Chapter-5
Demand & Supply of Transportation System
Chapter-5

Demand & Supply of Transportation System

What is Transportation Economics?

Transport Economics studies the movement of people and goods over space and time. Historically it has been thought of as located at the intersection of microeconomics and civil engineering, as shown on the left. However, if we think about it, traditional microeconomics is just a special case of transport economics, fixing space and time, and where the good being moved is money, as illustrated on the right.

Topics traditionally associated with Transport Economics include Privatization, Nationalization, Regulation, Pricing, Economic Stimulus, Financing, Funding, Expenditures, Demand, Production, and Externalities.

Demand Curve

How much would people pay for a final grade of an A in a transportation class?
- How many people would pay Rs.5000 for an A?
- How many people would pay Rs.500 for an A?
- How many people would pay Rs.50 for an A?
- How many people would pay Rs.5 for an A?

If we draw out these numbers, with the price on the Y-axis, and the number of people willing to pay on the X-axis, we trace out a demand curve. Unless you run into an exceptionally ethical (or hypocritical) group, the lower the price, the more people are willing to pay for an "A". We can of course replace an "A" with any other good or service, such as the price of petrol and get a similar though not identical curve.

Demand and Budgets in Transportation
We often say "travel is a derived demand". There would be no travel but for the activities being undertaken at the trip ends. Travel is seldom consumed for its own sake, the occasional "Sunday Drive" or walk in the park excepted. On the other hand, there seems to be some innate need for people to get out of the house, a 20-30 minute separation between the home and workplace is common, and 60 - 90 minutes of travel per day total is common, even for non-workers. We do know that the more expensive something is, the less of it that will be consumed. E.g. if gas prices were doubled there will be less travel overall. Similarly, the longer it takes to get from A to B, the less likely it is that people will go from A to B.

In short, we are dealing with a downward sloping demand curve, where the curve itself depends not only on the characteristics of the good in question, but also on its complements or substitutes.

The Shape of Demand

What we need to estimate is the shape of demand (is it linear or curved, convex or concave, what function best describes it), the sensitivity of demand for a particular thing (a mode, an origin destination pair, a link, a time of day) to price and time (elasticity) in the short run and the long run.

- Are the choices continuous (the number of miles driven) or discrete (car vs. bus)?
- Are we treating demand as an absolute or a probability?
- Does the probability apply to individuals (disaggregate) or the population as a whole (aggregate)?
- What is the trade-off between money and time?
- What are the effects on demand for a thing as a function of the time and money costs of competitive or complementary choices (cross elasticity).

Supply Curve

How much would a person need to pay you to write an A-quality 20 page term paper for a given transportation class?

- How many would write it for Rs.100,000?
- How many would write it for Rs.10,000?
- How many would write it for Rs.1,000?
- How many would write it for Rs.100?
- How many would write it for Rs.10?

If we draw out these numbers for all the potential entrepreneurial people available, we trace out a supply curve. The lower the price, the fewer people are willing to supply the paper-writing service.

The Supply and Demand for Transportation
What are the differences between a Boeing 747, an oil tanker, a car and a bicycle? Extensive indeed, but they each share the common goal of fulfilling a derived transport demand, and they thus all fill the purpose of supporting mobility. Transportation is a service that must be utilized immediately since it cannot be stored. Mobility must occur over transport infrastructures, providing a transport supply. In several instances, transport demand is answered in the simplest means possible, notably by walking. However, in some cases elaborate and expensive infrastructures and modes are required to provide mobility, such as for international air transportation.

An economic system including numerous activities located in different areas generates movements that must be supported by the transport system. Without movements infrastructures would be useless and without infrastructures movements could not occur, or would not occur in a cost efficient manner. This interdependency can be considered according to two concepts, which are transport supply and demand:

**Transport supply.** The capacity of transportation infrastructures and modes, generally over a geographically defined transport system and for a specific period of time. Therefore, supply is expressed in terms of infrastructures (capacity), services (frequency) and networks. The number of passengers, volume (for liquids or containerized traffic), or mass (for freight) that can be transported per unit of time and space is commonly used to quantify transport supply.

**Transport demand.** Transport needs, even if those needs are satisfied, fully, partially or not at all. Similar to transport supply, it is expressed in terms of number of people, volume, or tons per unit of time and space.

Transport supply and demand have a reciprocal but asymmetric relation. While a realized transport demand cannot take place without a corresponding level of transport supply, a transport supply can exist without a corresponding transport demand. This is common in infrastructure projects that are designed with a capacity fulfilling an expected demand level, which may or may not materialize, or may take several years to do so. Scheduled transport services, such as a public transit or airlines, are offering a transport supply that runs even if the demand is insufficient. Infrastructures also tend to be designed at a capacity level higher than the expected base scenario in case that demand turns out to be is higher than anticipated. In other cases, the demand does not materialize, often due to improper planning or unexpected socioeconomic changes.

There is a simple statistical way to measure transport supply and demand for passengers or freight:

The **passenger-km** (or passenger-mile) is a common measure expressing the realized passenger transport demand as it compares a transported quantity of passengers with a distance over which it gets carried. The **ton-km** (or ton-mile) is a common measure expressing the realized freight transport demand. Although both the passenger-km and ton-km are most commonly used to measure realized demand, the measure can equally apply for transport supply.

For instance, the transport supply of a Boeing 747-400 flight between New York and London would be 426 passengers over 5,500 kilometers (with a transit time of about 6 hours). This implies a transport supply of 2,343,000 passenger-kms. In reality, there could be a demand of 450
passengers for that flight, or of 2,465,000 passenger-km, even if the actual capacity would be of only 426 passengers (if a Boeing 747-400 with optimal seating configuration is used). In this case the realized demand would be 426 passengers over 5,500 kilometers out of a potential demand of 450 passengers, implying a system where demand is at 105% of capacity.

Transport demand is generated by the economy, which is composed of persons, institutions and industries and which generates movements of people and freight. When these movements are expressed in space they create a pattern, which reflects mobility and accessibility. The location of resources, factories, distribution centers and markets is obviously related to freight movements. Transport demand can vary under two circumstances that are often concomitant; the quantity of passengers or freight increases or the distance over which these passengers or freight are carried increases. Geographical considerations and transport costs account for significant variations in the composition of freight transport demand between countries. For the movements of passengers, the location of residential, commercial and industrial areas tells a lot about the generation and attraction of movements.

Supply and Demand Functions

Transport supply can be simplified by a set of functions representing what are the main variables influencing the capacity of transport systems. These variables are different for each mode. For road, rail and telecommunications, transport supply is often dependent on the capacity of the routes and vehicles (modal supply) while for air and maritime transportation transport supply is strongly influenced by the capacity of the terminals (intermodal supply).

- **Modal supply.** The supply of one mode influences the supply of others, such for roads where different modes compete for the same infrastructure, especially in congested areas. For instance, transport supply for cars and trucks is inversely proportional since they share the same road infrastructure.

- **Intermodal supply.** Transport supply is also dependent of the transshipment capacity of intermodal infrastructures. For instance, the maximum number flights per day between Montreal and Toronto cannot be superior to the daily capacity of the airports of Montreal and Toronto, even though the Montreal - Toronto air corridor has potentially a very high capacity.

Transport demand tends to be expressed at specific times that are related to economic and social activity patterns. In many cases, transport demand is stable and recurrent, which allows a good approximation in planning services. In other cases, transport demand is unstable and uncertain, which makes it difficult to offer an adequate level of service. For instance, commuting is a recurring and predictable pattern of movements, while emergency response vehicles such as ambulances are dealing with an unpredictable demand. Transport demand functions vary according to the nature of what is to be transported:

- **Passengers.** For the road and air transport of passengers, demand is a function of demographic attributes of the population such as income, age, standard of living, race and sex, as well as modal preferences.
- **Freight.** For freight transportation, the demand is a function of the nature and the importance of economic activities (GDP, commercial surface, number of tons of ore extracted, etc.) and of modal preferences. Freight transportation demand is more complex to evaluate than passengers.

- **Information.** For telecommunications, the demand can be a function of several criteria including the population (telephone calls) and the volume of financial activities (stock exchange). The standard of living and education levels are also factors to be considered.

### Supply / Demand Relationships

Relationships between transport supply and demand continually change, but they are mutually interrelated. From a conventional economic perspective, transport supply and demand interact until an equilibrium is reached between the quantity of transportation the market is willing to use at a given price and the quantity being supplied for that price level. However, several considerations are specific to the transport sector which make supply / demand relationships more complex:

- **Entry costs.** These are the costs incurred to operate at least one vehicle in a transport system. In some sectors, notably maritime, rail and air transportation, entry costs are very high, while in others such as trucking, they are very low. High entry costs imply that transport companies will consider seriously the additional demand before adding new capacity or new infrastructures (or venturing in a new service). In a situation of low entry costs, the market sees companies coming in or dropping, fluctuating with the demand. When entry costs are high, the emergence of a new player is uncommon while dropping out is often a dramatic event linked to a large bankruptcy. Consequently, transport activities with high entry costs tend to be oligopolistic while transport activities with low entry costs tend to have many competitors.

- **Public sector.** Few other sectors of the economy have seen such a high level of public involvement than transportation, which creates many disruptions in conventional price mechanisms. The provision of transport infrastructures, especially roads, was massively funded by governments, namely for the sake of national accessibility and regional equity. Transit systems are also heavily subsidized, namely to provide accessibility to urban populations and more specifically to the poorest segment judged to be deprived in mobility. As a consequence, transport costs are often considered as partially subsidized. Government control (and direct ownership) was also significant for several modes, such as rail and air transportation in a number of countries. The recent years have however been characterized by less governmental involvement and deregulation.

- **Elasticity.** Refers to the variation of demand in response to a variation of cost. For example, an elasticity of -0.5 for vehicle use with respect to vehicle operating costs means that an increase of 1% in operating costs would imply a 0.5% reduction in vehicle mileage or trips. Variations of transport costs have different consequences for different modes, but transport demand has a tendency to be inelastic. While commuting tends to be inelastic in terms of costs, it is elastic in terms of time. For economic sectors where freight costs are a small component of the total production costs, variations in transport costs have limited
consequences on the demand. For air transportation, especially the tourism sector, price variations have significant impacts on the demand.

As transport demand is a derived demand from individuals, groups and industries it can be desegregated into series of partial demands fulfilled by the adaptation and evolution of transport techniques, vehicles and infrastructures to changing needs. Moreover, the growing complexity of economies and societies linked with technological changes force the transport industry to constant changes. This leads to growing congestion, a reduction in transport safety, a degradation of transport infrastructures and growing concerns on environmental impacts.

4. Transportation Yield Management

Generally, transport demand is variable in time and space whereas transport supply is fixed. When demand is lower than supply, transit times are stable and predictable, since the infrastructures are able to support their load. When transport demand exceeds supply for a period in time, there is congestion with significant increases in transit times and higher levels of unpredictability. A growth of the transport demand increases the load factor of a transport network until transport supply is reached. Speed and transit times drop afterwards. The same journey can thus have different durations according to the time of the day.

Conventionally, congestion tended to have limited impacts on the fare structure as many transport operators were state owned or highly regulated. With deregulation, transport companies were able to establish a level of service reflecting market forces, as well as being able to expand, or rationalize, their capacity. Subsidies were removed, implying that the fare structure would be the dominant source of income to provide for operating and capital costs of the transport service. A common issue is that while the transport supply is relatively well known, often a scheduled service, the transport demand remains predictable, but subject to volatility. Many transport providers, particularly airline companies, have responded to the complexity of predicting transport demand with yield management approaches.

**Transportation yield management** is the process of managing the usage price of a transport asset, such as the fare paid by users, in view of continuous changes in the demand. The goal of such an approach is to maximize profit in the context where the transport supply is fixed.

Yield management leans on three conditions:

- **A fixed transport capacity** implying that transport demand is the only function that can effectively vary. For instance, the capacity of a scheduled flight or of a containership is fixed (known value) and cannot be readily changed.
- **Unused transport capacity loses all of its utility** implying that transport suppliers cannot store for another time the services that have not been used. Once an aircraft or a ship has departed, its transport capacity is lost for the concerned airport or port. Any unused capacity is therefore a loss of revenue.
Transport users are **willing to pay different rates for the same capacity or service**, implying that they value transportation differently based upon their priorities. For instance, a business traveller needing to attend a meeting values differently the same airplane seat than a tourist would. The former would be willing to pay a high price to secure a seat on a specific site while the latter tends to seek discounted values and would be unwilling to bid above a certain price threshold. Also, time dependent users of cargo services are willing to pay more for the same capacity than those who are less time dependent.

