction between demand and supply is an important one and is crucial to the formulation of an appropriate travel forecasting methodology.

Usually, demand is regarded incorrectly as the number of trips that will be made within, for example, the urban region at some future date. Demand is regarded, therefore, as the need or the requirement that must be met. This view of demand is somewhat analogous to the concept implicit in the trip-generation phase of most current forecasting models. As such, all that remains to be determined is between which zonal pairs and by what mode and route these trips will be made.

Contrarily, demand should be viewed as a statement of people's trip-making propensities; that is, it should be viewed as a demand function or conditional trip-making relationship. Thus, a demand function represents the dependence of the demanded quantity of trip-making on the price of or service afforded by trip-making. Implied, of course, is that more trips (in the absolute and relative sense) will be demanded (or generated) if either the price is reduced or the service is increased, whereas increased traffic congestion will tend to reduce trip-making. (As will be discussed in the next section, the demand function characterizes much more than trip generation; the demand function simultaneously incorporates the trip distribution, modal-split, and perhaps even route assignment phases as well.) Clearly, though, this functional and behavioral view of demand should give one pause when thinking about the common notions of "meeting the demand," or "needs," or "requirements," or constant trip-generation rates.

Another aspect of demand pertains to changes or shifts in demand. Over the short run, demand will not shift or increase; and thus changes in the amount of trip-making that occur in response to price or service changes should be regarded as movements along the demand function (or demand schedule) rather than as increases in demand. By contrast, increases in demand or shifts of the demand function will stem from long-run changes in population, income, tastes, and so forth.

The concepts of supply and demand are useful mainly because of the analogies that can be drawn from microeconomic theory, particularly in terms of specifying the interaction between supply and demand and of determining equilibrium prices and quantities demanded. Although a direct analogy can be made between the economist's and the transport analyst's characterization of travel demand and between their equilibration of supply and demand, there is only an approximate analogy for supply when applied to transport networks or links of a network.

In microeconomic theory, the term "supply" refers to the supply schedule—the amount of a product that will be supplied by the industry at different price levels. It is the amount supplied collectively by all firms producing that same product or service while assuming marginal cost pricing. In the context of this paper, "supply" is meant to characterize either the dependent relationship between travel service and the usage on the travel facility or that between travel price where the combined money and nonmoney time, effort, and expense of travel are placed on a commensurate value scale and usage. Alternatively, these expressions may be viewed as performance or service functions and are entirely analogous to capacity-restraint functions that have often been used in travel forecasting processes.

Employing the concepts of demand and performance functions to forecast trip-making at some point in time and for a given land-use plan and transport system requires that we follow a three-step process:

1. Describe trip-making behavior; i.e., specify demand functions (rather than point estimates or projections) of the form $q = f$(price, service, socioeconomic characteristics), where q is the quantity of trip-making demanded for the price, service, and other specified conditions.

2. Describe system service or performance; i.e., specify service-performance functions of the form p = f(system capacity, technology, controls, operating and price policies, volume of usage), where p is price resulting from the volume, capacity, and other specified conditions.

3. Interrelate supply and demand; i.e., equilibrate demand and performance functions for the region and transport network in question, so that point estimates of actual or equilibrium volumes and service or price levels can be determined.

In other words, we must find the values of p_x and q_x that will satisfy the following constraints:

$$q_x = f\left(\underline{p}_x, \underline{SE}_x\right)$$

$$\underline{p}_x = f\left(q_x, C_x, OP_x, PP_x\right)$$

where $\underline{p}_x$ is a vector of the price or performance conditions occurring in year x; q_x is the flow occurring in year x; C_x is the system capacity in year x; OP_x is the operating or control policy in year x; PP_x is the pricing policy in year x; and $\underline{SE}_x$ is a vector of the socioeconomic conditions in year x. The resultant $\underline{p}_x$ and q_x values are the equilibrium prices (or service levels) and flows and thus are the forecast for that system, that pricing policy, that year, etc. Figure 2 shows this interaction and the resultant or equilibrium price and volume levels.

Simplistically, the equilibrium flows and prices for a facility before and after improvement (for a one-link facility) can be as shown in Figure 3. As noted before, the induced traffic or increase in equilibrium flow from V_A to V_B that stemmed from the improvement and reduction in congestion or price should be regarded as a movement along the demand function or as an increase in the quantity of trips demanded rather than as an increase or shift in demand.

Demand, however, can and usually does shift or increase over time as a result of increases in population, income, etc., and because of changes in taste that generally affect the equilibrium flows and prices as shown in Figure 4. Thus, yearly increases in flow will stem from shifts in demand. Each yearly increase is generally slightly less than that for the previous year because of the exponential nature of queueing delays and thus the price-volume or performance curve.

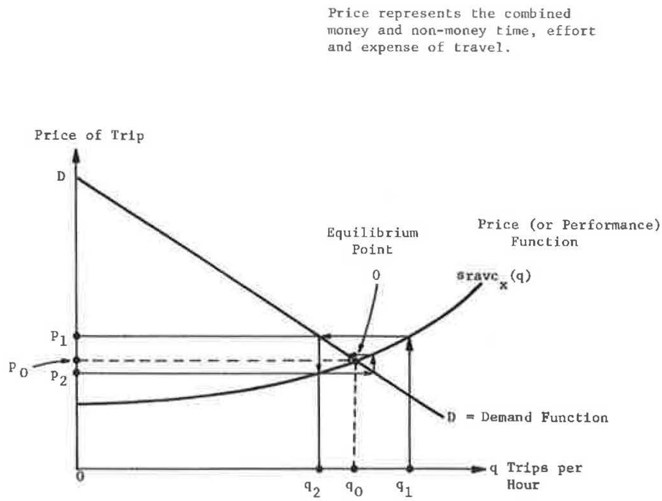


Figure 2. Simplified equilibrium relationships.

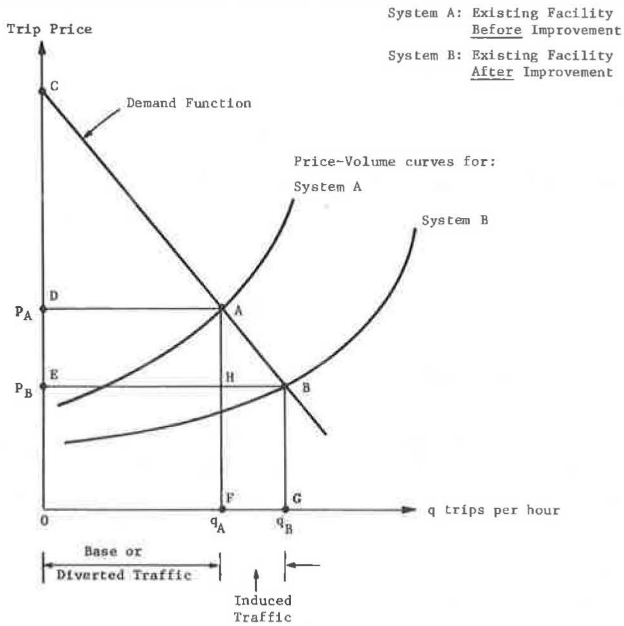


Figure 3. Equilibrium conditions for different facilities.

These simple notions about supply and demand and their interaction can be extended to explain hour-to-hour differences in trip-making and the peaking phenomenon that accompanies them. First, the demand for travel (as distinct from the equilibrium or actual flow) will fluctuate from hour to hour in response to people's preferences for traveling at specific times of day; in general, the demand for travel at starting-to-work or going-home-from-work times will be higher and less sensitive to congestion than that for other times of day. For illustrative purposes, then, demand throughout the day

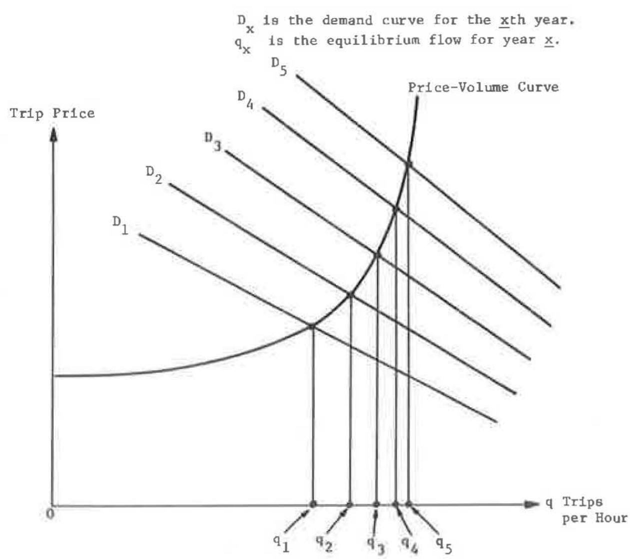


Figure 4. Intertemporal demand and price-volume relationships.

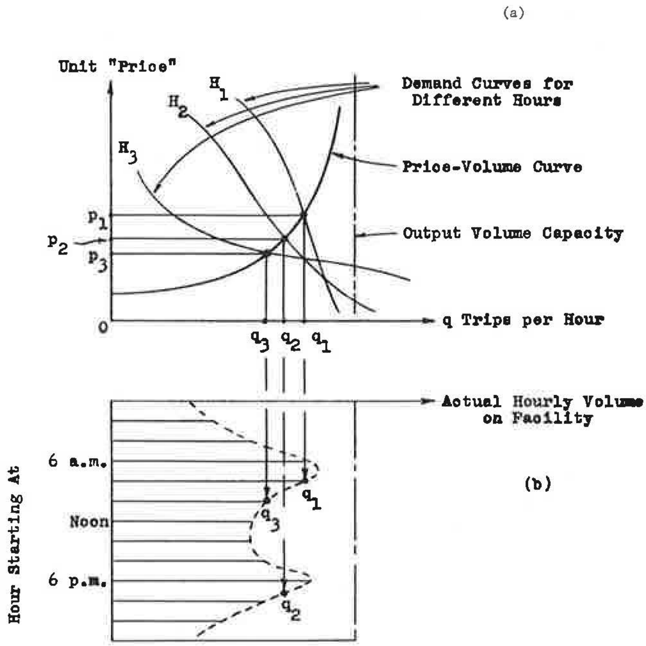


Figure 5. Short term intratemporal demand and price-volume relationships. For simplicity, demand curves for only 3 hours are shown. Also, for demand functions, the dependent and independent variable axes have been reversed. H_x is the demand for travel during hour x.

may be represented by a series of hour-by-hour demand functions as shown in Figure 5(a). (These relationships are oversimplified in some important ways which will be clarified in a later section.) Second, the interaction of the hour-by-hour demand functions with the price-volume curve determines the hour-to-hour equilibrium volumes which then may be plotted as volume versus time of day, as shown in Figure 5(b).

Considerably more discussion about the equilibration problem and about the full-scale development of appropriate demand and performance functions is given in the following section. These opening remarks are intended merely to introduce the aspects of forecasting and to indicate how demand and performance functions are related in the overall forecasting process.

DEVELOPMENT OF DEMAND FUNCTIONS

Demand When Considered in a Behavioral Context

The following points will be the basis of the model development.

First, demand for travel is a derived demand; that is, it is derived from a fundamental desire to do something else rather than travel without purpose. Thus, an understanding of travel relations must rely, to some degree, on the commodities and services being acquired at the trip destination. More simply, the value of a trip and, therefore, the extent to which it will be demanded depend on the importance of that trip to that individual. Is it a work trip? A pleasure trip? A doctor's visit? Is it important? Is the trip of little consequence and can it be foregone easily? Even work trips are given up when travel conditions are bad enough. These remarks suggest, at a minimum, that demand should be stratified by trip purpose.

Second, an individual's demand for goods and services depends on his social characteristics: family size, tastes, upbringing, income, etc. Stratification by income at least seems important.

Third, destinations differ in terms of the services and goods offered or number of opportunities; or they may differ locationally, aside from travel; or there may be just "perceived differences." We will need to specify demand, therefore, in terms of specific destination.

Fourth, in deciding whether to travel and what mode to choose, trip-makers invariably consider the circumstances for both directions of the trip. One may not go downtown by commuter railroad, for example, if he cannot come back until after the last train leaves.