**Supply and Demand Equilibrium**

As with earning grades and cheating, transportation is not free, it costs both time and money. These costs are represented by a supply curve, which rises with the amount of travel demanded. As described above, demand (e.g. the number of vehicles which want to use the facility) depends on the price, the lower the price, the higher the demand. These two curves intersect at an equilibrium point. In the example figure, they intersect at a toll of Rs.0.50 per km, and flow of 3000 vehicles per hour. Time is usually converted to money (using a Value of Time), to simplify the analysis. Costs may be *variable* and include users’ time, out-of-pockets costs (paid on a per trip or per distance basis) like tolls, gasolines, and fares, or *fixed* like insurance or buying an automobile, which are only borne once in a while and are largely independent of the cost of an individual trip.
Classic Transport Demand / Supply Function

Many transport systems behave in accordance with supply and demand, which are influenced by cost variations. On the above figure the demand curve assumes that if transport costs are high, demand is low as the consumers of a transport service (either freight or passengers) are less likely to use it. If transport costs are low, the demand would be high as users would get more services for the same cost. The supply curve behaves inversely. If costs are high, transport providers would be willing to supply high quantities of services since high profits are likely to arise under such circumstances. If costs are low, the quantity of transport services would be low as many providers would see little benefits operating at a loss.

The equilibrium point represents a compromise between what users are willing to pay and what providers are willing to offer. Under such circumstances, an amount of traffic T1 would flow at an operating cost C1. If because of an improvement a larger amount of service is possible for the same cost (the supply curve moves from S1 to S2), a new equilibrium will be reached with a quantity of traffic T2 at a price C2. Elasticity refers to the variation of the demand in accordance to the variation of the price. The higher it is, the more the traffic in a transport system is influenced by costs variations.

Equilibrium in a Negative Feedback System

Supply and Demand comprise the economists view of transportation systems. They are equilibrium systems. What does that mean? It means the system is subject to a negative feedback process:
An increase in $A$ begets a decrease in $B$. An increase $B$ begets an increase in $A$.
Example: $A$: Traffic Congestion and $B$: Traffic Demand ... more congestion limits demand, but more demand creates more congestion.

Disequilibrium

However, many elements of the transportation system do not necessarily generate an equilibrium. Take the case where an increase in $A$ begets an increase in $B$. An increase in $B$ begets an increase in $A$. An example where $A$ an increase in Traffic Demand generates more fuel Tax Revenue ($B$) more fuel Tax Revenue generates more Road Building, which in turn increases traffic demand. (This example assumes the gas tax generates more demand from the resultant road building than costs in sensitivity of demand to the price, i.e. the investment is worthwhile). This is dubbed a positive feedback system, and in some contexts a "Virtuous Circle", where the "virtue" is a value judgment that depends on your perspective.

Similarly, one might have a "Vicious Circle" where a decrease in $A$ begets a decrease in $B$ and a decrease in $B$ begets a decrease in $A$. A classic example of this is where ($A$) is Transit Service and ($B$) is Transit Demand. Again "vicious" is a value judgment. Less service results in fewer transit riders, fewer transit riders cannot make as a great a claim on transportation resources, leading to more service cutbacks.

These systems of course interact: more road building may attract transit riders to cars, while those additional drivers pay gas taxes and generate more roads. One might ask whether positive feedback systems converge or diverge. The answer is "it depends on the system", and in particular where or when in the system you observe. There might be some point where no matter how many additional roads you built, there would be no more traffic demand, as everyone already consumes as much travel as they want to. We have yet to reach that point for roads, but we have for lots of goods. If you live in most parts of India, the price of water at your house probably does not affect how much water you use, and a lower price for tap water would not increase your water consumption; you might use substitutes if their prices were lower (or tap water were costlier). Price might affect other behaviors such as lawn watering and car washing though.

Provision

Transportation services are provided by both the public and private sectors.

Properties

The specific properties of highway transportation include:
- Users commit a significant amount of their own time to the consumption of the final good (a trip). While the contribution of user time is found in all sectors to some extent, this fact is a dominant feature of highway travel.
Links are collected into large bundles which comprise the route. Individual links may only be a small share of the bundle of links. If we begin by assuming each link is “autonomous”, than the final consumption bundle includes a large number of (imperfect) complements.

Highway networks have a very specialized geometry. Competition, in the form of alternative routes between origin and destination is almost always present. Nevertheless there are large degrees of spatial monopoly, each link uniquely occupies space, and spatial location affects the user contribution of time.

There are significant congestion effects, which occur in the absence of pricing and potentially in its presence.

Users are choosing not only a route for a trip, but whether to make that trip, choose a different destination, or not travel on the highway network (at a given time). These choices are determined by the quality of that trip and all others.

Individual links may serve multiple markets (origin-destination pairs). There are economies achieved by using the same links on routes serving different markets. This is one factor leading to a hierarchy of roads.

Quantity cannot be controlled in the short term. Once a road is deployed, it is in the network, its entire capacity available for use. However, roads are difficult to deploy, responses to demand are necessarily slow, and for all practical purposes, these decisions are irrevocable.

**Trip Generation**

**Trip Generation** is the first step in the conventional four-step transportation forecasting process (followed by Destination Choice, Mode Choice, and Route Choice), widely used for forecasting travel demands. It predicts the number of trips originating in or destined for a particular traffic analysis zone.

Every trip has two ends, and we need to know where both of them are. The first part is determining how many trips originate in a zone and the second part is how many trips are destined for a zone. Because land use can be divided into two broad category (residential and non-residential) we have models that are household based and non-household based (e.g. a function of number of jobs or retail activity).

For the residential side of things, trip generation is thought of as a function of the social and economic attributes of households (households and housing units are very similar measures, but sometimes housing units have no households, and sometimes they contain multiple households, clearly housing units are easier to measure, and those are often used instead for models, it is important to be clear which assumption you are using).

At the level of the traffic analysis zone, the language is that of land uses "producing" or attracting trips, where by assumption trips are "produced" by households and "attracted" to non-households. Production and attractions differ from origins and destinations. Trips are produced by households even when they are returning home (that is, when the household is a destination). Again it is important to be clear what assumptions you are using.
Activities

People engage in activities, these activities are the "purpose" of the trip. Major activities are home, work, shop, school, eating out, socializing, recreating, and serving passengers (picking up and dropping off). There are numerous other activities that people engage on a less than daily or even weekly basis, such as going to the doctor, banking, etc. Often less frequent categories are dropped and lumped into the catchall "Other". Every trip has two ends, an origin and a destination. Trips are categorized by purposes, the activity undertaken at a destination location. These categories are:

- Work
- Work related
- Attending school
- Other school activities
- Child care, day care, after school care
- Shopping
- Visit friends or relatives
- Personal business
- Eat meal outside home
- Entertainment, recreation, fitness
- Civic or religious
- Pick up or drop off passengers
- With another person at their activities

We can collect the data based on these categories in a city according to gender and can get some observations like:

- Men and women behave differently on average, splitting responsibilities within households, and engaging in different activities,
- Most trips are not work trips, though work trips are important because of their peaked nature (and because they tend to be longer in both distance and travel time),
- The vast majority of trips are not people going to (or from) work.

Specifying Models

How do we predict how many trips will be generated by a zone? The number of trips originating from or destined to a purpose in a zone are described by trip rates (a cross-classification by age or demographics is often used) or equations. First, we need to identify what we think are the relevant variables.

Home-end

The total number of trips leaving or returning to homes in a zone may be described as a function of:

Home-End Trips are sometimes functions of:

- Housing Units
- Household Size
• Age
• Income
• Accessibility
• Vehicle Ownership
• Other Home-Based Elements

Work-end

At the work-end of work trips, the number of trips generated might be a function as below:
Work-End Trips are sometimes functions of:
• Jobs
• Area of Workspace
• Occupancy Rate
• Other Job-Related Elements

Shop-end

Similarly shopping trips depend on a number of factors:
Shop-End Trips are sometimes functions of:
• Number of Retail Workers
• Type of Retail Available
• Area of Retail Available
• Location
• Competition
• Other Retail-Related Elements

Input Data

A forecasting activity conducted by planners or economists, such as one based on the concept of economic base analysis, provides aggregate measures of population and activity growth. Land use forecasting distributes forecast changes in activities across traffic zones.

Estimating Models

Which is more accurate: the data or the average? The problem with averages (or aggregates) is that every individual’s trip-making pattern is different.

Home-end

To estimate trip generation at the home end, a cross-classification model can be used, this is basically constructing a table where the rows and columns have different attributes, and each cell in the table shows a predicted number of trips, this is generally derived directly from data.
In the example cross-classification model: The dependent variable is trips per person. The independent variables are dwelling type (single or multiple family), household size (1, 2, 3, 4, or 5+ persons per household), and person age.

**Non-home-end**

The trip generation rates for both “work” and “other” trip ends can be developed using Ordinary Least Squares (OLS) regression (a statistical technique for fitting curves to minimize the sum of squared errors (the difference between predicted and actual value) relating trips to employment by type and population characteristics.

The variables used in estimating trip rates for the work-end are Employment in Offices (\(E_{\text{off}}\)), Retail (\(E_{\text{ret}}\)), and Other (\(E_{\text{oth}}\)).

A typical form of the equation can be expressed as

\[ T_i = a_1 E_{\text{off},i} | a_2 E_{\text{oth},i} | a_3 E_{\text{ret},i} \]

Where:
- \(T_i\) - Person trips attracted per worker in the ith zone
- \(E_{\text{off},i}\) - office employment in the ith zone
- \(E_{\text{oth},i}\) - other employment in the ith zone
- \(E_{\text{ret},i}\) - retail employment in the ith zone
- \(a_1, a_2, a_3\) - model coefficients

**Normalization**

For each trip purpose (e.g. home to work trips), the number of trips originating at home must equal the number of trips destined for work. Two distinct models may give two results. There are several techniques for dealing with this problem. One can either assume one model is correct and adjust the other, or split the difference.

It is necessary to ensure that the total number of trip origins equals the total number of trip destinations, since each trip interchange by definition must have two trip ends.

The rates developed for the home end are assumed to be most accurate.

The basic equation for normalization:

\[
T_{ij}' = \frac{\sum_{i=1}^{l} T_{ij}}{\sum_{j=1}^{J} T_{j}}
\]

**Sample Problems**

- Problem (Solution)

**Variables**

- \(T_i\) - Person trips originating in Zone i
- \(T_{ij}\) - Person Trips destined for Zone j
- \(T_{ij}'\) - Normalized Person trips originating in Zone i
• T'j- Rs. Normalized Person Trips destined for Zone j
• Th- Person trips generated at home end (typically morning origins, afternoon destinations)
• Tw- Person trips generated at work end (typically afternoon origins, morning destinations)
• Ts- Person trips generated at shop end
• Hi- Number of Households in Zone i
• Eoff,i- office employment in the ith zone
• Eret,i- retail employment in the ith zone
• Eoth,i- other employment in the ith zone
• Bn- model coefficients

Planners have estimated the following models for the AM Peak Hour

\[ T'_i = 1.5 \times H_i \]
\[ T_j = (1.5 \times E_{off,j}) + (1 \times E_{oth,j}) + (0.5 \times E_{ret,j}) \]

Where:

\[ T'_i = \text{Person Trips Originating in Zone } i \]
\[ T_j = \text{Person Trips Destined for Zone } j \]
\[ H_i = \text{Number of Households in Zone } i \]

You are also given the following data
### Data

<table>
<thead>
<tr>
<th>Variable</th>
<th>Delhi</th>
<th>Gurgaon</th>
</tr>
</thead>
<tbody>
<tr>
<td>H</td>
<td>10000</td>
<td>15000</td>
</tr>
<tr>
<td>E_{off}</td>
<td>8000</td>
<td>10000</td>
</tr>
<tr>
<td>E_{oth}</td>
<td>3000</td>
<td>5000</td>
</tr>
<tr>
<td>E_{ret}</td>
<td>2000</td>
<td>1500</td>
</tr>
</tbody>
</table>

A. What are the number of person trips originating in and destined for each city?  
B. Normalize the number of person trips so that the number of person trip origins = the number of person trip destinations. Assume the model for person trip origins is more accurate.

**Solution:**

A. What are the number of person trips originating in and destined for each city?  
B.

#### Solution to Trip Generation Problem Part A

<table>
<thead>
<tr>
<th></th>
<th>Households (H_i)</th>
<th>Office Employees (E_{off})</th>
<th>Other Employees (E_{oth})</th>
<th>Retail Employees (E_{ret})</th>
<th>Origins T_i=1.5 * H_i</th>
<th>Destinations T_j=(1.5 * E_{off,j}) + (1 * E_{oth,j}) + (0.5 * E_{ret,j})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delhi</td>
<td>10000</td>
<td>8000</td>
<td>3000</td>
<td>2000</td>
<td>15000</td>
<td>16000</td>
</tr>
<tr>
<td>Gurgaon</td>
<td>15000</td>
<td>10000</td>
<td>5000</td>
<td>1500</td>
<td>22500</td>
<td>20750</td>
</tr>
<tr>
<td>Total</td>
<td>25000</td>
<td>18000</td>
<td>8000</td>
<td>3000</td>
<td>37500</td>
<td>36750</td>
</tr>
</tbody>
</table>

B. Normalize the number of person trips so that the number of person trip origins = the number of person trip destinations. Assume the model for person trip origins is more accurate.

\[
T'_j = \frac{\sum_{i=1}^{T_j} T_i}{\sum_{j=1}^{J} T_j} = \frac{37500}{36750} = T_j \times 1.0204
\]

#### Solution to Trip Generation Problem Part B

<table>
<thead>
<tr>
<th></th>
<th>Origins (T_i)</th>
<th>Destinations (T_j)</th>
<th>Adjustment Factor</th>
<th>Normalized Destinations (T'_j)</th>
<th>Rounded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delhi</td>
<td>15000</td>
<td>16000</td>
<td>1.0204</td>
<td>16326.53</td>
<td>16327</td>
</tr>
<tr>
<td>Gurgaon</td>
<td>22500</td>
<td>20750</td>
<td>1.0204</td>
<td>21173.47</td>
<td>21173</td>
</tr>
<tr>
<td>Total</td>
<td>37500</td>
<td>36750</td>
<td>1.0204</td>
<td>37500</td>
<td>37500</td>
</tr>
</tbody>
</table>

**Destination Choice**

*Everything is related to everything else, but near things are more related than distant things.* - Waldo Tobler's 'First Law of Geography'
**Destination Choice** (or trip distribution or zonal interchange analysis), is the second component (after Trip Generation, but before Mode Choice and Route Choice) in the traditional four-step transportation forecasting model. This step matches trip makers’ origins and destinations to develop a “trip table”, a matrix that displays the number of trips going from each origin to each destination. Historically, trip distribution has been the least developed component of the transportation planning model.

**Table: Illustrative Trip Table**

<table>
<thead>
<tr>
<th>Origin \ Destination</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$T_{i1}$</td>
<td>$T_{i2}$</td>
<td>$T_{i3}$</td>
<td>$T_{iZ}$</td>
</tr>
<tr>
<td>2</td>
<td>$T_{21}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>$T_{31}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td></td>
<td></td>
<td>$T_{Z1}$</td>
<td>$T_{ZZ}$</td>
</tr>
</tbody>
</table>

Where: $T_{ij}$ = Trips from origin $i$ to destination $j$. Work trip distribution is the way that travel demand models understand how people take jobs. There are trip distribution models for other (non-work) activities, which follow the same structure.

**Fratar Models**

The simplest trip distribution models (Fratar or Growth models) simply extrapolate a base year trip table to the future based on growth, $T_{ijy+1}=g*T_{ijy}$

where:

- $T_{ijy}$ - Trips from to in year $y$
- $g$ - growth factor

Fratar Model takes no account of changing spatial accessibility due to increased supply or changes in travel patterns and congestion.

**Gravity Model**

The gravity model illustrates the macroscopic relationships between places (say homes and workplaces). It has long been posited that the interaction between two locations declines with increasing (distance, time, and cost) between them, but is positively associated with the amount of activity at each location (Isard, 1956). In analogy with physics, Reilly (1929) formulated Reilly's law of retail gravitation, and J. Q. Stewart (1948) formulated definitions of demographic gravitation, force, energy, and potential, now called accessibility (Hansen, 1959). The distance decay factor of 1/distance has been updated to a more comprehensive function of generalized cost, which is not necessarily linear - a negative exponential tends to be the preferred form. In analogy with Newton’s law of gravity, a gravity model is often used in transportation planning.

The gravity model has been corroborated many times as a basic underlying aggregate relationship (Scott 1988, Cervero 1989, Levinson and Kumar 1995). The rate of decline of the interaction
(called alternatively, the impedance or friction factor, or the utility or propensity function) has to be empirically measured, and varies by context.

Limiting the usefulness of the gravity model is its aggregate nature. Though policy also operates at an aggregate level, more accurate analyses will retain the most detailed level of information as long as possible. While the gravity model is very successful in explaining the choice of a large number of individuals, the choice of any given individual varies greatly from the predicted value.

As applied in an urban travel demand context, the disutilities are primarily time, distance, and cost, although discrete choice models with the application of more expansive utility expressions are sometimes used, as is stratification by income or auto ownership.

Mathematically, the gravity model often takes the form:

\[ T_{ij} = K_i K_j T_i T_j f(C_{ij}) \]

\[ \sum_j T_{ij} = T_i, \sum_i T_{ij} = T_j \]

\[ K_i = \frac{1}{\sum_j K_j T_j f(C_{ij})}, K_j = \frac{1}{\sum_i K_i T_i f(C_{ij})} \]

where

- \( T_{ij} = \) Trips between origin and destination
- \( T_i = \) Trips originating at
- \( T_j = \) Trips destined for
- \( C_{ij} = \) travel cost between and
- \( K_i, K_j = \) balancing factors solved iteratively.
- \( f = \) impedance or distance decay factor

It is doubly constrained so that Trips from i to j equal number of origins and destinations.
Balancing a matrix

1. Assess Data, you have $T_i$, $T_j$, $C_{ij}$
2. Compute $f(C_{ij})$, e.g.
   - $f(C_{ij}) = 1/C_{ij}^2$
   - $f(C_{ij}) = e^{-\beta C_{ij}}$
3. Iterate to Balance Matrix
   (a) Multiply Trips from Zone $i$ ($T_i'$) by Trips to Zone $j$ ($T_j'$) by Impedance in Cell $i,j$ ($f(C_{ij})$) for all $i,j$
   (b) Sum Row Totals $T_i'$, Sum Column Totals $T_j'$
   (c) Multiply Rows by $N_i = T_i/T_i'$
   (d) Sum Row Totals $T_i'$, Sum Column Totals $T_j'$
   (e) Compare $T_i$ and $T_i'$, $T_j$ and $T_j'$ if within tolerance stop, Otherwise goto (f)
   (f) Multiply Columns by $N_j = T_j/T_j'$
   (g) Sum Row Totals $T_i'$, Sum Column Totals $T_j'$
   (h) Compare $T_i$ and $T_i'$, $T_j$ and $T_j'$ if within tolerance stop, Otherwise goto (b)

Feedback

One of the key drawbacks to the application of many early models was the inability to take account of congested travel time on the road network in determining the probability of making a trip between two locations. Although Wohl noted as early as 1963 research into the feedback mechanism or the “interdependencies among assigned or distributed volume, travel time (or travel ‘resistance’) and route or system capacity”, this work has yet to be widely adopted with rigorous tests of convergence or with a so-called “equilibrium” or “combined” solution (Boyce et al. 1994). Haney (1972) suggests internal assumptions about travel time used to develop demand should be consistent with the output travel times of the route assignment of that demand. While small methodological inconsistencies are necessarily a problem for estimating base year conditions, forecasting becomes even more tenuous without an understanding of the feedback between supply and demand. Initially heuristic methods were developed by Irwin and Von Cube (as quoted in Florian et al. (1975) ) and others, and later formal mathematical programming techniques were established by Evans (1976).

Feedback and time budgets

A key point in analyzing feedback is the finding in earlier research by Levinson and Kumar (1994) that commuting times have remained stable over the past thirty years in the Washington Metropolitan Region, despite significant changes in household income, land use pattern, family structure, and labor force participation. Similar results have been found in the Twin Cities by Barnes and Davis (2000). The stability of travel times and distribution curves over the past three decades gives a good basis for the application of aggregate trip distribution models for relatively long term forecasting. This is not to suggest that there exists a constant travel time budget.
In terms of time budgets:
• 1440 Minutes in a Day
• Time Spent Traveling: ~ 100 minutes + or -
• Time Spent Traveling Home to Work: 20 - 30 minutes + or –
Research has found that auto commuting times have remained largely stable over the past forty years, despite significant changes in transportation networks, congestion, household income, land use pattern, family structure, and labor force participation. The stability of travel times and distribution curves gives a good basis for the application of trip distribution models for relatively long term forecasting.

You are given the travel times between zones, compute the impedance matrix \( f(C_{ij}) \), assuming \( f(C_{ij}) = 1/C_{ij}^2 \).

<table>
<thead>
<tr>
<th>Origin Zone</th>
<th>Destination Zone 1</th>
<th>Destination Zone 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

Compute impedances (\( f(C_{ij}) \))

Solution:

**Impedance Matrix**

<table>
<thead>
<tr>
<th>Origin Zone</th>
<th>Destination Zone 1</th>
<th>Destination Zone 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \frac{1}{2^2} = 0.25 )</td>
<td>( \frac{1}{5^2} = 0.04 )</td>
</tr>
<tr>
<td>2</td>
<td>( \frac{1}{2^2} = 0.04 )</td>
<td>( \frac{1}{2^2} = 0.25 )</td>
</tr>
</tbody>
</table>

You are given the travel times between zones, trips originating at each zone (zone 1 = 15, zone 2 = 15) trips destined for each zone (zone 1 = 10, zone 2 = 20) and asked to use the classic gravity model \( f(C_{ij}) = 1/C_{ij}^2 \).
### Travel Time OD Matrix

<table>
<thead>
<tr>
<th>Origin Zone</th>
<th>Destination Zone 1</th>
<th>Destination Zone 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

#### Solution:

(a) Compute impedances \( f(C_{ij}) \)

### Impedance Matrix

<table>
<thead>
<tr>
<th>Origin Zone</th>
<th>Destination Zone 1</th>
<th>Destination Zone 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.25</td>
<td>0.04</td>
</tr>
<tr>
<td>2</td>
<td>0.04</td>
<td>0.25</td>
</tr>
</tbody>
</table>

(b) Find the trip table

#### Balancing Iteration 0 (Set-up)

<table>
<thead>
<tr>
<th>Origin Zone</th>
<th>Trips Originating</th>
<th>Destination Zone 1</th>
<th>Destination Zone 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>0.25</td>
<td>0.04</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>0.04</td>
<td>0.25</td>
</tr>
</tbody>
</table>

#### Balancing Iteration 1

<table>
<thead>
<tr>
<th>Origin Zone</th>
<th>Trips Originating</th>
<th>Destination Zone 1</th>
<th>Destination Zone 2</th>
<th>Row Total</th>
<th>Normalizing Factor ( N_i = T_i / \sum T_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>37.50</td>
<td>12</td>
<td>49.50</td>
<td>0.303</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>6</td>
<td>75</td>
<td>81</td>
<td>0.185</td>
</tr>
<tr>
<td>Column Total</td>
<td>43.50</td>
<td>87</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Balancing Iteration 2

<table>
<thead>
<tr>
<th>Origin Zone</th>
<th>Trips Originating</th>
<th>Destination Zone 1</th>
<th>Destination Zone 2</th>
<th>Row Total</th>
<th>Normalizing Factor ( N_i = T_i / \sum T_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>11.36</td>
<td>3.64</td>
<td>15.00</td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>1.11</td>
<td>13.89</td>
<td>15.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Column Total</td>
<td>12.47</td>
<td>17.53</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalizing Factor ( N_j = T_j / \sum T_j )</td>
<td>0.802</td>
<td>1.141</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Over the years, modelers have used several different formulations of trip distribution. The first was the Fratar or Growth model (which did not differentiate trips by purpose). This structure extrapolated a base year trip table to the future based on growth, but took no account of changing spatial accessibility due to increased supply or changes in travel patterns and congestion.