Fifth, trip-makers choose modes on the basis of service and price differences and their value scales as they perceive them. Because the analyst's differentiation by service and price is not sufficient to explain trip-making behavior, we must assume that some influence variables are overlooked or improperly measured. Along this line of reasoning, mode-specific stratification should treat drive-alone car and car pooling as two separate modes.

Sixth, travelers probably view the route selection problem in a fashion somewhat analogous to choosing modes, though it is conceivable that route switching occurs along the route as events or information along the way changes one's perception. Route stratification does, however, seem in order.

Seventh, the hour of day for both ends of a trip appears to be an important consideration. The time of travel and the mode chosen are independent neither of one's time-of-day preferences nor of the travel conditions during the preferred and other-than-preferred times of day.

Characteristics of Simplified Demand Functions and Some of Their Forms

One essential characteristic of demand function is sensitivity or elasticity. For example, the "elasticity of demand with respect to price" is a dimensionless measure of the degree to which travelers respond to price changes. Specifically, the elasticity (e_p or η_p) is defined as the percent change in quantity demanded that accompanies a 1 percent change in price; or:

$$e_p \text{ or } \eta_p = \frac{\text{relative change in quantity}}{\text{relative change in price}} = \frac{\partial q/q}{\partial p/p} \text{ or } \frac{\Delta q/q}{\Delta p/p}$$

The elasticities for two forms of single-variable demand functions are shown in Figure 6. For a linear demand model the elasticity varies over its entire range whereas for a nonlinear model of the hyperbolic form the elasticity is constant.

From the designer's point of view, the measure of elasticity permits determination of the changes in toll revenues and volume and thus road capacity (or toll booths) which stem from altering the toll structure. For a transit operator, changes in the number of buses needed and gross revenues can be calculated.

The practical usefulness of knowledge about elasticities, which is virtually unused in urban transport circles, cannot be overstated. If transit fares, for example, are

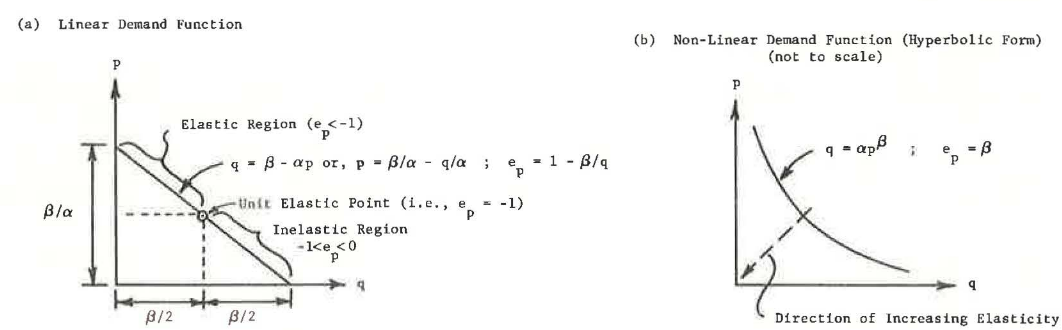


Figure 6. Single variable linear and nonlinear demand functions.

currently within the elastic portion of the demand curve, a decrease in price (without a change in service or schedule frequency) will increase ridership and gross revenues. (An increase in net revenues may or may not result, depending on the increase in costs stemming from the extra ridership.) On the other hand, if fares are presently within the inelastic region of the demand function, an increase in fare (without changing service or schedule frequency) will increase the gross and net revenues but decrease ridership. Similarly, the utility of such knowledge (that is, knowing if the demand is elastic or inelastic and to what extent) to toll authorities, railroads, airlines, etc., is all too obvious. Clearly, though, this knowledge can be exploited fully only by having similar types of information on the accompanying cost changes.

It is also important to recognize that utilization of this type of demand function, in contrast to the more usual trip-generation/trip-distribution approach, permits the analyst to assess directly the effect of price or performance changes. He is able to note increases or decreases in congestion caused by a change in technology or operating policy on the overall amount of trip-making or trip generation. Trip generation thus need not be calculated independently of or insensitive to the transport system performance characteristics and improvement. Furthermore, once these simplified single-variable demand models have been extended to incorporate multimodal and multi-time-period aspects, the ability to ascertain both the amount of trip-making by mode and by time of day and the shifts among modes and times of day can be achieved and the changes in trip-making, as well as in its modal and time-of-day distribution, that stem from changes to the transport system (or other influence variables) can be reflected.

Both of the previously mentioned linear and hyperbolic type of nonlinear demand models (Fig. 6), as well as a host of others, are in a form convenient for estimating parameter values. In both cases, linear regression techniques can be employed for estimating the parameter values (that is, for estimating α and β). For the nonlinear model, though, it is first necessary to convert the primary demand function into its log/log form. That is, given

$$q = \alpha p^{\beta} \tag{1}$$

and taking the logarithm of both sides of the equation, we get

$$\log q = \log \alpha + \beta \log p \tag{2}$$

which is a linear model (though in log form) that can be employed for estimating parameter values.

A third form that can be used for the demand model is the exponential in which

$$q = \alpha e^{\beta p} \tag{3}$$

Taking the natural log of both sides, we get

$$\ln q = \ln \alpha + \beta p \tag{4}$$

thus again providing a linear form for parameter estimation.

Extensions of Demand Models to Account for Modal Split and Peaking

First, demand models may be constructed to handle multiple service, price, or performance variables, such as enroute (or line-haul) travel time, access travel time, money price, and number of transfers. For example, if the money expense p and total trip time t were the only influence variables, the demand model might be formulated in one of the following ways:

$$q = \alpha - \beta p - \gamma t \tag{5}$$

or

$$q = \alpha p^{\beta} t^{\gamma} \tag{6}$$

Our interest then will focus on elasticity with respect to price and elasticity with respect to travel times. Quite properly, this more complete form (to include as many service variables as is appropriate) implies that the package of service-price levels is what influences trip-making. A priori, we would expect both the time and price elasticities to be nonpositive; in advance of gathering field information, however, nothing can be said about whether they fall within the elastic or inelastic region of the demand function or about whether price or time elasticities are largest. Knowledge about the elasticities in both absolute and relative terms is vital, however, because it will permit the analyst, designer, or operator to judge whether changes in service or price will beneficially affect ridership or revenues.

Second, and in a similar vein, these types of demand models may be extended to handle modal choice. One may differentiate between modes in one of two ways:

1. In terms of the technological difference, as usually is done, thus identifying bus, rail, and auto modes, for example; or
2. In terms of the service, price and performance differences, thus classifying modes only by differences in the service-price-performance package.

Should two technologies have identical service-price-performance characteristics, then travelers will be indifferent in choosing between them so long as all the price, service, and performance variables which influence the trip-making and modal choice have been incorporated. However, because travelers are not indifferent to the available modal choices but show a preference for one or another even though the price and service levels for the modes as defined and measured by the analyst are identical, then one may assume either that all the price-service-performance variables influencing behavior were not included or that the measurements of these variables were incorrect. In such a case, it will be necessary to abandon the latter type of differentiation and make use of the technological classification.

As a simple example, consider the demand functions for modes 1 and 2 with the price of each mode being the only measurable influence variable. Then, to take two alternative formulations,

1. Linear model form:

$$q_1 = \beta_1 - \alpha_1 p_1 - \gamma_1 p_2 \tag{7}$$

$$q_2 = \beta_2 - \alpha_2 p_1 - \gamma_2 p_2 \tag{8}$$

where $\alpha_1 p_1$ and $\gamma_2 p_2$ are direct demand relations, and $\alpha_2 p_1$ and $\gamma_1 p_2$ are the cross relations, which reflect the substitutability; and

2. Nonlinear model form:

$$q_1 = \alpha_1 p_1^{\beta_1} p_2^{\gamma_1} \tag{9}$$

$$q_2 = \alpha_2 p_1^{\beta_2} p_2^{\gamma_2} \tag{10}$$

For both model forms, the direct elasticities are respectively

$$e^1_{p_1} \text{ and } e^2_{p_2}, \text{ or } \frac{\partial q_1/q_1}{\partial p_1/p_1} \text{ and } \frac{\partial q_2/q_2}{\partial p_2/p_2}$$

The cross elasticities reflect the substitutability of one mode for another and are defined as the percentage change in quantity of travel demanded for one mode which accompanies a 1 percent change in the price of another mode; e.g., the cross elasticity of demand for mode 1 with respect to price of mode 2 is

$$e^1_{p_2} = \frac{\partial q_1/q_1}{\partial p_2/p_2}$$

It is important to emphasize that this type of demand model (in contrast to that implicit in the usual trip-generation/trip-distribution/model-split/route-assignment process) directly accounts for the following "real world" facts:

1. The travelers' decisions to travel or not and to select one mode or another are treated simultaneously. That is, one does not decide to travel irrespective of the alternatives afforded him and their service characteristics.
2. Both the amount of trip-making (summed over all modes) and the split among modes can and do vary with changes in travel service or price.

Third, but only to the extent that the amount or nature of trip-making is affected, it will be necessary for our demand models to incorporate the socioeconomic conditions of potential travelers who may originate trips at zone i and of the opportunities at a potential destination zone j. That is, q (ij, m) = f(transport service-price and socioeconomic variables), where q(ij, m) is the quantity of travel going from zone i to j by mode m. Just one of many possible model forms might be

$$q(ij, m) = \alpha_m (Y_i)^{\beta m} (P_i)^{\gamma m} \left(E_j\right)^{\delta m} \prod_{x=1}^{M} \left(p_{ij}^{x}\right)^{\theta m, x} \prod_{x=1}^{M} \left(t_{ij}^{x}\right)^{\varphi m, x} \qquad (11)$$

where Y_i is an income measure for zone i travelers, P_i is a population measure for zone i, E_j is an employment measure for zone j, p_{ij}^{x} is the money price for trips from i to j by mode x, t_{ij}^{x} is the travel time for trips from i to j by mode x, and M is the number of travel modes available for trips from i to j.

For this hyperbolic form, as before, the exponents are the elasticities with respect to the particular variables; e.g., β_m is the elasticity of demand for travel by mode m with respect to the zone i income measure. Similarly, $\theta_{m, x}$ (for $x \neq m$) is the cross elasticity of demand for travel from i to j by mode m with respect to the money price for travel from i to j by mode x. The exponent $\theta_{m, m}$ is the direct elasticity of demand with respect to the money price for travel from i to j by mode m. Also, to clarify,

$$\prod_{x=1}^{M} \left(p_{ij}^{x}\right)^{\theta m, x} = \left(p_{ij}^{1}\right)^{\theta m, 1} \left(p_{ij}^{2}\right)^{\theta m, 2} \cdots \left(p_{ij}^{M}\right)^{\theta m, M}$$

Fourth, the most important part of the demand analysis and travel forecasting problem concerns peaking; that is, the ability to differentiate travel by time of day and to measure the magnitude of peak loads, how long they last, and the extent of the accompanying congestion. No presently available methodology adequately copes with this aspect of travel forecasting, at least not when examined from a conceptual and behavioral point of view.

This is to suggest that the use of trip-purpose models, coupled with peak-hour factoring, is an unreliable technique for predicting peak-hour as well as peak-period travel conditions. Rather than attempt a critique of present-day methods and of their strengths and weaknesses, in the paragraphs that follow I shall attempt to discuss the aspects that conceptually, at least, should be incorporated in our demand models if peaking is to be reliably predicted.

Thus, it will be well to consider the various aspects contributing to or influencing the times of day at which people travel, as well as their modal choices (where they appear to be linked).

At the outset, one may hypothesize that three aspects are of prime importance to any discussion of peaking:

1. Trip purpose;
2. Institutional and physical system constraints, including transit scheduling and transport capacity; and
3. Time-of-day preference, both as related to and independent of trip purpose and institutional constraints.