The next models developed were the gravity model and the intervening opportunities model. The most widely used formulation is still the gravity model.

So while the matrix is not strictly balanced, it is very close, well within a 1% threshold, after 16 iterations.

**Destination Choice/Background**

Over the years, modelers have used several different formulations of trip distribution. The first was the Fratar or Growth model (which did not differentiate trips by purpose). This structure extrapolated a base year trip table to the future based on growth, but took no account of changing spatial accessibility due to increased supply or changes in travel patterns and congestion. The next models developed were the gravity model and the intervening opportunities model. The most widely used formulation is still the gravity model.
While studying traffic in Baltimore, Maryland, Alan Voorhees developed a mathematical formula to predict traffic patterns based on land use. This formula has been instrumental in the design of numerous transportation and public works projects around the world. He wrote "A General Theory of Traffic Movement," (Voorhees, 1956) which applied the gravity model to trip distribution, which translates trips generated in an area to a matrix that identifies the number of trips from each origin to each destination, which can then be loaded onto the network.

Evaluation of several model forms in the 1960s concluded that "the gravity model and intervening opportunity model proved of about equal reliability and utility in simulating the 1948 and 1955 trip distribution for Washington, D.C." (Heanue and Pyers 1966). The Fratar model was shown to have weakness in areas experiencing land use changes. As comparisons between the models showed that either could be calibrated equally well to match observed conditions, because of computational ease, gravity models became more widely spread than intervening opportunities models.

Some theoretical problems with the intervening opportunities model were discussed by Whitaker and West (1968) concerning its inability to account for all trips generated in a zone which makes it more difficult to calibrate, although techniques for dealing with the limitations have been developed by Ruiter (1967).

With the development of logit and other discrete choice techniques, new, demographically disaggregate approaches to travel demand were attempted. By including variables other than travel time in determining the probability of making a trip, it is expected to have a better prediction of travel behavior. The logit model and gravity model have been shown by Wilson (1967) to be of essentially the same form as used in statistical mechanics, the entropy maximization model. The application of these models differ in concept in that the gravity model uses impedance by travel time, perhaps stratified by socioeconomic variables, in determining the probability of trip making, while a discrete choice approach brings those variables inside the utility or impedance function. Discrete choice models require more information to estimate and more computational time.

Ben-Akiva and Lerman (1985) have developed combination destination choice and mode choice models using a logit formulation for work and non-work trips. Because of computational intensity, these formulations tended to aggregate traffic zones into larger districts or rings in estimation. In current application, some models, including for instance the transportation planning model used in Portland, Oregon use a logit formulation for destination choice. Allen (1984) used utilities from a logit based mode choice model in determining composite impedance for trip distribution.

However, that approach, using mode choice log-sums implies that destination choice depends on the same variables as mode choice. Levinson and Kumar (1995) employ mode choice probabilities as a weighting factor and develops a specific impedance function or “f-curve” for each mode for work and non-work trip purposes.

**Mathematics**

At this point in the transportation planning process, the information for zonal interchange analysis is organized in an origin-destination table. On the left is listed trips produced in each zone. Along the top are listed the zones, and for each zone we list its attraction. The table is $n \times n$, where $n$ = the number of zones.

Each cell in our table is to contain the number of trips from zone $i$ to zone $j$. We do not have these within cell numbers yet, although we have the row and column totals. With data organized this way, our task is to fill in the cells for tables headed $t=1$ through say $t=n$. 

157
Actually, from home interview travel survey data and attraction analysis we have the cell information for $t = 1$. The data are a sample, so we generalize the sample to the universe. The techniques used for zonal interchange analysis explore the empirical rule that fits the $t = 1$ data. That rule is then used to generate cell data for $t = 2, t = 3, t = 4$, etc., to $t = n$.

The first technique developed to model zonal interchange involves a model such as this:

$$T_{ij} = T_i \frac{T_{ij} \cdot f \left( C_{ij} \right) \cdot K_{ij}}{\sum_{j=1}^{n} T_{ij} \cdot f \left( C_{ij} \right) \cdot K_{ij}}$$

where:

- $T_{ij}$: trips from $i$ to $j$.
- $T_i$: trips from $i$, as per our generation analysis
- $T_j$: trips attracted to $j$, as per our generation analysis
- $f(C_{ij})$: travel cost friction factor, say = $C_{ij}^{tb}$
- $K_{ij}$: Calibration parameter

Zone $i$ generates $T_i$ trips; how many will go to zone $j$? That depends on the attractiveness of $j$ compared to the attractiveness of all places; attractiveness is tempered by the distance a zone is from zone $i$. We compute the fraction comparing $j$ to all places and multiply $T_i$ by it.

The rule is often of a gravity form:

$$T_{ij} = \frac{P_i \cdot P_j}{C_{ij}^{tb}}$$

where:

- $P_i$, $P_j$: populations of $i$ and $j$
- $a$, $b$: parameters

But in the zonal interchange mode, we use numbers related to trip origins ($) and trip destinations ($) rather than populations.

There are lots of model forms because we may use weights and special calibration parameters, e.g., one could write say:

$$T_{ij} = a \frac{T_{ij}\cdot T_{ij}^{cd}}{C_{ij}^{tb}}$$

or

$$T_{ij} = \frac{c \cdot T_{ij} \cdot d \cdot T_{ij}}{C_{ij}^{tb}}$$

where:

- $a$, $b$, $c$, $d$ are parameters
- $C_{ij}$: travel cost (e.g. distance, money, time)
- $T_i$: inbound trips, destinations
- $T_i$: outbound trips, origin
Mode Choice

**Mode choice analysis** is the third step in the conventional four-step transportation forecasting model, following Trip Generation and Destination Choice but before Route Choice. While trip distribution's zonal interchange analysis yields a set of origin destination tables which tells where the trips will be made, mode choice analysis allows the modeler to determine what mode of transport will be used.

The early transportation planning model developed by the Chicago Area Transportation Study (CATS) focused on transit, it wanted to know how much travel would continue by transit. The CATS divided transit trips into two classes: trips to the CBD (mainly by subway/elevated transit, express buses, and commuter trains) and other (mainly on the local bus system). For the latter, increases in auto ownership and use were traded off against bus use; trend data were used. CBD travel was analyzed using historic mode choice data together with projections of CBD land uses. Somewhat similar techniques were used in many studies. Two decades after CATS, for example, the London study followed essentially the same procedure, but first dividing trips into those made in inner part of the city and those in the outer part. This procedure was followed because it was thought that income (resulting in the purchase and use of automobiles) drove mode choice.

**Diversion Curve techniques**

The CATS had diversion curve techniques available and used them for some tasks. At first, the CATS studied the diversion of auto traffic from streets and arterial to proposed expressways. Diversion curves were also used as bypasses were built around cities to establish what percentage of the traffic would use the bypass. The mode choice version of diversion curve analysis proceeds this way: one forms a ratio, say:

\[
\frac{C_{\text{transit}}}{C_{\text{auto}}} = R
\]

where:
- \(C_m\) = travel time by mode \(m\) and
- \(R\) is empirical data in the form:

Given the \(R\) that we have calculated, the graph tells us the percent of users in the market that will choose transit.

A variation on the technique is to use costs rather than time in the diversion ratio. The decision to use a time or cost ratio turns on the problem at hand.

Transit agencies developed diversion curves for different kinds of situations, so variables like income and population density entered implicitly.

Diversion curves are based on empirical observations, and their improvement has resulted from better...
(more and more pointed) data. Curves are available for many markets. It is not difficult to obtain data and array results. Expansion of transit has motivated data development by operators and planners. Yacov Zahavi's UMOT studies contain many examples of diversion curves. In a sense, diversion curve analysis is expert system analysis. Planners could "eyeball" neighborhoods and estimate transit ridership by routes and time of day. Instead, diversion is observed empirically and charts can be drawn.

**Disaggregate Travel Demand models**

Travel demand theory was introduced in the appendix on traffic generation. The core of the field is the set of models developed following work by Stan Warner in 1962 (Strategic Choice of Mode in Urban Travel: A Study of Binary Choice). Using data from the CATS, Warner investigated classification techniques using models from biology and psychology. Building from Warner and other early investigators, disaggregate demand models emerged. Analysis is disaggregate in that individuals are the basic units of observation, yet aggregate because models yield a single set of parameters describing the choice behavior of the population. Behavior enters because the theory made use of consumer behavior concepts from economics and parts of choice behavior concepts from psychology. Researchers at the University of California, Berkeley (especially Daniel McFadden, who won a Nobel Prize in Economics for his efforts) and the Massachusetts Institute of Technology (Moshe Ben-Akiva) (and in MIT associated consulting firms, especially Cambridge Systematics) developed what has become known as choice models, direct demand models (DDM), Random Utility Models (RUM) or, in its most used form, the multinomial logit model (MNL).

Choice models have attracted a lot of attention and work; the Proceedings of the International Association for Travel Behavior Research chronicles the evolution of the models. The models are treated in modern transportation planning and transportation engineering textbooks. One reason for rapid model development was a felt need. Systems were being proposed (especially transit systems) where no empirical experience of the type used in diversion curves was available. Choice models permit comparison of more than two alternatives and the importance of attributes of alternatives. There was the general desire for an analysis technique that depended less on aggregate analysis and with a greater behavioral content. And, there was attraction too, because choice models have logical and behavioral roots extended back to the 1920s as well as roots in Kelvin Lancaster’s consumer behavior theory, in utility theory, and in modern statistical methods.
Psychological roots
Early psychology work involved the typical experiment: Here are two objects with weights, w1 and w2, which is heavier? The finding from such an experiment would be that the greater the difference in weight, the greater the probability of choosing correctly. Graphs similar to the one on the right result.
Louis Leon Thurstone proposed (in the 1920s) that perceived weight, \( w = v + e \), where is the true weight and is random with \( E(e) = 0 \).
The assumption that is normally and identically distributed (NID) yields the binary probit model.

Econometric formulation
Economists deal with utility rather than physical weights, and say that observed utility = mean utility + random term.
Utility in this context refers to the total satisfaction (or happiness) received from making a particular choice or consuming a good or service.
The characteristics of the object, x, must be considered, so we have \( u(x) = v(x) + e(x) \).
If we follow Thurston's assumption, we again have a probit model.
An alternative is to assume that the error terms are independently and identically distributed with a Weibull, Gumbel Type I, or double exponential distribution. They are much the same, and differ slightly in their tails (thicker) from the normal distribution). This yields the multinomial logit model (MNL). Daniel McFadden argued that the Weibull had desirable properties compared to other distributions that might be used. Among other things, the error terms are normally and identically distributed. The logit model is simply a log ratio of the probability of choosing a mode to the probability of not choosing a mode.

\[
\log \left( \frac{P_i}{1 - P_i} \right) = u(x_i)
\]
Observe the mathematical similarity between the logit model and the S-curves we estimated earlier, although here share increases with utility rather than time. With a choice model we are explaining the share of travelers using a mode (or the probability that an individual traveler uses a mode multiplied by the number of travelers).
The comparison with S-curves is suggestive that modes (or technologies) get adopted as their utility increases, which happens over time for several reasons. First, because the utility itself is a function of network effects, the more users, the more valuable the service, higher the utility
associated with joining the network. Second, because utility increases as user costs drop, which happens when fixed costs can be spread over more users (another network effect). Third, technological advances, which occur over time and as the number of users increases, drive down relative cost.

An illustration of a utility expression is given:

where

\( P_i \) = Probability of choosing mode i.
\( P_A \) = Probability of taking auto
\( c_A, c_T \) = cost of auto, transit
\( t_A, t_T \) = travel time of auto, transit
\( I \) = income
\( N \) = Number of travelers

With algebra, the model can be translated to its most widely used form:

\[
\frac{P_A}{1 - P_A} = e^{v_A}
\]

\[
P_A = e^{v_A} - P_A e^{v_A}
\]

\[
P_A (1 + e^{v_A}) = e^{v_A}
\]

\[
P_A = \frac{e^{v_A}}{1 + e^{v_A}}
\]

It is fair to make two conflicting statements about the estimation and use of this model:

1. It's a "house of cards", and
2. Used by a technically competent and thoughtful analyst, it's useful.

The "house of cards" problem largely arises from the utility theory basis of the model specification. Broadly, utility theory assumes that (1) users and suppliers have perfect information about the market; (2) they have deterministic functions (faced with the same options, they will always make the same choices); and (3) switching between alternatives is costless. These assumptions don't fit very well with what is known about behavior. Furthermore, the aggregation of utility across the population is impossible since there is no universal utility scale.