Stratification of demand by trip purpose clearly helps to explain trip-making behavior. Work travelers, for example, generally will tolerate more congestion, higher trip prices, and more inconvenience than will shoppers simply because the work trip will provide them with more net value (in whatever terms and whether in earnings or job satisfaction). Travelers, therefore, would suffer greater net losses by foregoing a work trip than by foregoing a shopping trip. Furthermore, one would expect work travel to be less elastic than shopping travel (i.e., the percentage change in work trips caused by 1 percent change in travel time or price should be less than that for shopping trips). Both of these hypothesized conditions are shown in Figure 7 and can be extended to all other trip purposes. A rough validation of these hypotheses can be inferred from the analysis and data incorporated in a report to the U.S. Department of Transportation (1).

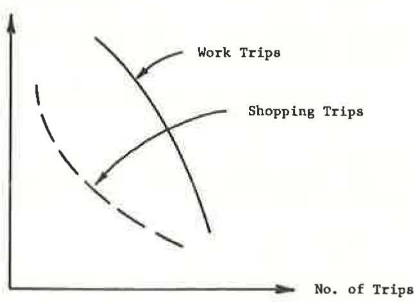


Figure 7. Hypothetical demand functions for work and shopping trips.

Institutional and physical system constraints (here broadly defined) influence trip-making behavior in two important ways. First, work and school schedules (and any attendant flexibility) and opening-closing hours for businesses, professionals, and shops all significantly affect and limit the times of day at which trips of different purposes are made. Second, both the transit schedules and the transport system capacity can and often do constrain and influence the times at which trips are made. Changes either in the hours for various activities or in the transit schedules and available capacity can lessen or increase peaking, can either reduce or increase the total amount of trip-making, and can shift trip-making among modes.

In the same fashion, as travelers make tradeoffs among mode and route choices, based on their preferences of relative and absolute travel service and price conditions, they also must make them among different time-of-day choices. For instance, workers can often choose between getting to work on time but "fighting traffic" and getting to work early (or late) but avoiding congestion. In any case, a wide range of travel times and time-of-day scheduling choices will be available to travelers and must be matched with their preferences and tradeoffs, thus affecting both the amount and extent of peaking as well as the modal choices.

Some of the more practical situations relating to the three aspects noted previously can be explained by a number of illustrations and examples. To begin, consider the effects of increasing highway capacity. Three possibilities (or some combination thereof) come to mind:

1. As more capacity is added, the same amount of daily auto trip-making can take place with the same time-of-day distribution, thus leading to a reduction in congestion, particularly during peak hours and peak periods;
2. The same amount of daily auto trip-making can take place but some trip-makers formerly traveling before or after the peak period (of some defined length) will shift into the peak period, thus changing the time-of-day distribution; in this case, congestion during off-peak periods will be reduced and that during peak periods may or may not be reduced (depending on the extent of shifts, on volume levels, and on the capacity); and
3. An increased amount of daily auto trip-making (whether from car pool to drive-alone shifts, from "induced" trips, or from modal shifts) can take place, some or all of which can occur during the peak period; also, shifts from one time-of-day period to another can occur; congestion may or may not be reduced either during peak or off-peak periods.

TABLE 1

HYPOTHETICAL TRANSIT SCHEDULE AND TRAVEL CONDITIONS FOR ZONE i TO ZONE j TRAVELERS

Conditions	Morning Bus Schedule	Total Enroute Time (min)	Arrival at Destination		Schedule Delay[a] (min)
			Expected Time	Preferred Time	
No standees	7:30	25	7:55	8:45	+50
Standees	8:00	45	8:45	8:45	0
No standees	8:30	30	9:00	8:45	-15

[a]Schedule delay is equal to the preferred minus the expected time of arrival at one's destination.

Clearly, though, none of these possibilities (the last of which is without question the most likely one) can be forecast properly without having available models that simultaneously treat the amount of trip-making, the modal split, and the time of day at which trips are made. Furthermore, it should be evident that the first and most unrealistic of these three possibilities characterizes the assumptions implicit in our present travel forecasting techniques. The major reasons are that trip generation is not made functionally dependent on the equilibrium service, price, and performance conditions that will result from a system change, and that time-of-day preferences and constraints are not incorporated in the methodology and are not related simultaneously to trip-making and modal split.

The necessity of incorporating time-of-day preferences and constraints of the sort described earlier, and their relationship both to trip-making and modal choice, can be emphasized by more concrete examples.

Transit Example No. 1—Assume that travelers potentially going from zone i to zone j (and each having identical preferred time of arrival at zone j) are faced with the bus schedule and travel conditions given in Table 1. Even if we assume that these travelers were going to travel by bus, regardless of the conditions for travel by auto, it is hardly clear which of the three buses would be preferable and to whom—at least not without having demand functions that incorporate time-of-day and service preferences. For instance, some travelers may be willing to arrive slightly later than preferred in order to avoid the possibility of standing and the extra enroute time. Others may feel quite strongly about getting to their destination "just on time," even at the expense of spending extra time enroute and standing. Another group may be particularly impatient about enroute delays and thus choose the earliest bus even though the arrival was 50 minutes earlier than preferred. In sum, one can scarcely deal with the practical real-life forecasting problems by failing to consider the full range of service differentials or of travelers' time-of-day preferences, both of which are related to the choices available to them; nor can these problems be dealt with by simply comparing enroute travel times.

Transit Example No. 2—The bus schedules and travel conditions are identical to those outlined in Transit Example No. 1 except that the high demand for an 8:45 arrival time (relative to scheduled bus capacity) causes P percent of those trying to catch the 8:00 bus to fail in getting either a seat or standing room. Those attempting to gain space on the 8:00 bus, therefore, have to consider the probability of the bus being full, resulting in an extra 30 minutes waiting time and in their being 15 minutes late. Some travelers will be willing to gamble and accept the penalties; others will find the risk too costly to accept, thus shifting either to the earlier or later bus. Again, without knowing travelers' time-of-day preferences and without accounting for all service conditions influencing trip-making behavior, it seems unlikely that we can successfully predict how much travel will take place, when it will occur, and by what mode.

Auto Expample—The following example should demonstrate that the sorts of issues and problems that arose with the transit case (and involved the tradeoffs among time-of-day preferences, service conditions, bus schedules, and capacity) also are experienced with auto travel and thus affect both peaking and modal choices.

The auto example concerns one-way traffic flow across a bridge during the morning peak period. The following assumptions are given:

1. Fifteen hundred travelers start to work in zone A at times that are spread uniformly between 8 and 9 a.m.;
2. All workers drive alone and must cross the bridge to get to work in zone B;
3. Arrivals at the bridge entrance (or end of the queue) are uniformly spaced over time period T (in hours);
4. The bridge can service arrivals at a constant rate μ of 1,000 vph; and
5. It takes exactly 15 minutes to cross the bridge (after gaining entrance) and to complete the work trip.

Under these simplified conditions, one of the following (or some combination thereof) would result:

1. All 1,500 workers would arrive at the bridge entrance uniformly between 7:45 and 8:45. The arrival rate λ would be 1,500 vph for a time period T of 1 hour, and the service rate μ would be 1,000 vph. The average queueing delay for the arrivals during time T is approximately (2)

$$\bar{t}_q \doteq \left(\frac{\lambda}{\mu} - 1\right)\frac{T}{2} \text{ for } \frac{\lambda}{\mu} > 1$$

As a consequence, virtually all workers would suffer queueing delays, and (depending on how workers ordered themselves in the queue) all could (and most would) arrive 15 minutes late to work, on the average. In this case, the average enroute travel time of 30 minutes (to include queueing delays) would automatically incorporate the average schedule delay (or preferred minus the expected time of arrival at destination).
2. The 1,500 workers will adjust to the potential queueing delays and thus will arrive uniformly at the bridge entrance between 7:15 and 8:45. In this case, the arrival rate will fall to 1,000 vph for an arrival time period T of $1\frac{1}{2}$ hours. There will be no queueing delays, and 500 of the workers will arrive at work 15 minutes early, on the average. Note, however, that the enroute travel time of 15 minutes will not include or account for the schedule delays.
3. The 1,500 workers will adjust to the potential queueing delays and thus will arrive uniformly at the bridge entrance between 7:30 and 9:00. In this case, there also will be no queueing delays, but 250 of the workers will arrive at work about 8 minutes early and 250 of the workers will arrive 8 minutes late; note again that the enroute travel time of 15 minutes will not account for these schedule delays.

In auto situations of this sort, which are typical for peak periods in many large cities, it is not clear how the traveling public will adjust. Some will prefer to arrive either early or late to avoid congestion and queueing delays; others will decide to shift to other modes of travel or to car pooling; and so forth. But without constructing demand models that simultaneously incorporate time-of-day preferences, the full range of service conditions, and modal possibilities, it seems evident that neither modal split nor the extent of peaking can be forecast appropriately.

Inadequacy of Traditional Models—To bring these points closer to reality, one might ask why the traditional modal-split and peak factoring models are unsuitable for predicting the split and peaking, both as a general case and as applied to a city like New York. First, most if not all modal-split models and peak factoring techniques make use only of enroute travel times (including waiting times, queueing delays, and transfer times) and thus ignore the inconvenience (termed herein as "schedule delay") that results from arriving early or late to avoid or reduce congestion. Second, present models and forecasting techniques do not account for the way in which modal split and peaking are related and the extent to which they are affected by the strengths of travelers' time-of-day and service tradeoff preferences.

Turning to the first of these points, it should be evident that modal choice is not based simply on the enroute travel conditions of the alternatives and certainly does not discount the enroute travel conditions during alternative times of day. For example, why do most downtown New York workers who commute by auto travel to work during peak periods and endure extremely high enroute travel times when they could be significantly reduced if they would only travel either before or after the rush period? The answer depends jointly on the knowledge that the commuting time could be significantly reduced only by starting to work very early or very late (relative to work starting times) and on their distaste for these other time-of-day alternatives (i.e., on a strong preference for leaving home no sooner than necessary and for arriving at work no sooner or later than necessary). On the other hand, if the peak period were not so lengthy, some peak travelers would shift to off-peak hours, the extent of switching depending of course on their time-of-day preferences and on the reduction in enroute travel time versus increase in schedule delay.

In a similar vein, one can begin to understand both modal split and peaking as well as their interrelationship. Clearly, as the level of congestion and as the length of the peak period for auto travel increases, shifts from auto to transit and from drive-alone auto to car pool will occur. Specifically, as more intense and longer peak periods occur, the amount of schedule delay generally will increase because the working, shopping, and business hours appear to remain virtually the same. Increased schedule delay, coupled with travelers' time-of-day preferences, will lead some people to shift to those facilities capable of handling higher peak loads, however uncomfortable or inconvenient they may be. In New York City, for example, it is extremely doubtful that the existing modal split (much less the level and extent of peaking) can be explained satisfactorily by making modal-split curves that employ the usual enroute travel times (to include allowances for waiting and transfers), money expenses, and income differentials. For example, transit riders in New York are probably aware that a shift to auto, in addition to entailing an arduous and lengthy trip that would permit avoidance of the subway "crush", would probably be accompanied by an early or late arrival at work.

Furthermore, it is of considerable importance to note that the data used to compare the travel times, to compute travel time ratios, and so forth, are usually incorrect. The modal-split percentages, which are based on empirical data and are incorporated in the curves for predicting future splits, often are computed on the basis of one set of data and then applied while making use of a different set. For example, empirically based modal-split percentages by trip purpose often have been calculated for travel time ratios that are based on the actual origin-destination (O-D) travel times of the travelers having that trip purpose. When this model is applied to future trip-making, different travel times are used for computing the ratios. More specifically, suppose that the empirical modal-split percentages for work trips were based on actual travel time data for work trips. The travel times then would be heavily concentrated during the peak hours, in the order of 65 to 75 percent of the total daily work trips occurring during the 4 peak hours. Given this basis for the modal-split model, it would be incorrect to use (as is often the case) off-peak or average daily travel time data or to use other travel time data for a different time period to calculate future modal splits.

Along similar lines, other inaccuracies arise because the O-D modal travel time data used in determining modal-split curves often are derived from different time periods and then are applied to still different ones. Transit work trips, for example, are usually more peaked than auto work trips. The degree of peaking depends on the extent and duration of highway congestion and on the availability of transit capacity. The O-D travel times for auto work trips thus are spread over a longer peak period than are those for transit work trips, and we can be assured that modal splits are being computed either for people having different working hours or for those having different amounts of schedule delay.