Suppose an option has a net utility \( u_{jk} \) (option \( k \), person \( j \)). We can imagine that having a systematic part \( v_{jk} \) that is a function of the characteristics of an object and person \( j \), plus a random part \( e_{jk} \), which represents tastes, observational errors, and a bunch of other things (it gets murky here). (An object such as a vehicle does not have utility, it is characteristics of a vehicle that have utility.) The introduction of \( e \) lets us do some aggregation. As noted above, we think of observable utility as being a function:

\[
v_A = \beta_0 + \beta_1 (c_A - c_T) + \beta_2 (t_A - t_T) + \beta_3 I + \beta_4 N
\]

where each variable represents a characteristic of the auto trip. The value \( \beta_0 \) is termed an alternative specific constant. Most modelers say it represents characteristics left out of the
equation (e.g., the political correctness of a mode, if I take transit I feel morally righteous, so $\beta 0$ may be negative for the automobile), but it includes whatever is needed to make error terms NID.

**Econometric estimation**

Turning now to some technical matters, how do we estimate $v(x)$? Utility ($v(x)$) isn’t observable. All we can observe are choices (say, measured as 0 or 1), and we want to talk about probabilities of choices that range from 0 to 1. (If we do a regression on 0s and 1s we might measure for $j$ a probability of 1.4 or -0.2 of taking an auto.) Further, the distribution of the error terms wouldn’t have appropriate statistical characteristics.

The MNL approach is to make a maximum likelihood estimate of this functional form. The likelihood function is:

$$L^* = \prod_{n=1}^{N} f(y_n | x_n, \theta)$$

we solve for the estimated parameters $\hat{\theta}$ that max $L^*$. This happens when:

$$\frac{\partial L}{\partial \hat{\theta}_N} = 0$$

The log-likelihood is easier to work with, as the products turn to sums:

$$\ln L^* = \sum_{n=1}^{N} \ln f(y_n | x_n, \theta)$$

Consider an example adopted from John Bitzan’s Transportation Economics Notes. Let $X$ be a binary variable that is $gamma$ and 0 with probability (1- $gamma$). Then $f(0) = (1- gamma)$ and $f(1) = gamma$. Suppose that we have 5 observations of X, giving the sample {1,1,1,0,1}. To find the maximum likelihood estimator of $gamma$ examine various values of $gamma$, and for these values determine the probability of drawing the sample {1,1,1,0,1} If $gamma$ takes the value 0, the probability of drawing our sample is 0. If $gamma$ is 0.1, then the probability of getting our sample is:

$$f(1,1,1,0,1) = f(1)f(1)f(1)f(0)f(1) = 0.1*0.1*0.1*0.9*0.1=0.00009.$$

We can compute the probability of
obtaining our sample over a range of gamma – this is our likelihood function. The likelihood function for \( n \)
independent observations in a logit model is

\[
I_i^* = \prod_{i=1}^{n} P_i^{Y_i} (1 - P_i)^{1-Y_i}
\]

where: \( Y_i = 1 \) or 0 (choosing e.g. auto or not-auto) and \( P_i = \) the probability of observing \( Y_i=1 \)
The log likelihood is thus:

\[
\ell = \ln L^* = \sum_{i=1}^{n} \left[ Y_i \ln P_i + (1 - Y_i) \ln (1 - P_i) \right]
\]

In the binomial (two alternative) logit model,

\[
P_{auto} = \frac{e^{v(x_{auto})}}{1 + e^{v(x_{auto})}}, \text{ so }
\]

\[
\ell = \ln L^* = \sum_{i=1}^{n} \left[ Y_i v(x_{auto}) - \ln \left( 1 + e^{v(x_{auto})} \right) \right]
\]

so

The log-likelihood function is maximized setting the partial derivatives to zero:

\[
\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \left( Y_i - \hat{P}_i \right) = 0
\]

The above gives the essence of modern MNL choice modeling.

**Independence of Irrelevant Alternatives (IIA)**

Independence of Irrelevant Alternatives is a property of Logit, but not all Discrete Choice models. In brief, the implication of IIA is that if you add a mode, it will draw from present modes in proportion to their existing shares. (And similarly, if you remove a mode, its users will switch to other modes in proportion to their previous share). To see why this property may cause problems, consider the following example: Imagine we have seven modes in our logit mode choice model (drive alone, carpool 2 passenger, carpool 3+ passenger, walk to transit, auto driver to transit (park and ride), auto passenger to transit (kiss and ride), and walk or bike). If we eliminated Kiss and Ride, a disproportionate number may use Park and Ride or carpool.

Consider another example. Imagine there is a mode choice between driving and taking a red bus, and currently each has 50% share. If we introduce another mode, let's call it a blue bus with identical attributes to the red bus, the logit mode choice model would give each mode 33.3% of the market, or in other words, buses will collectively have 66.7% market share. Logically, if the mode is truly identical, it would not attract any additional passengers (though one can imagine scenarios where adding capacity would increase bus mode share, particularly if the bus was capacity constrained.)
There are several strategies that help with the IIA problem. Nesting of choices allows us to reduce this problem. However, there is an issue of the proper Nesting structure. Other alternatives include more complex models (e.g. Mixed Logit) which are more difficult to estimate.

**Returning to roots**

The discussion above is based on the economist’s utility formulation. At the time MNL modeling was developed there was some attention to psychologist's choice work (e.g., Luce’s choice axioms discussed in his Individual Choice Behavior, 1959). It has an analytic side in computational process modeling. Emphasis is on how people think when they make choices or solve problems (see Newell and Simon 1972). Put another way, in contrast to utility theory, it stresses not the choice but the way the choice was made. It provides a conceptual framework for travel choices and agendas of activities involving considerations of long and short term memory, effectors, and other aspects of thought and decision processes. It takes the form of rules dealing with the way information is searched and acted on. Although there is a lot of attention to behavioral analysis in transportation work, the best of modern psychological ideas are only beginning to enter the field. (e.g. Golledge, Kwan and Garling 1984; Garling, Kwan, and Golledge 1994).

<table>
<thead>
<tr>
<th>You are given this mode choice model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where:</td>
</tr>
<tr>
<td>• $C_c/u =$ cost of mode (cents) / wage rate (in cents per minute)</td>
</tr>
<tr>
<td>• $C_{tvl} =$ travel time in-vehicle (min)</td>
</tr>
<tr>
<td>• $C_{otvl} =$ travel time out-of-vehicle (min)</td>
</tr>
<tr>
<td>• $D =$ mode specific dummies: (dummies take the value of 1 or 0)</td>
</tr>
<tr>
<td>• $D_1 =$ driving,</td>
</tr>
<tr>
<td>• $D_2 =$ transit with walk access, [base mode]</td>
</tr>
<tr>
<td>• $D_3 =$ transit with auto access,</td>
</tr>
<tr>
<td>• $D_4 =$ carpool</td>
</tr>
<tr>
<td>With these inputs:</td>
</tr>
</tbody>
</table>
Route Choice

Route assignment, route choice, or traffic assignment concerns the selection of routes (alternative called paths) between origins and destinations in transportation networks. It is the fourth step in the conventional transportation forecasting model, following Trip Generation, Destination Choice, and Mode Choice. The zonal interchange analysis of trip distribution provides origin-destination trip tables. Mode choice analysis tells which travelers will use which mode. To determine facility needs and costs and benefits, we need to know the number of travelers on each route and link of the network (a route is simply a chain of links between an origin and destination). We need to undertake traffic (or trip) assignment. Suppose there is a network of highways and transit systems and a proposed addition. We first want to know the present pattern of travel times and flows and then what would happen if the addition were made.
**Link Performance Function**

The cost that a driver imposes on others is called the marginal cost. However, when making decisions, a driver only faces his own cost (the average cost) and ignores any costs imposed on others (the marginal cost).

\[
\text{Average Cost} = \frac{C_T}{Q} \\
\text{Marginal Cost} = \frac{\delta C_T}{\delta Q}
\]

where \( C_T \) is the total cost, and \( Q \) is the flow.

**BPR Link Performance Function**

Suppose we are considering a highway network. For each link there is a function stating the relationship between resistance and volume of traffic. The Bureau of Public Roads (BPR) developed a link (arc) congestion (or volume-delay, or link performance) function, which we will term \( S_a(v_a) \)

\[
S_a(v_a) = t_a \left( 1 + 0.15 \left( \frac{v_a}{c_a} \right)^4 \right)
\]

\( t_a \) = free flow travel time on link \( a \) per unit of time

\( v_a \) = volume of traffic on link \( a \) per unit of time (somewhat more accurately: flow attempting to use link \( a \))

\( c_a \) = capacity of link \( a \) per unit of time

\( S_a(v_a) \) is the average travel time for a vehicle on link \( a \)

There are other congestion functions. The CATS has long used a function different from that used by the BPR, but there seems to be little difference between results when the CATS and BPR functions are compared.

**Can Flow Exceed Capacity?**

On a link, the capacity is thought of as “outflow.” Demand is inflow.

If inflow > outflow for a period of time, there is queueing (and delay).

For Example, for a 1 hour period, if 2100 cars arrive and 2000 depart, 100 are still there. The link performance function tries to represent that phenomenon in a simple way.

**Wardrop's Principles of Equilibrium**

**User Equilibrium**

Each user acts to minimize his/her own cost, subject to every other user doing the same. Travel times are equal on all used routes and lower than on any unused route.

**System optimal**

Each user acts to minimize the total travel time on the system.

**Price of Anarchy**

The reason we have congestion is that people are selfish. The cost of that selfishness (when people behave according to their own interest rather than society's) is the *price of anarchy*. 
The ratio of system-wide travel time under User Equilibrium and System Optimal conditions. 
Price of Anarchy = \( \frac{U_E}{S_O} > 1 \)
For a two-link network with linear link performance functions (latency functions), Price of 
Anarchy is \(< \frac{4}{3} \).
Is this too much? Should something be done, or is 33% waste acceptable? [The loss may be larger 
/ smaller in other cases, under different assumptions, etc.]

**Conservation of Flow**
An important factor in road assignment is the conservation of flow. This means that the number of 
vehicles entering the intersection (link segment) equals the number of vehicles exiting the 
intersection for a given period of time (except for sources and sinks).
Similarly, the number of vehicles entering the back of the link equals the number exiting the front 
(over a long period of time).

**Auto assignment**

**Long-standing techniques**
The above examples are adequate for a problem of two links, however real networks are much 
more complicated.
The problem of estimating how many users are on each route is long standing. Planners started 
looking hard at it as freeways and expressways began to be developed. The freeway offered a 
superior level of service over the local street system and diverted traffic from the local system. At 
first, diversion was the technique. Ratios of travel time were used, tempered by considerations of 
costs, comfort, and level of service.
One result was the Moore algorithm for finding shortest paths on networks.
The issue the diversion approach didn’t handle was the feedback from the quantity of traffic on 
links and routes. If a lot of vehicles try to use a facility, the facility becomes congested and travel 
time increases. Absent some way to consider feedback, early planning studies (actually, most in 
the period 1960-1975) ignored feedback. They used the Moore algorithm to determine shortest 
paths and assigned all traffic to shortest paths. That’s called all or nothing assignment because 
either all of the traffic from \( i \) to \( j \) moves along a route or it does not.
The all-or-nothing or shortest path assignment is not trivial from a technical-computational view. 
Each traffic zone is connected to \( n - 1 \) zones, so there are numerous paths to be considered. In 
addition, we are ultimately interested in traffic on links. A link may be a part of several paths, and 
traffic along paths has to be summed link by link.
An argument can be made favoring the all-or-nothing approach. It goes this way: The planning 
study is to support investments so that a good level of service is available on all links. Using the 
travel times associated with the planned level of service, calculations indicate how traffic will flow 
once improvements are in place. Knowing the quantities of traffic on links, the capacity to be 
supplied to meet the desired level of service can be calculated.