This aspect becomes of extreme importance in those situations having high and lengthy traffic congestion, particularly when considerable transit capacity is available. In Washington, D.C., for example, where transit capacity is somewhat limited (and in much the same way as auto travel is limited by congestion and street capacity), the percentage of daily transit work trips arriving at work during the morning peak hour is roughly 22 percent as compared to about 18 percent for daily auto driver work trips during the same hour (3). (Based on 1955 survey data, these results apply to work trips for the entire region rather than solely for the downtown sector; if similar percentages were available for downtown work trips, the transit percentage would probably be slightly higher and the auto percentage somewhat lower.)

In downtown New York City, however, where congestion periods are extremely intense and lengthy and which is served by very high peak-period transit capacity, about 16 percent of the daily work trips leave work during the peak 10 minutes and about 31 percent depart during the peak hour (4). (For the morning, the corresponding percentages are 10 and 31.) Although these figures represent the combined peaking for auto and transit work travel, they mainly reflect the peaking patterns for transit travel which account for almost 95 percent of the downtown work trips.

A Workable Demand Model—To formulate a workable demand model capable of incorporating the most significant of the modal-split and peaking aspects is, of course, no mean task. Moreover, the demand models, to be fully operational and meaningful, should not be formulated without considering the related problems of formulating consistent and compatible price and performance functions and of equilibrating the two sets of functions. Even so, before discussing these latter two aspects, it will be useful to propose two forms of demand models that treat peak and off-peak conditions. The first does so in a highly simplistic way; and the second, in a more satisfying and complete way (conceptually, at least).

Simplified Peak/Off-Peak Demand Model—The peak-period demand model has the form

$$q_p = \alpha_p + \beta_p t_p + \gamma_p t_o \tag{12}$$

and the off-peak-period demand model may be represented as

$$q_o = \alpha_o + \beta_o t_p + \gamma_o t_o \tag{13}$$

where q_p is the hourly volume of trips demanded during the peak period, q_o is that demanded during the off-peak period, t_p is the peak-period travel time, and t_o is the off-peak-period travel time. We would expect the parameters α_p, α_o, γ_p, and β_o to be nonnegative and β_p and γ_o to be nonpositive.

This model expresses some simple though important and logical relations. First, as congestion (i.e., travel time) during the peak period increases (while that during the off-peak period remains unchanged), some peak-period trip-making will be discouraged; some people will cancel trips altogether and others will shift to off-peak hours. If travel conditions during the peak-period are improved (but those during off-peak hours are unchanged), the peak-period flow will be increased and that during the off-peak period will be reduced. Second, both the total amount of daily flow and the split of the flow among peak and off-peak hours can change in response to changes in travel conditions during either or both of the time periods.

Composite Multimode and Time-of-Day Demand Model—Among the many ways of specifying significant influence variables, model forms, demand relations, and cross relations, the following general formulation seems sufficiently complete and logical to serve as a point of departure for further exploration and study. (In form, this demand model is not unlike the intercity and multimode passenger demand model which was developed for the Northeast Corridor by Gerald Kraft (5).)

$$q_{ij}^{m,t} = \alpha_m (Y_i)^{\beta_m} (P_i)^{\gamma_m} \left(E_{j,t}\right)^{\delta_{m,t}} \prod_{\forall x,y} \left(c_{ij}^{x,y}\right)^{\theta_{m,t,x,y}} \prod_{\forall x,y} \left(f_{ij}^{x,y}\right)^{\varphi_{m,t,x,y}} \tag{14}$$

where

$q_{ij}^{m,t}$ = quantity of trip-making between zones i and j by mode m during time period t;

Y_i = income measure for zone i residents;

P_i = population measure for zone i residents;

$E_{j,t}$ = employment measure for zone j during time period t;

$\prod_{\forall x,y}$ = the product of terms for all values of x and y ranging from 1 to M and T respectively; the expression therefore represents the product of M·T terms (of course, some elasticity values can be zero, thus reducing the number of terms);

$c_{ij}^{x,y}$ = congestion measure for travel between zones i and j by mode x during time period y;

$f_{ij}^{x,y}$ = fare or money cost measure for travel between zones i and j by mode x during time period y; and

$\beta_m, \gamma_m, \delta_{m,t}$ = the demand elasticities for mode m (or mode m and time period t) travel with respect to the income, population, and employment measures respectively. (Some of these measures will be stated in absolute terms and others in relative terms, though for this discussion it will not be necessary to be more specific.) Finally,

$\theta_{m,t,x,y}$ = elasticity of demand for mode m during time period t with respect to congestion on mode x during time period y

$$= \frac{\partial q_{ij}^{m,t} / q_{ij}^{m,t}}{\partial c_{ij}^{x,y} / c_{ij}^{x,y}}; \text{ and}$$

$\varphi_{m,t,x,y}$ = elasticity of demand for mode m during time period t with respect to fare or money cost on mode x during time period y.

The two elasticities, as x and y vary from 1 to M and T respectively, will represent the cross elasticities (i.e., they will reflect the percent change in quantity of travel demanded for one mode and time period with respect to the percent change in congestion or cost of another mode and time period). When x and y are equal to m and t respectively, however, the elasticities then will represent the direct demand elasticities. We would expect the direct elasticities to be nonpositive and the cross elasticities to be nonnegative. It is likely that empirical analysis will show that some if not many of the cross elasticity values will be zero, thus reducing materially the number of terms in the individual demand functions and the complexities to be confronted in equilibrating demand and performance functions for transport networks. For example, suppose the 24-hour day could be suitably represented by five time periods (i.e., time period 1, 7 to 9 a.m.; time period 2, 9 a.m. to 3 p.m.; time period 3, 3 to 6 p.m.; time period 4, 6 to 9 p.m.; and time period 5, 9 p.m. to 7 a.m.). For such a breakdown, it can be argued that the demand for travel during period 1, for instance, would be particularly sensitive to travel conditions during that time period and somewhat sensitive to those during periods 2 and 5. The same demand would be practically insensitive to the travel conditions during time periods 3 and 4. The cross elasticities for demand during time period 1 with respect to travel conditions during time periods 3 and 4, therefore, will be zero (or at least will be small enough to be ignored).

Equation 14 gives the demand for only one mode and time period combination out of M·T possible combinations. Thus, M·T demand functions will be required to fully specify the demand for each zonal or ij pair. Clearly, then, in situations where many modes are available the number of combinations and demand functions necessary to explain trip-making can become cumbersome, particularly when many time periods are required to reasonably explain people's time-of-day tradeoffs and preferences. Without a considerable amount of data analysis and parameter estimation it is difficult to even guess how finely the modes, submodes, and time-of-day periods should be stratified. For cities having these alternatives available, at least six modes probably should be specified: bus transit, rail transit, commuter railroad, taxi, drive-alone auto, and car pool. It seems that there are significant service and/or price differentials among these choices (that is, significant from the standpoint of influencing the amount of trip-making, the time-of-day in which trips are made, or the modal choice), and that by aggregating modes in the usual fashion (i.e., all auto versus all transit) the differentials are made much less sharp, thus obfuscating the modal-choice question and the ability to differentiate between modes and to predict future choices.

Probably the worst aspect of this type of aggregation involves lumping drive-alone auto and car pool trips together in a single auto mode. Drive-alone auto travel has service and price characteristics that are distinctly different from both transit modes and car pool travel. Car pool travel, however, is not unlike transit travel with respect to service and price. Also, although drive-alone has service features that are all clearly superior to those for transit, car pool has some important service features that are far worse than those for transit. For instance, car poolers are restricted to a single work-trip time schedule and to a single O-D pair, whereas transit riders can take earlier or later buses to and from work and can stop off at intermediate zones or change their final destination. Aggregating the two auto submodes thus produces an average auto trip that is difficult to differentiate from a transit trip.

The problem of specifying the different modal possibilities and of defining how finely they must be stratified for forecasting purposes does not end here. Different people choose to use different modes for different parts of their door-to-door trip. For example, when comparing auto to rail transit, the service and price differentials would depend in part on whether the people traveled to the rail transit station by foot, by feeder bus, by "kiss-and-ride" auto, or by "park-and-ride" auto. Modes should be defined, therefore, by the overall modal combination for the door-to-door trip. Extending the modal-choice definition in this fashion, however, can easily triple and perhaps quadruple the number of transit modes and double the number of demand functions required. Consideration of these practical sorts of problems is hardly trivial, and careful data analysis will be required to reach firm conclusions about which modal combinations are significant and worth inclusion.

Specifying the number of time-of-day periods necessary to accurately portray the time-of-day volume variation pattern is no less difficult and certainly no less important than delineating the modal breakdown. If too few time periods are specified, it is likely that substantial inaccuracies will occur in predicting travel volumes and travel times during different times of day. This will result because the aggregated data used for estimating parameter values will mask and shift the peaks. On the other hand, if many time-of-day periods are specified, the number of demand functions will be multiplied enormously, making the task of equilibrating demand and performance functions for multimodal transport networks virtually impossible (from a computational standpoint); furthermore, the data requirements for parameter estimation would be enormous if

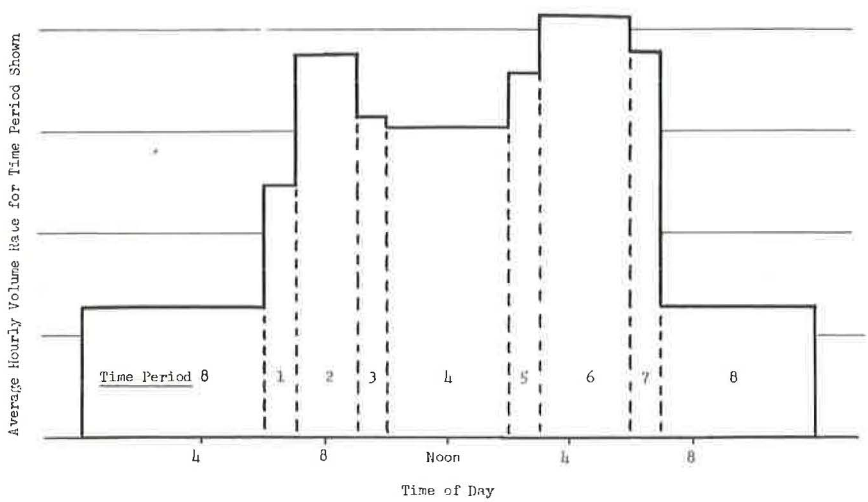


Figure 8. A suggested breakdown for time-of-day demand periods.

not out of reach for the data that presently are available. Ideally, at least eight time-of-day periods would be incorporated as shown in Figure 8. A time period on each side of the morning and afternoon major rush periods would be necessary to identify and account for shifting peak situations in which some workers (or perhaps rush-period shoppers) would go to work either early or late because of congestion during the more preferable time-of-day period. Analysis of empirical data will be necessary, however, to establish the necessity of different numbers and lengths of time-of-day periods.

Finally, it should be pointed out that to properly identify and measure time-of-day elasticities and cross elasticities, it probably will be necessary to stratify demand by trip purpose, as well as by mode and by time-of-day period. Although more clarity and accuracy will be provided by this additional stratification, the tasks of parameter estimation and of network equilibration also will be compounded considerably. No firm statements can be made about the fineness to which trip purposes should be defined; but at a minimum, two purposes—work trips and nonwork trips—should be used.

DEVELOPMENT OF PRICE-SERVICE-PERFORMANCE FUNCTIONS

The essence of the problem is to develop functions that will express the price, service, and performance conditions as a function of the facility design, vehicle technology, volume and character of usage (percent of tructs, etc.), operating and pricing policies, and so forth. Our concern is with the representation of performance and prices as viewed by the traveler.

To approach this problem, the performance function may be characterized in one of two ways:

1. Use a vector of service and price characteristics (such as time enroute, waiting time, schedule delay, and out-of-pocket money payments), or
2. Use a single price or performance variable that represents the cumulative value of money and nonmoney service and price components. In this case, it is necessary to establish commensurate values for the various components of service and price.

Whichever type of performance function is adopted, single variable or multivariate, it must be consistent and compatible with the demand function. That is, if $q = f(p,$ socioeconomic conditions, etc.), where p represents the combined money and nonmoney time, effort, and expense of travel, only a single performance function is needed or, for example, $p = f(q, C_X, \ldots)$, where C_X is a capacity measure for facility type x. However, if $q = f(p, t_1, t_2, \ldots,$ socioeconomic conditions, etc.), where p represents only the money expenses, t_1 is the access travel time, and t_2 is the line-haul travel time, then the following set of performance functions is required:

$$p = f(C_X, q, \ldots)$$

$$t_1 = f(C_X, q, \ldots)$$

$$t_2 = f(C_X, q, \ldots)$$

$$\vdots$$

Characteristics of Performance Functions

The essential aspects and characteristics of performance functions can be illustrated by using the simpler single-variable performance function rather than the more complete function involving both service and price variables. For this discussion, then, assume that

$$p = f\left(q, C_X, P_y\right) \tag{15}$$

where C_X is a capacity measure for facility x, and P_y represents the yth pricing policy in use for the facility. For this single-variable model, the performance or price p

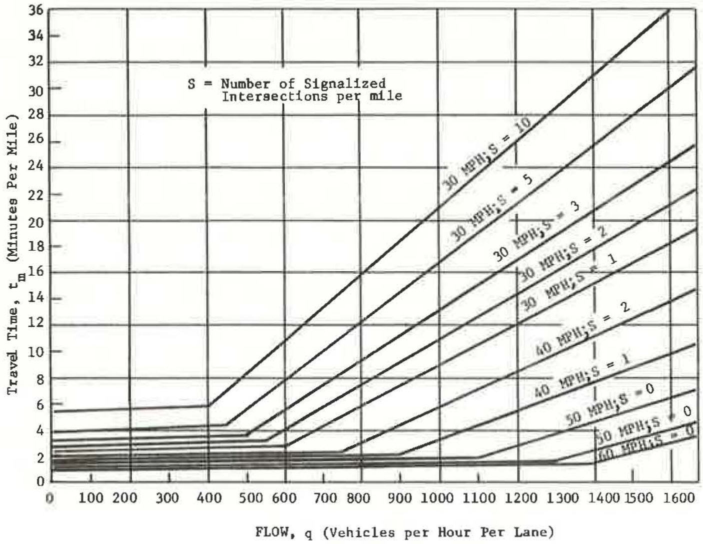


Figure 9. Capacity-restraint function curves. (From N. A. Irwin, Norman Dodd, and H. G. Von Cube. Capacity Restraint in Assignment Programs. HRB Bull. 297, 1961, pp. 109-127.)

represents the combined value (in dollars or other commensurate terms) of the money payments, time, effort, and hazards of travel, but only to the extent perceived by travelers. All travelers, however, are assumed to be homogeneous with respect to the values of the service and performance components. (This assumption is adopted merely for the purposes of illustration rather than for its realism.)

A rather typical example of a set of travel time versus volume (or so-called capacity restraint) curves is shown in Figure 9; these would be applicable when travel time is the only significant price or service variable affecting demand. The representation in Figure 9 is incomplete in two respects:

1. It fails to apply certain capacity-reducing or bottleneck types of facilities; and
2. It fails to relate travel time delay to the time interval or period over which the volume rate is sustained.

Though somewhat tangential, these two points are important enough to be clarified.

In capacity-reducing type of facilities, the service rate (or capacity) can be reduced by the shock waves produced when the traffic volume reaches a critical level. For such situations—as occurs at uncontrolled intersections and expressways, or at uncontrolled merging points—flow and the resultant performance is unstable where demand is high. Figure 10 depicts performance functions for both capacity-reducing and non-capacity-reducing types of facilities. At present, though, capacity-restraint functions used in traffic assignment have failed to represent the more complex and dynamic performance function for the former type of facility.

The second aspect—the time interval or period over which volume rate is sustained—arises partially because of a failure to differentiate between steady-state and non-steady-state queueing situations. (The effects of queues existing at the start of time periods are discussed in a later section.) To deal with the latter, the following functional form is necessary:

$$p = f\left(q,\ C_x,\ P_y,\ T\right) \tag{16}$$

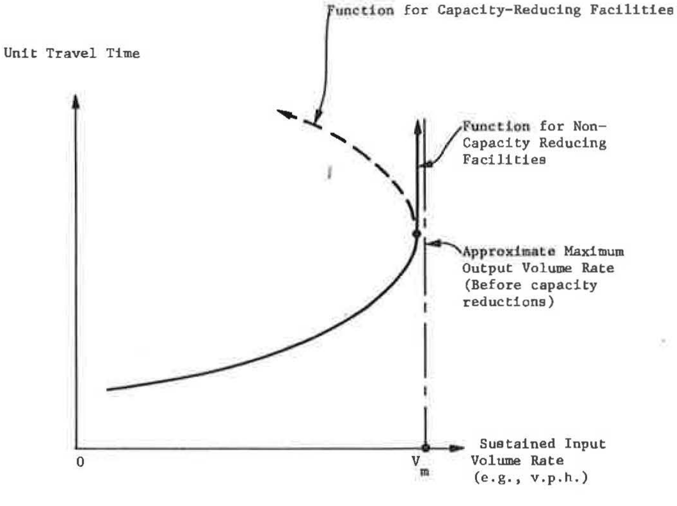


Figure 10. Approximate (non-time-dependent) relationships for travel time (or cost) versus flow for different types of facilities.

where q and C_X are stated in flow rates (e.g., vph) and T is the time interval or period over which flow rate q is sustained. In contrast, steady-state queueing, and most capacity-restraint functions, assume that λ (the arrival rate) is constant for all time and that $\lambda < \mu$ (where μ is the service rate). (The switch in notation from q to λ and from C_X to μ was made so that the terms would correspond to those common to the queueing theory literature. To avoid confusion, however, it should be emphasized that there is a one-to-one correspondence between q and λ and between C_X and μ; a differentiation and distinction is made simply because the transportation planner is more familiar with one set of definitions and notations and the queueing theorist with another.)

Figures 10 and 11 show the usual steady-state queuing model. Clearly, though, λ often does exceed μ for 1- to 2-hour rush periods (sometimes more, sometimes less) in downtown areas and on radials, thus building long queues which are worked off during later time periods when the arrival rate declines. Delays for these peak periods are not infinitely large, as implied by steady-state queueing relationships.

Thus, we need transient queueing functions, particularly ones for dealing with the exploding queue case (i.e., with the $\lambda > \mu$ case). Take a simple example in which

1. Intersection capacity or service rate = μ = 1,000 vph,
2. Arrival rate = λ = 2,000 vph, and
3. Service is uniform or constant and arrivals are equally spaced.

Case 1: Let the arrival rate of 2,000 vph be sustained for a time period of ½ hour, and let there be no queue at the beginning of the time period. Then, it will take 1 full hour to clear all those arriving during the ½-hour period, and their average wait will be ¼ hour to clear the facility.

Case 2: Let the arrival rate of 2,000 vph be sustained for a time period of 1 full hour, and let there be no queue at the beginning. Then, it will take 2 full hours to clear all

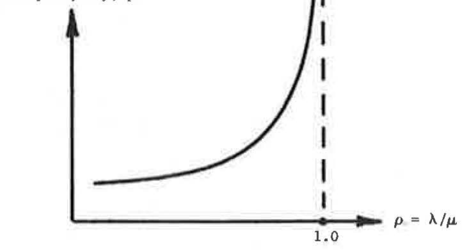


Figure 11. Generalized queueing curve for steady-state conditions.

$\bar{t} = t_o + \bar{t}_q$ where t_o is service time for $\lambda < \mu$

$\bar{t}_q$ = Time-In-Queue $\doteq (\frac{\lambda}{\mu} - 1)\frac{T}{2}$ for $\lambda/\mu > 1$

μ = 600 vehicles per hour

T = Time period over which λ is sustained; also, initial queue is zero

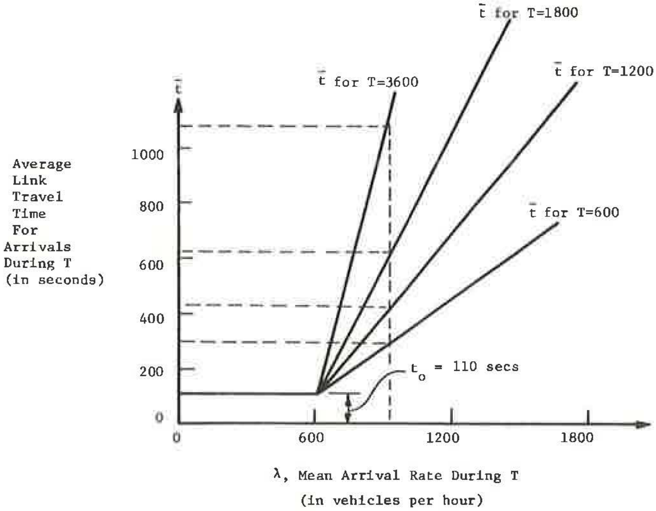


Figure 12. Time-dependent travel-time-versus-volume relationships for constant service and uniform arrival case.

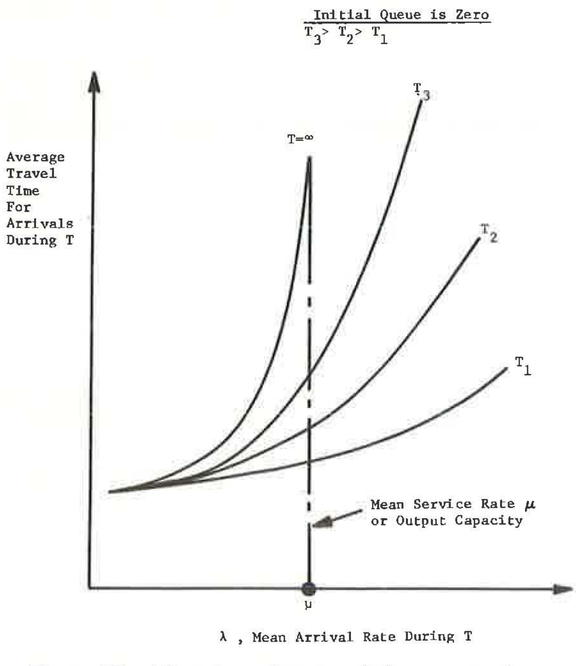


Figure 13. Time-dependent travel-time-versus-volume relationships for random arrivals and service.

those arriving during the 1-hour period, and their average wait till clearance will be ½ hour.

In short, even though λ was greater than μ (i.e., $\rho > 1$), delay was not infinite; and, quite importantly, delay obviously is a function of the time period over which the arrival rate is sustained.

Figure 12 shows these time-dependent, non-steady-state queueing relationships for the uniform service and uniform or equally spaced arrival case, and Figure 13 shows them for the random arrival and service case.

This discussion and the examples emphasize that travel time and price functions (for each link) must incorporate the time dependency and must match and be compatible with the length or time period of the corresponding demand interval. For example, if the demand function for the ith time-of-day period covers 2 hours, then the travel time versus delay or capacity-restraint function must represent the travel conditions that occur for volume rates sustained over 2 hours. Of equal importance, it should be recognized that the traffic engineer, when gathering field data to establish capacity-restraint functions, must not gather data for different time periods and then factor the data to hourly rates and incorporate them into a single travel-time-versus-hourly-volume rate function. Note also that the arrival volumes and travel times of interest here are those for vehicles arriving at the upstream side of the intersection or facility during the time period in question, regardless of whether they clear the intersection or facility during that or a later time period. Also, the engineer when recording the field data should include information on the length of the queue in existence at the start of the time period.