**Heuristic procedures**
To take account of the affect of traffic loading on travel times and traffic equilibria, several 
heuristic calculation procedures were developed. One heuristic proceeds incrementally. The traffic 
to be assigned is divided into parts (usually 4). Assign the first part of the traffic. Compute new 
travel times and assign the next part of the traffic. The last step is repeated until all the traffic is 
assigned. The CATS used a variation on this; it assigned row by row in the O-D table.
The heuristic included in the FHWA collection of computer programs proceeds another way.

- Step 0: Start by loading all traffic using an all or nothing procedure.
- Step 1: Compute the resulting travel times and reassign traffic.
- Step 2: Now, begin to reassign using weights. Compute the weighted travel times in the previous two loadings and use those for the next assignment. The latest iteration gets a weight of 0.25 and the previous gets a weight of 0.75.
- Step 3. Continue.

These procedures seem to work “pretty well,” but they are not exact.

Frank-Wolfe algorithm
Dafermos (1968) applied the Frank-Wolfe algorithm [1] (1956, Florian 1976), which can be used to deal with the traffic equilibrium problem.

Equilibrium Assignment
To assign traffic to paths and links we have to have rules, and there are the well-known Wardrop equilibrium (1952) conditions. The essence of these is that travelers will strive to find the shortest (least resistance) path from origin to destination, and network equilibrium occurs when no traveler can decrease travel effort by shifting to a new path. These are termed user optimal conditions, for no user will gain from changing travel paths once the system is in equilibrium.

The user optimum equilibrium can be found by solving the following nonlinear programming problem

\[
\min \sum_a \int_0^{v_a} S_a(x)dx
\]

subject to:

\[v_a = \sum_i \sum_j \sum_r \alpha_{ij}^r x_{ij}^r\]

\[\sum_r x_{ij}^r = T_{ij}\]

\[v_a \geq 0, \quad x_{ij}^r \geq 0\]

Where \(x_{ij}^r\) is the number of vehicles on path \(r\) from origin \(i\) to destination \(j\). So constraint (2) says that all travel must take place \(-i = 1 ... n; j = 1 ... n\)

\[\alpha_{ij}^{ar} = 1 \text{ if link } a \text{ is on path } r \text{ from } i \text{ to } j; \text{ zero otherwise.}\]

So constraint (1) sums traffic on each link. There is a constraint for each link on the network. Constraint (3) assures no negative traffic.

Integrating travel choices
The urban transportation planning model evolved as a set of steps to be followed and models evolved for use in each step. Sometimes there were steps within steps, as was the case for the first statement of the Lowry model. In some cases, it has been noted that steps can be integrated. More generally, the steps abstract from decisions that may be made simultaneously and it would be desirable to better replicate that in the analysis.

Disaggregate demand models were first developed to treat the mode choice problem. That problem assumes that one has decided to take a trip, where that trip will go, and at what time the trip will be made. They have been used to treat the implied broader context. Typically, a nested model will be developed, say, starting with the probability of a trip being made, then examining the choice among places, and then mode choice. The time of travel is a bit harder to treat.

Wilson’s doubly constrained entropy model has been the point of departure for efforts at the aggregate level. That model contains the constraint $t_ij = C$

where the $c_{ij}$ are the link travel costs, $t_{ij}$ refers to traffic on a link, and $C$ is a resource constraint to be sized when fitting the model with data. Instead of using that form of the constraint, the monotonically increasing resistance function used in traffic assignment can be used. The result determines zone-to-zone movements and assigns traffic to networks, and that makes much sense from the way one would imagine the system works. Zone-to-zone traffic depends on the resistance occasioned by congestion.

Alternatively, the link resistance function may be included in the objective function (and the total cost function eliminated from the constraints).

A generalized disaggregate choice approach has evolved as has a generalized aggregate approach. The large question is that of the relations between them. When we use a macro model, we would like to know the aggregate implications of the analysis. If we are doing a micro analysis, we would like to know the aggregate behavior it represents. If we are doing a micro analysis, we would like to know the aggregate implications of the analysis.

Wilson derives a gravity-like model with weighted parameters that say something about the attractiveness of origins and destinations. Without too much math we can write probability of choice statements based on attractiveness, and these take a form similar to some varieties of disaggregate demand models.

**Evaluation**

A **benefit-cost analysis** (BCA)[1] is often required in determining whether a project should be approved and is useful for comparing similar projects. It determines the stream of quantifiable economic benefits and costs that are associated with a project or policy. If the benefits exceed the costs, the project is worth doing; if the benefits fall short of the costs, the project is not. Benefit-cost analysis is appropriate where the technology is known and well understood or a minor change from existing technologies is being performed. BCA is not appropriate when the technology is new and untried because the effects of the technology cannot be easily measured or predicted.

However, just because something is new in one place does not necessarily make it new, so benefit-cost analysis would be appropriate, e.g., for a light-rail or commuter rail line in a city without rail, or for any road project, but would not be appropriate (at the time of this writing) for something truly radical like teleportation.

The identification of the costs, and more particularly the benefits, is the chief component of the “art” of Benefit-Cost Analysis. This component of the analysis is different for every project.
Furthermore, care should be taken to avoid double counting; especially counting cost savings in both the cost and the benefit columns. However, a number of benefits and costs should be included at a minimum. In transportation these costs should be separated for users, transportation agencies, and the public at large. Consumer benefits are measured by consumer’s surplus. It is important to recognize that the demand curve is downward sloping, so there a project may produce benefits both to existing users in terms of a reduction in cost and to new users by making travel worthwhile where previously it was too expensive. Agency benefits come from profits. But since most agencies are non-profit, they receive no direct profits. Agency construction, operating, maintenance, or demolition costs may be reduced (or increased) by a new project; these cost savings (or increases) can either be considered in the cost column, or the benefit column, but not both. Society is impacted by transportation project by an increase or reduction of negative and positive externalities. Negative externalities, or social costs, include air and noise pollution and accidents. Accidents can be considered either a social cost or a private cost, or divided into two parts, but cannot be considered in total in both columns. If there are network externalities (i.e. the benefits to consumers are themselves a function of the level of demand), then consumers’ surplus for each different demand level should be computed. Of course this is easier said than done. In practice, positive network externalities are ignored in Benefit Cost Analysis.

**Background**

**Early Beginnings**

When Benjamin Franklin was confronted with difficult decisions, he often recorded the pros and cons on two separate columns and attempted to assign weights to them. While not mathematically precise, this “moral or prudential algebra”, as he put it, allowed for careful consideration of each “cost” and “benefit” as well as the determination of a course of action that provided the greatest benefit. While Franklin was certainly a proponent of this technique, he was certainly not the first. Western European governments, in particular, had been employing similar methods for the construction of waterway and shipyard improvements. Ekelund and Hebert (1999) credit the French as pioneers in the development of benefit-cost analyses for government projects. The first formal benefit-cost analysis in France occurred in 1708. Abbe de Saint-Pierre attempted to measure and compare the incremental benefit of road improvements (utility gained through reduced transport costs and increased trade), with the additional construction and maintenance costs. Over the next century, French economists and engineers applied their analysis efforts to canals (Ekelund and Hebert, 1999). During this time, The Ecole Polytechnique had established itself as France’s premier educational institution, and in 1837 sought to create a new course in “social arithmetic”: “…the execution of public works will in many cases tend to be handled by a system of concessions and private enterprise. Therefore our engineers must henceforth be able to evaluate the utility or inconvenience, whether local or general, or each enterprise; consequently they must have true and precise knowledge of the elements of such investments.” (Ekelund and Hebert, 1999, p. 47). The school also wanted to ensure their students were aware of the effects of currencies, loans, insurance, amortization and how they affected the probable benefits and costs to enterprises.
In the 1840s French engineer and economist Jules Dupuit (1844, tr. 1952) published an article “On Measurement of the Utility of Public Works”, where he posited that benefits to society from public projects were not the revenues taken in by the government (Aruna, 1980). Rather the benefits were the difference between the public’s willingness to pay and the actual payments the public made (which he theorized would be smaller). This “relative utility” concept was what Alfred Marshall would later rename with the more familiar term, “consumer surplus” (Ekelund and Hebert, 1999).

Vilfredo Pareto (1906) developed what became known as Pareto improvement and Pareto efficiency (optimal) criteria. Simply put, a policy is a Pareto improvement if it provides a benefit to at least one person without making anyone else worse off (Boardman, 1996). A policy is Pareto efficient (optimal) if no one else can be made better off without making someone else worse off. British economists Kaldor and Hicks (Hicks, 1941; Kaldor, 1939) expanded on this idea, stating that a project should proceed if the losers could be compensated in some way. It is important to note that the Kaldor-Hicks criteria states it is sufficient if the winners could potentially compensate the project losers. It does not require that they be compensated.

Modern Benefit-cost Analysis
During the 1960s and 1970s the more modern forms of benefit-cost analysis were developed. Most analyses required evaluation of:
1. The present value of the benefits and costs of the proposed project at the time they occurred
2. The present value of the benefits and costs of alternatives occurring at various points in time (opportunity costs)
3. Determination of risky outcomes (sensitivity analysis)
4. The value of benefits and costs to people with different incomes (distribution effects/equity issues) (Layard and Glaister, 1994)

Probabilistic Benefit-Cost Analysis
In recent years there has been a push for the integration of sensitivity analyses of possible outcomes of public investment projects with open discussions of the merits of assumptions used. This “risk analysis” process has been suggested by Flyvbjerg (2003) in the spirit of encouraging more transparency and public involvement in decision-making. The Treasury Board of Canada’s Benefit-Cost Analysis Guide recognizes that implementation of a project has a probable range of benefits and costs. It posits that the “effective sensitivity” of an outcome to a particular variable is determined by four factors:
• the responsiveness of the Net Present Value (NPV) to changes in the variable;
• the magnitude of the variable’s range of
plausible values;
• the volatility of the value of the variable (that is, the probability that the value of the variable will move within that range of plausible values); and
• the degree to which the range or volatility of the values of the variable can be controlled.
It is helpful to think of the range of probable outcomes in a graphical sense, as depicted in Figure 1 (probability versus NPV).
Once these probability curves are generated, a comparison of different alternatives can also be performed by plotting each one on the same set of ordinates. Consider for example, a comparison between alternative A and B (Figure 2).
In Figure 2, the probability that any specified positive outcome will be exceeded is always higher for project B than it is for project A. The decision maker should, therefore, always prefer project B over project A. In other cases, an alternative may have a much broader or narrower range of NPVs compared to other alternatives (Figure 3).
Some decision-makers might be attracted by the possibility of a higher return (despite the possibility of greater loss) and therefore might choose project B. Risk-averse decision-makers will be attracted by the possibility of lower loss and will therefore be inclined to choose project A.

Discount rate
Both the costs and benefits flowing from an investment are spread over time. While some costs are one-time and borne up front, other benefits or operating costs may be paid at some point in the future, and still others received as a stream of payments collected over a long period of time. Because of inflation, risk, and uncertainty, a dollar received now is worth more than a dollar received at some time in the future.
Similarly, a dollar spent today is more onerous than a dollar spent tomorrow. This reflects the concept of time preference that we observe when people pay bills later rather than sooner. The existence of real interest rates reflects this time preference. The appropriate discount rate depends on what other opportunities are available for the capital. If simply putting the money in a government insured bank account earned 10% per year, then at a minimum, no investment earning less than 10% would be worthwhile. In general, projects are undertaken with those with the highest rate of return first, and then so on until the cost of raising capital exceeds the benefit from using that capital. Applying this efficiency argument, no project should be undertaken on cost-benefit grounds if another feasible project is sitting there with a higher rate of return.
Three alternative bases for the setting the government’s test discount rate have been proposed:
1. The social rate of time preference recognizes that a dollar's consumption today will be more valued than a dollar's consumption at some future time for, in the latter case, the dollar will be subtracted from a higher income level. The amount of this difference per dollar over a year gives the annual rate. By this method, a project should not be undertaken unless its rate of return exceeds the social rate of time preference.

2. The opportunity cost of capital basis uses the rate of return of private sector investment, a government project should not be undertaken if it earns less than a private sector investment. This is generally higher than social time preference.

3. The cost of funds basis uses the cost of government borrowing, which for various reasons related to government insurance and its ability to print money to back bonds, may not equal exactly the opportunity cost of capital.

Typical estimates of social time preference rates are around 2 to 4 percent while estimates of the social opportunity costs are around 7 to 10 percent.