Pricing Policy as a Determinant of the "Price" or Performance Function

It was suggested earlier that the price perceived by users was made of certain money and nonmoney payments which reflected their money, hazard, time, and discomfort "expenses." In a rough sense, one might assume that the price function now in existence on public roads and streets is equivalent to the short-run average variable cost function. Essentially, this implies the following:

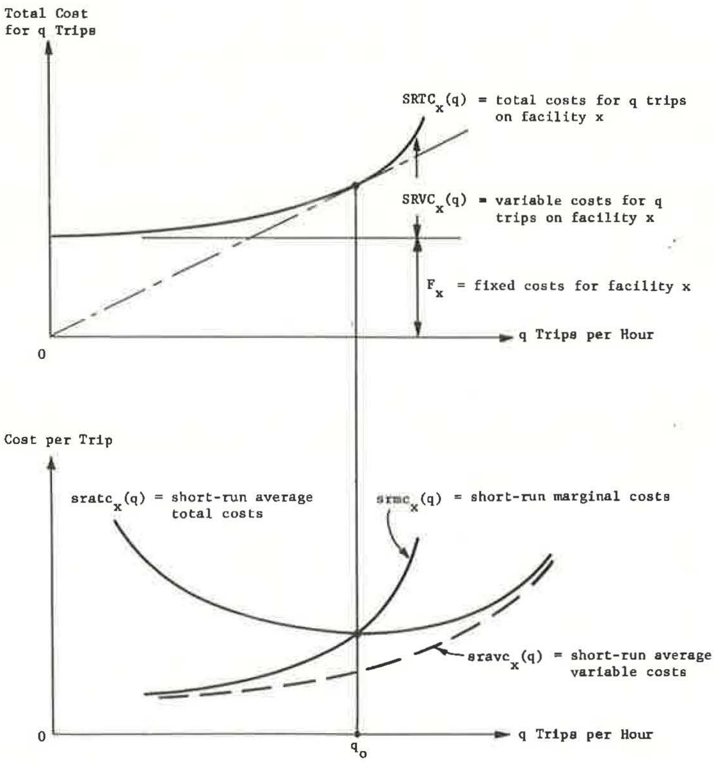
1. Perceived vehicle operating pavements ≡ variable vehicle costs;
2. Perceived parking fee payments ≡ variable parking costs;
3. Perceived user gas tax payments ≡ variable highway costs; and
4. Perceived time, effort, hazard and discomfort expenditures or payment ≡ variable time, effort, hazard, and discomfort costs.

For this discussion, short-run cost functions are those applying to time periods that are too short to alter the capacity of transport systems or links; thus the facility is fixed and neither the capacity nor the capital investment (as well as the overall travel costs for a given volume q) can be altered. In the long run, however, the facility capacity and cost relationships can be altered (both upward and downward), and thus can change the short-run cost functions. The distinction between fixed and variable costs also is important. For a particular facility, fixed costs are those that do not vary with changes in usage (i.e., with changes in q) over the short run, whereas variable costs are those that vary or change with changes in usage. Alternatively, the fixed costs may be viewed as those costs that are nonseparable with respect to nonzero volume levels and thus are common to all units of the volume using the facility.

For travel on public highways, for the existing user gas tax type of highway pricing, and for these four assumptions, the short-run average variable cost function (the $sravc_x(q)$ curve shown in Figure 14) would represent appropriately the price or performance function to be used in equilibration. Using this function presumes that travelers either do not pay or do not perceive any portion of the fixed vehicle or highway costs. Also, the curves shown in Figure 14 only apply to controlled highways and do not represent the relations applicable to uncontrolled roads and streets on which shock-wave action and backward-bending-delay-versus-volume situations often occur.

Similarly, one might argue that a comparable situation presently exists for most transit facilities and that the short-run average variable cost function can serve as an appropriate price or performance function for transit trip-making. To accept this assumption would imply that the transit fare just covers and equals the variable costs for operating and maintaining transit vehicles, trackage, stations, and maintenance facilities and that the transit system's fixed costs are covered through other revenue sources.

For other kinds of pricing policies, cost functions different from the short-run average variable cost curve would have to be used. For example, to represent the overall money and nonmoney price for highway toll facilities or for transit systems in which the toll or fare covers both the fixed and variable costs and in which the toll or fare is uniform throughout the day (that is, the fixed costs are distributed evenly among all daily users), use of the short-run average total cost function, as described by the $sratc_x(q)$ curve in Figure 14, might be appropriate. Should a peak-load pricing policy based on marginal costs be employed, then the marginal cost curve as shown by $srmc_x(q)$ would apply.



$$\mathrm{sratc}_x(q) = \mathrm{SRTC}_x(q) \div q$$

$$\mathrm{sravc}_x(q) = \mathrm{SRVC}_x(q) \div q$$

$$\mathrm{srmc}_x(q) = \frac{\partial \mathrm{SRTC}_x(q)}{\partial q} = \frac{\partial \mathrm{SRVC}_x(q)}{\partial q}$$

Figure 14. Basic short-run cost functions and relationships for a given transport facility.

The way in which these different cost and price functions can vary with facility size is shown in Figure 15.

EQUILIBRATION OF DEMAND AND PERFORMANCE FUNCTIONS

Once the demand and performance functions have been formulated, and their parameter values estimated, various techniques can be used to equilibrate them for given transport networks and pricing policies. In notational form, the task is to find the equilibrium volumes and prices (i.e., and p_i and q_i values for all n time-of-day periods) that will satisfy the following set of demand and performance or price functions:

$$q_i = f(p_1, \ldots, p_n; \text{ socioeconomic variables}) \tag{17}$$

and, for $i = 1, \ldots, n$,

$$p_i = f\left(q_i, C_x, P_y, T_i\right) \tag{18}$$

The effect of queues in existence at the start of time periods on performance or price is discussed in the next section and also by Kraft and Wohl (6).

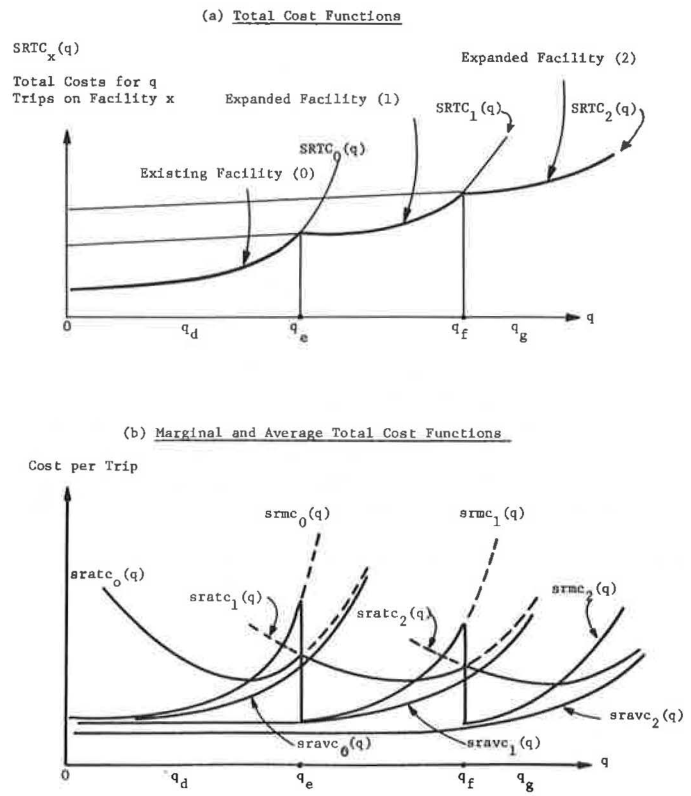


Figure 15. Short-run cost functions for three facility sizes.

It will be helpful to illustrate the interaction between these functions under different demand conditions and pricing policies; to do so, some oversimple demand and performance functions and a simple one-link unimodal transport network will be employed.

Example 1: Average Variable Cost Pricing Policy and Linear Demand

Let us consider two cases: one in which we assume that there are no hour-to-hour demand cross elasticities and that demand is linear (this last assumption is made to simplify the example); thus, for $i = 1, \ldots, n$ time-of-day periods,

$$q_i = \beta_i - \alpha_i p_i \tag{19}$$

and the second in which we assume there are hour-to-hour demand cross elasticities; thus, for $i = 1, \ldots, n$,

$$q_i = \beta_i - \alpha_{i,\,1}p_1 - \ldots - \alpha_{i,\,n}p_n \tag{20}$$

For both equations, q_i is the quantity of trips demanded during the ith time-of-day period but expressed in trips per hour (averaged over the time period), and p_i is the average price experienced during the ith time period.

For both demand cases, a time-dependent price or performance function will be used to represent the short-run average variable cost relationship. For simplicity, let us employ the following relationship. For $q_i > C_x$,

$$p_i = p_{min} + \gamma_i\left(\frac{q_i}{C_x} - 1\right)\frac{T_i}{2} \tag{21}$$

and, for $q_i \leq C_x$,

$$p_i = p_{min} \tag{22}$$

in which p_{min} is the (constant) minimum overall trip price for unsaturated flow conditions, γ_i is the unit value of travel time and congestion during the ith time period, C_x is a capacity measure for the facility (expressed as an hourly rate), and T_i is the number of hours in the ith time-of-day period.

This clearly is an oversimple model, first because the effects of stochasticity are not incorporated, second because it suggests that the unit value of travel time and congestion is constant for all levels of travel time and congestion (i.e., an extra minute is worth the same to trips of 5 minutes as it is to those of 50 minutes), and third because it applies only to non-backward-bending flow situations. The concern at this stage of the discussion, however, is with the equilibration process and the interactions between demand and performance functions rather than with the validity of the particular models and functions employed.

Case 1: Without Hour-to-Hour Demand Cross Elasticities—For this case, congestion during the ith time-of-day period does not affect the extent of trip-making, the congestion, or the price during any other period; thus, there are no shifts of travel from one period to another. Direct solution therefore is possible, either analytically or graphically. For the former, after inverting the demand functions and equating to p_i (one unknown in each equation), the following equations (each having one unknown) are derived, and the price equation having the highest value and thus providing the smallest value of q is selected:

$$\frac{\beta_1}{\alpha_1} - \frac{q_1}{\alpha_1} = p_{min} \text{ or}$$
$$= p_{min} + \gamma_1\left(\frac{q_1}{C_x} - 1\right)\frac{T_1}{2}$$
$$\vdots$$
$$\frac{\beta_i}{\alpha_i} - \frac{q_i}{\alpha_i} = p_{min} \text{ or}$$
$$= p_{min} + \gamma_i\left(\frac{q_i}{C_x} - 1\right)\frac{T_i}{2}$$
$$\vdots$$
$$\frac{\beta_n}{\alpha_n} - \frac{q_n}{\alpha_n} = p_{min} \text{ or}$$
$$= p_{min} + \gamma_n\left(\frac{q_n}{C_x} - 1\right)\frac{T_n}{2}$$

The values of $q_1, \ldots, q_n$ can be determined directly; and those of $p_1, \ldots, p_n$, by substitution into Equation 21 where p_{min} does not apply.

Case 2: With Hour-to-Hour Demand Cross Elasticities—For this case, congestion and thus price during the ith demand period will affect flow during other periods or times of day; i.e., people are shifting from hour to hour, depending on the alternatives and their preferences, and thus the hour-to-hour and total amount of daily trip-making is affected. Stated functionally, though in linear form, the set of demand and price functions might be somewhat as follows:

1. n demand equations:

$$q_1 = \beta_1 - \alpha_{1,1}p_1 - \dots - \alpha_{1,n}p_n$$

$$\vdots$$

$$q_n = \beta_n - \alpha_{n,1}p_1 - \dots - \alpha_{n,n}p_n$$

2. n price equations (select the largest of each pair):

$$p_1 = p_{min} \text{ or}$$

$$= p_{min} + \gamma_1\left(\frac{q_1}{C_x} - 1\right)\frac{T_1}{2}$$

$$\vdots$$

$$p_n = p_{min} \text{ or}$$

$$= p_{min} + \gamma_n\left(\frac{q_n}{C_x} - 1\right)\frac{T_n}{2}$$

Alternatively, the n price equations can be stated as a series of inequalities as follows:

$$p_1 \geq p_{min} \text{ and } p_1 \geq p_{min} + \gamma_1\left(\frac{q_1}{C_x} - 1\right)\frac{T_1}{2}$$

$$\vdots$$

$$p_n \geq p_{min} \text{ and } p_n \geq p_{min} + \gamma_n\left(\frac{q_n}{C_x} - 1\right)\frac{T_n}{2}$$

The n demand functions and n price functions cannot be solved directly for the equilibrium q_i and p_i values because of the interdependencies that stem from the time-period-to-time-period cross elasticities. Thus, iterative numerical or programming techniques must be employed for their solution. For this simplified example, linear programming can serve as one practical technique for solving the equations simultaneously. A suitable linear programming format for accomplishing this is the following.

Determine the $q_1, \dots, q_n$ and $p_1, \dots, p_n$ values that will

$$\text{Maximize } Z = \sum_{i=1}^{n} k_i q_i$$

as subject to nonnegativity restrictions (i.e., all q_i and p_i values must be nonnegative) and to the following constraints:

$$q_1 = \beta_1 - \alpha_{1,1}p_1 - \dots - \alpha_{1,n}p_n$$

$$\vdots$$

$$q_n = \beta_n - \alpha_{n,1}p_1 - \dots - \alpha_{n,n}p_n$$

and

$$p_1 \geq p_{min}$$

$$p_1 \geq p_{min} + \gamma_1\left(\frac{q_1}{C_x} - 1\right)\frac{T_1}{2}$$

$$\vdots$$

$$p_n \geq p_{min}$$

$$p_n \geq p_{min} + \gamma_n\left(\frac{q_n}{C_x} - 1\right)\frac{T_n}{2}$$

The objective function (maximize Z) has no particular significance other than to bring about an appropriate equilibration between the demand and price functions. Also, k_i is the number of hours in the ith time-of-day period.

Example 2: Average Variable Cost Plus Uniform Fixed Facility Cost Toll Pricing Policy and Linear Demand

The pricing policy to be considered for this example is roughly equivalent to that experienced by travelers using toll facilities and, in some cases, transit systems (and may be viewed as an average total cost type of pricing policy). For this pricing policy, the overall money and nonmoney price (which reflects the total time, effort, and money expenses perceived by and expended by users) will include both the variable money and nonmoney components described in the discussion of the average variable cost pricing policy and the uniform money toll to cover fixed facility costs. Thus, p_i, the price during the ith demand period, will be equal to the toll plus the short-run average variable cost and may be regarded as roughly equivalent to the shortrun average total costs as shown earlier in Figure 14. If, for simplicity, we again assume that the time-dependent functions described by Equations 21 and 22 are suitable for representing the short-run average variable costs, then for this average total cost type of pricing policy the appropriate price or performance function for the ith time period would be as follows for $q_i \leq C_x$:

$$p_i = \text{toll} + p_{min} \tag{23}$$

and, for $q_i > C_x$:

$$p_i = \text{toll} + p_{min} + \gamma_i\left(\frac{q_i}{C_x} - 1\right)\frac{T_i}{2} \tag{24}$$

For situations in which the toll (or fare) was adjusted to raise revenues sufficiently to just cover the fixed facility costs (including interest), and in which the toll (or fare) remained uniform throughout the day, an average total cost pricing policy would result and the toll t would be

$$t = AFC_x \div 365 \sum_{i=1}^{n} k_i q_i = F_x \div \sum_{i=1}^{n} k_i q_i \tag{25}$$

where AFC_x is the annual fixed costs and F_x is the daily fixed costs for facility x, q_i is the hourly trip volume during the ith time-of-day period, k_i is the number of hours during the ith time period, and n is the number of time periods per day.

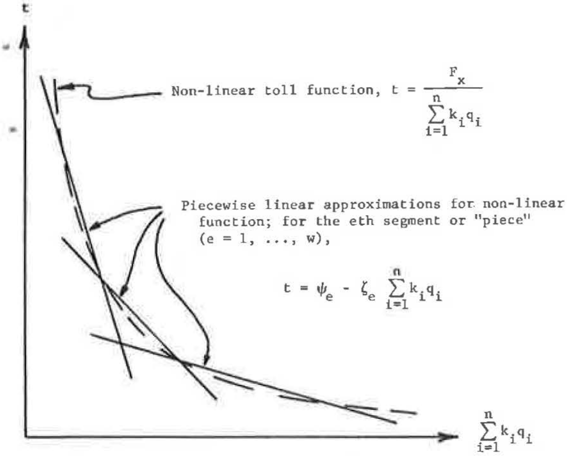


Figure 16. Linearization of nonlinear toll function.

Accompanying these price and toll functions will be demand functions that incorporate the time-of-day cross elasticities. Thus, the demand functions for the n time-of-day periods are

$$q_1 = \beta_1 - \alpha_{1,1} p_1 - \cdots - \alpha_{1,n} p_n$$
$$\vdots$$
$$q_n = \beta_n - \alpha_{n,1} p_1 - \cdots - \alpha_{n,n} p_n \tag{26}$$

To determine the equilibrium flows and prices ($q_1, \ldots, q_n$ and $p_1, \ldots, p_n$ respectively) that will simultaneously satisfy both the demand and price functions (Eqs. 23, 24, and 26 respectively) will require the use of iterative numerical or programming techniques because of the interactions resulting from time-of-day demand cross elasticities and from the toll function. Again, both these interactions and a solution technique capable of recognizing can be expressed by making use of a linear programming format. To accomplish this, however, the nonlinear toll functions must first be linearized; that is, the curvilinear toll function must be replaced or approximated by a series of piecewise linear functions. Then the price and toll functions must be expressed as inequalities, and an appropriate objective function must be chosen. Linearization of the toll function is illustrated in Figure 16; although only three linear segments or pieces were used to approximate the nonlinear function in Figure 16 (i.e., w = 3), the number of segments (w) can be increased without limit and can provide whatever accuracy is desired or necessary.

The linear programming format for this example would be as follows: Determine the $q_1, \ldots, q_n$, $p_1, \ldots, p_n$, and t values that will

$$\text{Maximize } Z = \sum_{i=1}^{n} k_i q_i$$

as subject to nonnegativity restrictions and to the following constraints:

$$q_1 = \beta_1 - \alpha_{1,1} p_1 - \cdots - \alpha_{1,n} p_n$$
$$\vdots$$
$$q_n = \beta_n - \alpha_{n,1} p_1 - \cdots - \alpha_{n,n} p_n \tag{27}$$

and

$$p_1 \geq t + p_{min}$$

$$p_1 \geq t + p_{min} + \gamma_1\left(\frac{q_1}{C_x} - 1\right)\frac{T_1}{2}$$
$$\vdots$$
$$p_n \geq t + p_{min}$$

$$p_n \geq t + p_{min} + \gamma_n\left(\frac{q_n}{C_x} - 1\right)\frac{T_n}{2} \tag{28}$$

and

$$t \geq \psi_1 - \zeta_1(k_1q_1 + \ldots + k_nq_n)$$
$$\vdots$$
$$t \geq \psi_w - \zeta_w(k_1q_1 + \ldots + k_nq_n) \tag{29}$$

For purposes of illustration, this format will be applied to a situation that is partially hypothetical and partially empirically based. The example is intended to be applicable to a six-lane, urban toll facility which is 5 miles in length; the total annual fixed costs were set at \$1.260 million (for all 5 miles), a figure that corresponds to a facility construction and right-of-way cost of roughly \$4 million a mile. The toll function associated with this fixed facility cost is shown in Figure 17 in both its nonlinear and linear form; for the piecewise linear approximation of the toll function, the form of which is described by Equation 29, six linear segments were used.

Eight time-of-day periods were used to delineate demand. The specific times of day and the time period lengths were as shown in Figure 8, and the hypothesized demand functions for the eight time periods are given in Table 2. Although all eight functions

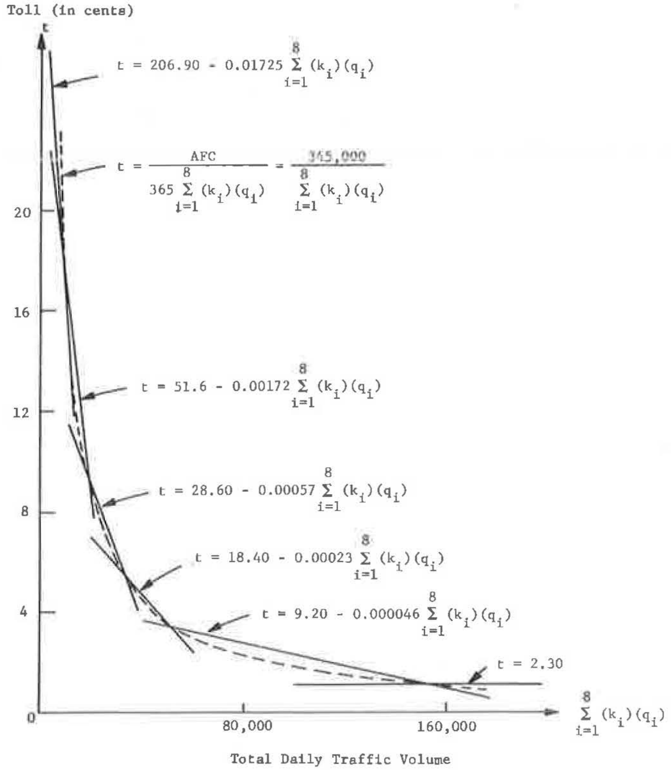


Figure 17. Linearization of average-toll-versus-daily-traffic-volume relationship for 5-mile, six-lane, urban toll facility.

TABLE 2

DEMAND FUNCTIONS FOR 5-MILE URBAN EXPRESSWAY EXAMPLE

$$
\begin{aligned}
q_1 &= 5{,}000 - 25p_1 + 19p_2 \\
q_2 &= 8{,}000 + 10p_1 - 21p_2 + 5p_3 \\
q_3 &= 7{,}200 + 6p_2 - 18p_3 \\
q_4 &= 7{,}500 + p_3 - 22p_4 + 2p_5 \\
q_5 &= 8{,}500 - 20p_5 + 4p_6 \\
q_6 &= 10{,}000 + 3p_5 - 25p_6 + 6p_7 \\
q_7 &= 8{,}600 + 10p_6 - 24p_7 \\
q_8 &= 3{,}300 - 10p_8
\end{aligned}
$$

incorporated direct demand relations (that is, the effect of the price during the ith time period on the amount of travel during the ith time period), it was not necessary to include demand cross relations (that is, to account for the effect of travel prices during other time periods on the amount of travel during the ith time period) for all cases. For example, it was assumed that nighttime travel between 7 p.m. and 6 a.m. was unaffected by peak-period travel prices and by those just before the morning peak and just after the evening peak. By contrast, travel during the morning peak-period is affected by the travel prices of both before and after the morning peak-period.

The price function used in this example represents the combined value of the toll plus the short-run average variable costs (Eq. 28); for convenience, the oversimple relationship described by Figure 12 and Equations 21 and 22 was used to represent the short-run average variable costs. (In brief, this relation ignores the effects of stochasticity; even so, neither the model structure nor the interactions are affected greatly.) The exact price functions (which embody the parameter values for γ_i and C_x) used for this purpose are given in Table 3; the values for γ_i, the unit value of travel time, congestion, and so forth, vary from time period to time period; and the value for C_x, the capacity measure, was set at 4,000 vph, and that for p_{min} was set as 64 cents.

By varying the γ_i values, it is implied that travelers during different periods react to increases in travel time and congestion to different degrees. For example, shoppers traveling during off-peak periods may be less willing or prone to tolerate some given level of congestion than will workers traveling during peak periods; or, put differently, those workers traveling to work early or late to avoid the heaviest rush periods may find travel time and increases in congestion more onerous than will those workers willing to endure the congestion occurring during the heaviest part of the rush period.