Generally, for Benefit-Cost studies an acceptable rate of return (the government’s test rate) will already have been established. An alternative is to compute the analysis over a range of interest rates, to see to what extent the analysis is sensitive to variations in this factor. In the absence of knowing what this rate is, we can compute the rate of return (internal rate of return) for which the project breaks even, where the net present value is zero. Projects with high internal rates of return are preferred to those with low rates.

**Determine a present value**

The basic math underlying the idea of determining a present value is explained using a simple compound interest rate problem as the starting point. Suppose the sum of Rs.100 is invested at 7 percent for 2 years. At the end of the first year the initial Rs.100 will have earned Rs.7 interest and the augmented sum (Rs.107) will earn a further 7 percent (or Rs.7.49) in the second year. Thus at the end of 2 years the Rs.100 invested now will be worth Rs.114.49.

The discounting problem is simply the converse of this compound interest problem. Thus, Rs.114.49 receivable in 2 years time, and discounted by 7 per cent, has a present value of Rs.100.

Present values can be calculated by the following equation:

\[
P = \frac{F}{(1 + i)^n}
\]

where:
- \(F\) = future money sum
- \(P\) = present value
- \(i\) = discount rate per time period (i.e. years) in decimal form (e.g. 0.07)
- \(n\) = number of time periods before the sum is received (or cost paid, e.g. 2 years)

Illustrating our example with equations we have:

\[
P = \frac{114.49}{(1 + 0.07)^2} = 100.00
\]

The present value, in year 0, of a stream of equal annual payments of Rs.A starting year 1, is given by the reciprocal of the equivalent annual cost. That is, by:
\[ P = A \left[ \frac{(1 + i)^n - 1}{i (1 + i)^n} \right] \]

where:

• \( A \) = Annual Payment

For example: 12 annual payments of Rs.500, starting in year 1, have a present value at the middle of year 0 when discounted at 7% of: Rs.3971

\[
P = \frac{A}{i (1 + i)^n} = \frac{500 \left[ (1 + 0.07)^{12} - 1 \right]}{0.07 (1 + 0.07)^{12}} = 3971
\]

The present value, in year 0, of \( m \) annual payments of Rs.\( A \), starting in year \( n + 1 \), can be calculated by combining discount factors for a payment in year \( n \) and the factor for the present value of \( m \) annual payments. For example: 12 annual mid-year payments of Rs.250 in years 5 to 16 have a present value in year 4 of Rs.1986 when discounted at 7%.

Therefore in year 0, 4 years earlier, they have a present value of Rs.1515

\[
P_{Y=4} = A \left[ \frac{(1 + i)^n - 1}{i (1 + i)^n} \right] = 250 \left[ \frac{(1 + 0.07)^{12} - 1}{0.07 (1 + 0.07)^{12}} \right] = 1986
\]

\[
P_{Y=0} = \frac{F}{(1 + i)^n} = \frac{P_{Y=4}}{(1 + i)^n} = \frac{1986}{(1 + 0.07)^{4}} = 1515
\]

**Evaluation criterion**

Three equivalent conditions can tell us if a project is worthwhile

1. The discounted present value of the benefits exceeds the discounted present value of the costs
2. The present value of the net benefit must be positive.
3. The ratio of the present value of the benefits to the present value of the costs must be greater than one.

However, that is not the entire story. More than one project may have a positive net benefit. From the set of mutually exclusive projects, the one selected should have the highest net present value. We might note that if there are insufficient funds to carry out all mutually exclusive projects with a positive net present value, then the discount used in computing present values does not reflect the true cost of capital. Rather it is too low.

There are problems with using the internal rate of return or the benefit/cost ratio methods for project selection, though they provide useful information. The ratio of benefits to costs depends on how particular items (for instance, cost savings) are ascribed to either the benefit or cost column. While this does not affect net present value, it will change the ratio of benefits to costs (though it cannot move a project from a ratio of greater than one to less than one).
Planning
Urban, city, and town planning integrates land use planning and transportation planning to improve the built, economic and social environments of communities. **Transportation planning** evaluates, assesses, designs and sites transportation facilities. There are two approaches to planning
- Planning determines the rules of the game or constitution. This is planning as regulator or referee
- Planning determines the outcome of the game. This is planning as designer or player.
Both are typically undertaken. In land markets, the first is often the case. In transportation, which is often government provided, the second often holds.

Rationales for Planning
Why plan?
- To prepare for future contingencies, lowering the cost (in money, time, political effort) of dealing with anticipatable future outcomes. If the forecast is for rain, it is prudent to carry an umbrella.
- To establish a vision of the future to guide present action. To graduate with a Bachelor's degree in Civil Engineering, I must take CE3201.

Why plan transportation?
Transportation is, for better or worse, a public enterprise with long lasting consequences for decisions. There exist economies of coordination which may (but not necessarily) be difficult to achieve in the absence of planning. For instance, we want to ensure that roads from two different counties meet at the county line.

Why plan land use?
Owners of developed land have a right to service from public and private infrastructure. As infrastructure is costly and the economics of financing it are crude (though they need not be), planning is a substitute for the market.

Economic rationales
To invoke microeconomic theory:
In a market, equilibrium \((P_p, Q_p)\) occurs where marginal private cost equals marginal willingness to pay or demand. However if there are externalities, costs that are not borne by the parties to the transaction (e.g. noise pollution, air pollution, and congestion), there is a marginal social cost that is higher than the marginal private cost, leading to overconsumption. The first best solution is to price goods at their marginal social cost, imposing an additional charge. However, the same effect can be achieved by establishing a restrictive quota on demand \((Q_s)\).

There are other approaches to the problem besides directly increasing taxes or restricting demand. The libertarian approach for instance is the use of lawsuits, and fear thereof, to bring about good behavior. If you create a nuisance for your neighbors, they could sue you and the courts could force you to behave better. This shifts the problem from legislators and bureaucrats to judges and juries, which may or may not be an improvement. However, while this is easy to do for some kinds of obvious nuisances, it poses more problems for widely distributed externalities like air pollution.
The favored transportation economist solution to congestion is road pricing, charging a higher toll in the peak to provide the right price signal to commuters about their true costs. More radical solutions include road privatization.

The central issue with most externalities is the lack of well defined property rights: Ronald Coase (1992) argues that the problem is that of actions of economic agents have harmful effects on others.

His theorem is restated from George Stigler (1966) as "... under perfect competition, private and social costs will be equal." This analysis extends and controverts the argument of Arthur Pigou (1920), who argued that the creator of the externality should pay a tax or be liable, what is now called The Polluter Pays Principle. Coase (1992) suggests the problem is lack of property rights, and notes that the externality is caused by both parties, the polluter and the receiver of pollution. In this reciprocal relationship, there would be no noise pollution externality if no-one was around to hear. This theory echoes the Zen question ``If a tree falls in the woods and no-one is around to hear, does it make a sound?. Moreover, the allocation of property rights to either the polluter or pollutee results in a socially optimal level of production, because in theory the individuals or firms could merge and the external cost would become internal. However, this analysis assumes zero transaction costs. If the transaction costs exceed the gains from a rearrangement of activities to maximize production value, then the switch in behavior won’t be made.

There are several means for internalizing these external costs. Pigou identifies the imposition of taxes and transfers, Coase (1992) suggests assigning property rights, while government most frequently uses regulation. To some extent all have been tried in various places and times. In dealing with air pollution, transferable pollution rights have been created for some pollutants. Fuel taxes are used in some countries to deter the amount of travel, with an added rationale being compensation for the air pollution created by cars. The US government establishes pollution and noise standards for vehicles, and requires noise walls be installed along highways in some areas.

**Property rights**

Government in general, and land use planning in particular, concerns itself with restricting individual property rights for the common good. Donald Krueckeberg (1995) cites Christman who defines 9 separable kinds of property rights:

• Possession,
• Use,
• Alienation,
• Consumption,
• Modification,
• Destruction,
• Management,
• Exchange,
• Profit Taking.

Belief that your property can easily be taken, or taken without recompense, will discourage long term investment.

**Problems in urban land markets**

Whitehead identifies a number of problems in urban land markets.

• Provision of public goods
• Q: Are transportation services truly public goods?
• Existence of locational externalities
• Q: Why do location externalities exist?
• Imperfect information on which to base individual decisions
• Q: Do planners really know better?
• Unequal division of market power among economic agents
• Q: Is unequal division of market power necessary to achieve economies of scale and scope, efficiencies from doing things big or broadly?
• Differences between how individuals and communities value future and current benefits
• Q: Do communities value anything?

Arrow’s Impossibility Theorem An illustration of the problem of aggregation of social welfare functions:
Three individuals each have well-behaved preferences. However, aggregating the three does not produce a well-behaved preference function:
• Person A prefers red to blue and blue to green
• Person B prefers green to red and red to blue
• Person C prefers blue to green and green to red.
Aggregating, transitivity is violated.
• Two people prefer red to blue
• Two people prefer blue to green, and
• Two people prefer green to red.
What does society want?
• Differences between individual and community risk perception
• Q: How should the economic discount rate, be computed. This rate may be lower for communities than individuals because communities can borrow more cheaply, they tend to be lower risk for lenders than individuals. But each investment, treated on its own, should have the same discount rate applied to it.
• Interdependence in utility arising from "merit goods," consumption of which by one individual benefits others
• Q: Do merit goods exist? While there may be an external benefit to some things, it is rare that there is not a much larger internal benefit. For instance, education or good housing benefit the recipients much more than society.
• Income redistribution
• Do income redistribution schemes reduce the incentive to work? What are the long term consequences of these schemes. While in the U.S. they have been kept manageable, these schemes have undermined the economy of many countries.

Operations
Almost anybody can build a road. Its safety and durability would likely come into question if done by an amateur, but roads are not a difficult thing to deploy. Management of these roads, however, is something that is far from simple. Safety, efficiency, and sensibility are all elements that come into play when managing a road. It is seldom possible to find a managing strategy that satisfies all
three of these categories, but often a "good" strategy can be found by balancing tradeoffs. This section introduces the fundamentals of traffic management and operation, discussing everything from queueing theory to traffic signals. The discussions here cover the fundamentals for traffic operations. To this day, research is still being conducted to perfect these elements.

Queueing
Queueing is the study of traffic behavior near a certain section where demand exceeds available capacity. Queues can be seen in a variety of situations and transportation systems. Its presence is one of the causes of driver delay.

Traffic Flow
Traffic Flow is the study of the movement of individual drivers and vehicles between two points and the interactions they make with one another. While traffic flow cannot be predicted with one-hundred percent certainty due to unpredictable driver behavior, drivers tend to behave within a reasonable range that can be represented mathematically. Relationships have been established between the three main traffic flow characteristics:
• Flow
• Density
• Velocity

Queueing and Traffic Flow
Queueing and Traffic Flow is the study of traffic flow behaviors surrounding queueing events. They merge concepts learned in the prior two sections and apply them to bottlenecks.

Shockwaves
Shockwaves are the byproducts of traffic congestion and queueing. They are defined as transition zones between two traffic states that are dynamic, meaning they generally have the ability to move. Most drivers can identify shockwaves as the propagation of brake lights stemming from a given incident.

Traffic Signals
Traffic Signals are one of the more familiar types of traffic control. Using either a fixed or adaptive schedule, traffic signals allow certain parts of the intersection to move while forcing other parts to wait, delivering instructions to drivers through a set of colorful lights. While many benefits come with traffic signals, they also come with a series of costs, such as increasing delay during the off-peak. But, traffic signals are generally a well-accepted form of traffic control for busy intersections and continue to be deployed.