TABLE 3

PRICE FUNCTIONS FOR 5-MILE SIX-LANE (RUN 1) AND EIGHT-LANE (RUN 2) URBAN TOLL FACILITY EXAMPLE

Price Functions	For Run 1	For Run 2
$p_1 \geq 64 + t$ and	$p_1 \geq 44 + t + 0.005q_1$	$p_1 \geq 44 + t + 0.00333q_1$
$p_2 \geq 64 + t$ and	$p_2 \geq 39 + t + 0.00625q_2$	$p_2 \geq 39 + t + 0.00416q_2$
$p_3 \geq 64 + t$ and	$p_3 \geq 44 + t + 0.005q_3$	$p_3 \geq 44 + t + 0.00333q_3$
$p_4 \geq 64 + t$ and	$p_4 \geq 32 + t + 0.008q_4$	$p_4 \geq 32 + t + 0.00534q_4$
$p_5 \geq 64 + t$ and	$p_5 \geq 44 + t + 0.005q_5$	$p_5 \geq 44 + t + 0.00333q_5$
$p_6 \geq 64 + t$ and	$p_6 \geq 34 + t + 0.0075q_6$	$p_6 \geq 34 + t + 0.005q_6$
$p_7 \geq 64 + t$ and	$p_7 \geq 44 + t + 0.005q_7$	$p_7 \geq 44 + t + 0.00333q_7$
$p_8 \geq 64 + t$ and	$p_8 \geq 28 + t + 0.009q_8$	$p_8 \geq 28 + t + 0.006q_8$

Note: These functions are of the form $p_i \geq t + p_{min} + \gamma_i\,[(q_i/C_x) - 1]\,(T_i/2)$. Although the γ_i and T_i values varied from time period to time period, p_{min} and C_x were held constant and set equal to 64 cents and 4,000 vph respectively for Run 1, and to 64 cents and 6,000 vph for Run 2.

forth are independent of transportation but to imply that the characterization of this interaction should be made within an intermediate and long-run time frame.

One objective of the travel forecasting process is to list the equilibrium flows and associated performance or price levels (or, perhaps, a vector of money and nonmoney prices if certain price and time elements are to be differentiated). If we assume that all money and nonmoney elements can be combined into a single price, the desired outputs for the system would be $q_{ij}^{m,h}$ and $p_{ij}^{m,h}$ for all values of m (the variable designating mode), for all values of h (the time-of-day period), and for all combinations of i and j (origin and destination zones) except for i equal to j.

For the system and region shown in Figure 18, there would be at least four modes—drive-alone auto, car pool, bus transit, and taxi—and thus $m = 1, \ldots, 4$. If the time-of-day demand periods outlined in Figure 8 are adopted, then $h = 1, \ldots, 8$. The number of combinations (of ij pairs) for z zones can be computed by using the following combinatorial formula:

$$\text{No. of ij pairs} = \frac{z!}{2!(z-2)!} = \frac{z(z-1)}{2}$$

However, the number of directed interzonal transfer possibilities (e.g., from i to j and from j to i) will be twice this number. For the four zones in Figure 18 having land uses, there will be six ij combinations (ab, ac, ad, bc, bd, and cd) and thus the number of directed trip pair combinations (e.g., trips from a to b) will be equal to 12. In all, there will be M times H demand functions to be used for each directed ij interzonal pair and a total of $M \cdot H \cdot z(z - 1)$ demand functions to be dealt with for the region.

The total flow for a directed ij zonal pair during time-of-day period h, or q_{ij}^h, would be

$$q_{ij}^h = \sum_{m=1}^{M} q_{ij}^{m,h} \tag{30}$$

where M is the number of modes. Similarly, the total daily flow by mode m, or q_{ij}^m, can be determined by summing over all H time-of-day periods; that is,

$$q_{ij}^m = \sum_{h=1}^{H} q_{ij}^{m,h} \tag{31}$$

Neither of these flow totals, however, can be determined directly or in advance of ascertaining the equilibrium flows by mode and time of day. Nor can the total daily flow originating or ending at any zone be determined exogenously or prior to forecasting equilibrium interzonal flows by mode and time of day. Thus, q_i, the total daily trips originating at zone i, would be

$$q_i = \sum_{j=1}^{z} \sum_{m=1}^{M} \sum_{h=1}^{H} q_{ij}^{m,h} \tag{32}$$

The most difficult problems in carrying out the forecasting process (once the demand functions have been specified) are those involving the determination of the path or route that has the best price and value characteristics. There are two aspects of this phase of the forecasting process that deserve special mention. First, the prices to be used in a demand function of the form $q_{ij}^{m,h}$ = f(price for i to j trips by different modes and during different times of day; socioeconomic characteristics of residence zone i and destination zone j) are the equilibrium prices that will result from demand and supply interaction for the entire system and region. These prices also result from the accumulated travel conditions over the various links of the travel paths between zones i and j. For example, assume that an drive-alone auto trip (m = 1) from zone a to zone b follows a route involving the links connecting intersections (2,4), (3,4), (3,3), and (3,2) as well as access between the zone centroids and the first and last intersections. (An intersection can be identified by its x and y or column and row coordinates respectively; thus, in Figure 18 intersection (4,1) is that at which the fourth column and first row intersect.) As a consequence, and if we let $p^{1,h}_{(x,y)-(x',y')}$ represent the price of traveling by mode 1 during time period h over the link between intersections (x,y) and (x',y'), the total price for trip from zone a to zone b, or $p^{1,h}_{ab}$, would be

$$p^{1,h}_{ab} = p^{1,h}_{a-(2,4)} + p^{1,h}_{(2,4)-(3,4)} + p^{1,h}_{(3,4)-(3,3)} + p^{1,h}_{(3,3)-(3,2)} + p^{1,h}_{(3,2)-b} \tag{33}$$

The first and last terms represent the price for gaining access to or from the origin and destination zone and the initial and last links of the route.

Second, and as noted in an earlier section, if demand is not stratified by route as well as by mode and time of day, then some route assignment procedure must be adopted for assigning trips to the best path between ij pairs in an all-or-nothing fashion or for splitting trips among the different paths. The conceptually correct procedure, of course, would be to stratify demand to include route choices, but it would greatly enlarge the number of demand functions and the complexity of forecasting equilibrium flows and prices. However, by adopting some route assignment procedure as an alternative to stratifying demand by route, one assumes implicitly that neither the amount of trip-making nor the modal or time-of-day choices are affected significantly. (Short of a full-scale system analysis, the assumption will be difficult if not impossible to test.) All things considered, though, the adoption of an arbitrary route assignment procedure seems to be the wisest course; this judgment is embodied in the travel forecasting process shown in Figure 19. Although this flow chart incorporates a minimum-path route assignment procedure, such a technique was adopted merely for computational and illustrative convenience rather than as a result of any hard analysis of alternative procedures.

Several things come to mind with respect to the travel forecasting process. First, the analyst would like to know how different route assignment procedures would affect the equilibrium travel volumes and associated trip prices as well as the data processing requirements. Second, it is important to ask how much accuracy is necessary or feasible in equilibrating the demand and performance or price functions (i.e., how many iterations are necessary for satisfactory closure). Third, one must wonder whether a system of demand and performance functions will be sufficient to define a unique equilibrium (that is, whether more than one set of travel volumes and prices will satisfy the demand and performance function constraints for a given transport system and land-use pattern).

Although there are day-to-day variations in travel volumes and prices on urban transport networks (because of fluctuations in weather, people's living habits, etc.), at any one point in time and on any one day there is a unique amount of flow and level of congestion. Our problem, of course, is to predict this unique flow and to represent the variance associated with the flow and travel conditions. The latter aspect has been ignored throughout this paper, as it is a separate aspect from that of predicting the unique equilibrium values.

CONCLUSIONS

The foregoing methodology for forecasting travel on urban transport networks is offered with the sincere hope of generating both dialogue and research on the subject and of leading to analysis that will permit the development of models capable of realistically forecasting peak and off-peak travel volumes. The data gathering and processing requirements for the development of models along the lines suggested will be formidable but necessary to develop our forecasting procedures to a higher and more fruitful level of achievement than has been possible heretofore.

It can be argued that a better or more fundamental conceptualization of the forecasting process (perhaps in the direction of that outlined herein) will be needed to improve materially our predictive capabilities. It cannot be argued with certainty, however, that the requisite effort, in the last analysis, will prove to have been worthwhile and to have improved significantly the decision-making process. For the present, and at best, one can only hope or judge that such will be the case. It is in this sense that I urge continued improvement of our forecasting methodologies and support the commitment of the research funds required for that improvement.

REFERENCES

1. Charles River Associates. An Evaluation of Free Transit Service. U.S. Department of Transportation, August 1968.
2. Wohl, M. Notes on Transient Queuing Behavior, Capacity Restraint Functions, and Their Relationship to Travel Forecasting. Papers of the Regional Science Association, Vol. 21, 1968.
3. Wilbur Smith and Associates. Mass Transportation Survey, Traffic Engineering Study, 1958.
4. Journey to Work. New York-New Jersey Transportation Agency, 1961.
5. Systems Analysis and Research Corporation. Demand for Intercity Passenger Travel in the Washington-Boston Corridor. U.S. Department of Commerce, 1963.
6. Kraft, G., and Wohl, M. New Directions for Passenger Demand Analysis and Forecasting. Transportation Research, Vol. 1, 1967, pp. 221 ff.

Appendix

TRAVEL FORECASTING PROCESS FOR MULTIMODAL TRANSPORT NETWORKS: PARTS A THROUGH G

A. Initialization

1. Land-use data for z zones,
2. Socioeconomic data for z zones,
3. Transport system definition and capacity measures for M modes and all links,
4. Pricing and/or control policies,
5. Parameter values for $M \cdot H \cdot z(z - 1)$ demand functions,
6. Parameter values for transport link performance (or price) functions, and
7. Initial estimates of equilibrium prices or $p^{m,h}_{(x,y)-(x',y')}$ values for transport system links.

B. Determine Minimum (i.e., Least Price) Paths

For each ij pair, by mode and time of day, accumulate link prices for each alternate connecting route and determine least-price path. $\left(\text{Output will be } p_{ij}^{m,h} \text{ values for all ij pairs, m and h.}\right)$

C. Determine Interzonal Volumes

Given functions of the form $q_{ij}^{m,h} = f\left(p_{ij}^{1,1}, p_{ij}^{1,2}, \ldots, p_{ij}^{M,H-1}, p_{ij}^{M,H}; \text{socioeconomic variables}\right)$, determine the volume of trips demanded between zone pairs by mode and time period for all ij pairs and for all values of m and h. $\left(\text{Output will be } q_{ij}^{m,h} \text{ values for ij pairs and for all values of m and h.}\right)$

D. Assign Volumes to Routes

Assign interzonal volumes $\left(\text{i.e., } q_{ij}^{m,h} \text{ values}\right)$, by mode and time of day, to minimum price route between ij pairs.

E. Compute Link Volumes

Determine the volumes, by mode and time of day, on the transport system links. (Output will be $q_{(x,y)-(x',y')}^{m,h}$ values for all links and all values of m and h. For modes using common facilities, such as drive-alone auto, car pool, bus transit and taxi, the modal-link volumes will be combined in commensurate vehicle or passenger flow units, whichever is appropriate.)

F. Compute Initial Queues

For each link, by mode and time of day, compute the initial queue lengths at the start of each time period. This can be accomplished by comparing link volumes $\left(\text{or } q_{(x,y)-(x',y')}^{m,h} \text{ values}\right)$ and link capacities $\left(\text{or } C_{(x,y)-(x',y')}^{m,h} \text{ values}\right)$. (Output will be $Q_{(x,y)-(x',y')}^{m,h}$ values for all links and all value of m and h.)

G. Performance Functions

Given functions of the form $p_{(x,y)-(x',y')}^{m,h} = f\left(q_{(x,y)-(x',y')}^{m,h}; Q_{(x,y)-(x',y')}^{m,h}; C_{(x,y)-(x',y')}^{m,h}; P_k; T_h; \text{etc.}\right)$ and given modal-capacity measures for all links $\left(\text{i.e., } C_{(x,y)-(x',y')}^{m,h} \text{ values}\right)$, pricing and/or control policies (P_k values), initial link queue lengths at start of time period $\left(Q_{(x,y)-(x',y')}^{m,h} \text{ values}\right)$, and time of day period lengths (T_h values), determine resultant link prices for all links, modes, and times of day. (Output will be $p_{(x,y)-(x',y')}^{m,h}$, values for all links and for all values of m and h.)