Additional Areas of Study
Level of Service
One of the basic assessments of roadway performance is by determining its Level of Service (LOS). These Levels of Service, typically ranging between A and F (A being good performance, F being poor performance), are assessed by certain predetermined thresholds for any of the three characteristics of traffic flow (flow, density, and/or velocity). Finding the correct LOS often requires reference to a table. When computing a value of flow to determine LOS, the following formula is used:
Similarly, LOS can be analyzed the anticipated speed of the corridor. The equation below produces the estimated free-flow speed for a strip of rural road, given certain characteristics about it:

\[ v_p = \frac{V}{PHF \times N \times f_{HV} \times f_p} \]

Where:
- \( v_p \) = Actual Volume per Lane
- \( V \) = Hourly Volume
- \( PHF \) = Peak Hour Factor
- \( N \) = Number of Lanes
- \( f_{HV} \) = Heavy Vehicle Factor
- \( f_p \) = Driver Familiarity Factor (generally, a value of 1 for commuters, less for out-of-town drivers)

Peak Hour Factor is found to be:

\[ PHF = \frac{V}{(4V_{15})} \]

Where:
- \( V \) = Hourly Volume
- \( V_{15} \) = Peak 15-Minute Flow

The Heavy Vehicle Factor is found to be:

\[ f_{HV} = \frac{1}{1 + P_T (E_T - 1) + P_R (E_R - 1)} \]

Where:
- \( P_T \) = Percentage of Trucks, in decimal
- \( E_T \) = Passenger Car Equivalent of Trucks (found through table)
- \( P_R \) = Percentage of Recreational Vehicles, in decimal
- \( E_R \) = Passenger Car Equivalent of Recreational Vehicles (found through table)

Similarly, LOS can be analyzed the anticipated speed of the corridor. The equation below produces the estimated free-flow speed for a strip of rural road, given certain characteristics about it:
Estimating Gaps

One area of transportation operations that is still being considered is the estimation of gaps between vehicles. This is applicable in a variety of situations, primarily dealing with arrivals of vehicles at a given intersection. Gaps can only be predicted through a probability, based on several known characteristics occurring at a given time or location.

One of the more popular models used for predicting arrivals is the Poisson Model. This model is shown here:

\[ FF_S = BFF_S - f_{LW} - f_{LC} - f_N - f_{ID} \]

Where:
- \( FF_S \) = Estimated Free Flow Speed
- \( BFF_S \) = Base Free Flow Speed
- \( f_{LW} \) = Adjustment for Lane Width (found in tables)
- \( f_{LC} \) = Adjustment for Lateral Clearance (found in tables)
- \( f_N \) = Adjustment for Number of Lanes (found in tables)
- \( f_{ID} \) = Adjustment for Interchange Density (found in tables)

For multilane urban highways, free-flow speed is computed with a similar equation:

\[ FF_S = BFF_S - f_{LW} - f_{LC} - f_M - f_A \]

Where:
- \( FF_S \) = Estimated Free Flow Speed
- \( BFF_S \) = Base Free Flow Speed
- \( f_{LW} \) = Adjustment for Lane Width (found in tables)
- \( f_{LC} \) = Adjustment for Lateral Clearance (found in tables)
- \( f_M \) = Adjustment for Median Type (found in tables)
- \( f_A \) = Adjustment for Access Points (found in tables)

Lateral Clearance is computed:

\[ TLC = LC_R + LC_L \]

Where:
- \( TLC \) = Total Lateral Clearance
- \( LC_R \) = Lateral Clearance on the Right Side of the Travel Lanes to Obstructions
- \( LC_L \) = Lateral Clearance on the Left Side of the Travel Lanes to Obstructions

Once TLC is computed, a table can provide the value for the Lateral Clearance Adjustment.

Estimating Gaps

One area of transportation operations that is still being considered is the estimation of gaps between vehicles. This is applicable in a variety of situations, primarily dealing with arrivals of vehicles at a given intersection. Gaps can only be predicted through a probability, based on several known characteristics occurring at a given time or location.

One of the more popular models used for predicting arrivals is the Poisson Model. This model is shown here:
Queueing

Queueing is the study of traffic behavior near a certain section where demand exceeds available capacity. Queues can be seen in many common situations: boarding a bus or train or plane, freeway bottlenecks, shopping checkout, exiting a doorway at the end of class, waiting for a computer in the lab, a hamburger at McDonald’s, or a haircut at the barber. In transportation engineering, queueing can occur at red lights, stop signs, bottlenecks, or any design-based or traffic-based flow constriction. When not dealt with properly, queues can result in severe network congestion or "gridlock" conditions, therefore making them something important to be studied and understood by engineers.

Cumulative Input-Output Diagram (Newell Curve)

Based on the departure rate and arrival rate pair data, the delay of every individual vehicle can be obtained. Using the input-output (I/O) queueing diagram shown in the side figure, it is possible to find the delay for every individual vehicle: the delay of the vehicle is time of departure - time of arrival (\( D(t) - A(t) \)). Total delay is the sum of the delays of each vehicle, which is the area in the triangle between the arrival \( A(t) \) and departure \( D(t) \) curves.
Distributions
Arrival Distribution - Deterministic (uniform) OR Random (e.g. Poisson)
Service Distribution - Deterministic OR Random Service Method:
• First Come First Serve (FCFS) or First In First Out (FIFO)
• Last Come First Served (LCFS) or Last In First Out (LIFO)
• Priority (e.g. HOV bypasses at ramp meters)

Little's Formula
This means that the average queue size (measured in vehicles) equals the arrival rate (vehicles per unit time) multiplied by the average waiting time (in units of time). This result is independent of particular arrival distributions and, while perhaps obvious, is an important fundamental principle that was not proven until 1961.

Uncapacitated Queues (M/D/1) and (M/M/1)
It has been shown that queue sizes, waiting times, and delays differ between M/D/1 and M/M/1 queueing. These differences are represented in formulas and shown below. Note the minor differences between the two.

Comparison of M/D/1 and M/M/1 queue properties

<table>
<thead>
<tr>
<th></th>
<th>M/D/1</th>
<th>M/M/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q (average queue size (#))</td>
<td>$Q = \frac{\rho^2}{2(1 - \rho)}$</td>
<td>$Q = \frac{\rho^2}{1 - \rho}$</td>
</tr>
<tr>
<td>w (average waiting time)</td>
<td>$w = \frac{\rho}{2\mu(1 - \rho)}$</td>
<td>$w = \frac{\lambda}{\mu(\mu - \lambda)}$</td>
</tr>
<tr>
<td>t (average total delay)</td>
<td>$t = \frac{2 - \rho}{2\mu(1 - \rho)}$</td>
<td>$t = \frac{1}{(\mu - \lambda)}$</td>
</tr>
</tbody>
</table>

Notes:
• Average queue size includes customers currently being served (in number of units)
• Average wait time excludes service time
• Average travel time (through the queue) is (wait time + service time). This is sometimes referred to average delay time due to the existence of the bottleneck.

Uncapacitated queues (M/M/1) (random arrival and random service)
In addition to the properties stated before, M/M/1 queueing have a few additional ones of which to take note.

Additional M/M/1 queue properties
Capacitated Queues (Finite)
Capacitated queues permit a maximum number of vehicles to wait, and thus have different properties than uncapacitated queues. For single channel undersaturated finite queues where the maximum number of units in system is specified as \( N \) and with random arrivals and departures \((\lambda, \mu)\) we have:

\[
P(n > N) = \rho^{N+1}
\]

Additional M/M/1 queue properties

<table>
<thead>
<tr>
<th>Probability of n Units in System</th>
<th>( P(n) = \frac{(1 - \rho)}{1 - \rho^{N+1}} \rho^n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected Number of Units in System</td>
<td>( E(n) = \frac{(\rho)}{(1 - \rho)} \frac{1 - (N + 1) \rho^N + N \rho^{N+1}}{1 - \rho^{N+1}} )</td>
</tr>
</tbody>
</table>

Real Life Causes of Queue Generation
For Roads:
- Geometric Bottlenecks (lane drops, hard curves, hills)
- Accidents and Incidents
- Traffic Signals and Intersection Controls
• At-Grade Crossings with other Modes (Railroad crossings, drawbridges, etc.)
• Toll Booths
• Ramp Meters
• "Gawker" Effect
• Inclement Weather

**For Trains and Transit:**
• Platform Capacities
• Fare Gates
• Ticket Windows/Ticket Machines
• Minimum Safe Separation for Trains
• Security Checkpoints
• Efficiency of Trains Entering and Leaving Station (number of tracks, switches, etc.)

**For Aviation and Airports:**
• Runways
• Designated Minimum Safe Following Distances for Planes (by government)
• Physical Minimum Safe Following Distance for Planes (creation of turbulence, etc.)
• Available Airspace for Approaches and Departures
• Ticketing Counters/Check-in Procedures
• Security Checkpoints
• Baggage Systems
• Terminal Capacity for Planes
• Internal Terminal Capacity for Passengers
• Inclement Weather

**For Water and Maritime:**
• Locks and Dams
• Port Capacities
• Minimum "Safe" Distances (determined by government and physics)
• Inclement Weather

**For Space Flight:**
• Launch Capacity
• Minimum Spacings between Orbital Vehicles
• Inclement Weather on Earth
• Unfavorable Celestial Conditions

**Problem:**
At the Krusty-Burger, if the arrival rate is 1 customer every minute and the service rate is 1 customer every 45 seconds, find the average queue size, the average waiting time, and average total delay. Assume an M/M/1 process.

**Solution:**
To proceed, we convert everything to minutes.
Service time:
Comparison of M/D/1 and M/M/1 queue properties

<table>
<thead>
<tr>
<th></th>
<th>M/D/1</th>
<th>M/M/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Q ) (average queue size (#))</td>
<td>1.125</td>
<td>2.25</td>
</tr>
<tr>
<td>( w ) (average waiting time)</td>
<td>1.125</td>
<td>2.25</td>
</tr>
<tr>
<td>( t ) (average total delay)</td>
<td>1.88</td>
<td>3</td>
</tr>
</tbody>
</table>

As can be seen, the delay associated with the more random case (M/M/1, which has both random arrivals and random service) is greater than the less random case (M/D/1), which is to be expected.

Traffic Flow

Traffic Flow is the study of the movement of individual drivers and vehicles between two points and the interactions they make with one another. Unfortunately, studying traffic flow is difficult because driver behavior is something that cannot be predicted with one-hundred percent certainty. Fortunately, drivers tend to behave within a reasonably consistent range and, thus, traffic streams tend to have some reasonable consistency and can be roughly represented mathematically. To better represent traffic flow, relationships have been established between the three main characteristics: (1) flow, (2) density, and (3) velocity. These relationships help in planning, design, and operations of roadway facilities.
Traffic flow theory
Time-Space Diagram

Traffic engineers represent the location of a specific vehicle at a certain time with a time-space diagram. This two-dimensional diagram shows the trajectory of a vehicle through time as it moves from a specific origin to a specific destination. Multiple vehicles can be represented on a diagram and, thus, certain characteristics, such as flow at a certain site for a certain time, can be determined.

**Flow and Density**
Flow \( (q) \) = the rate at which vehicles pass a fixed point (vehicles per hour)

\[
q = \frac{3600N}{t_{\text{measured}}}
\]

Density (Concentration) \( (k) \) = number of vehicles \( (N) \) over a stretch of roadway \( (L) \) (in units of vehicles per kilometer)

\[
k = \frac{N}{L}
\]

where:

- \( N \) = number of vehicles passing a point in the roadway in \( t_{\text{measured}} \) sec
- \( q \) = equivalent hourly flow
- \( L \) = length of roadway
- \( k \) = density

**Speed**
Measuring speed of traffic is not as obvious as it may seem; we can average the measurement of the speeds of individual vehicles over time or over space, and each produces slightly different results.

**Time mean speed**
Time mean speed \( (\bar{v}) \) = arithmetic mean of speeds of vehicles passing a point
Space mean speed

Space mean speed \( \bar{v}_s \) is defined as the harmonic mean of speeds passing a point during a period of time. It also equals the average speeds over a length of roadway.

\[
\bar{v}_s = \frac{N}{\sum_{n=1}^{N} \frac{1}{v_n}}
\]

Headway

Time headway

Time headway \( h_t \) = difference between the time when the front of a vehicle arrives at a point on the highway and the time the front of the next vehicle arrives at the same point (in seconds).

Average Time Headway \( \bar{h}_t \) = Average Travel Time per Unit Distance \(*\) Average Space Headway

\[
\bar{h}_t = \bar{t} \times \bar{h}_s
\]

Space headway

Space headway \( h_s \) = difference in position between the front of a vehicle and the front of the next vehicle (in meters).

Average Space Headway \( \bar{h}_s \) = Space Mean Speed \(*\) Average Time Headway

\[
\bar{h}_s = \bar{v}_s \times \bar{h}_t
\]

Note that density and space headway are related:

\[
k = \frac{1}{\bar{h}_s}
\]
Fundamental Diagram of Traffic Flow

The variables of flow, density, and space mean speed are related definitionally as:

\[ q = k \bar{v}_s \]

**Traditional Model (Parabolic)**

Properties of the traditional fundamental diagram.
- When density on the highway is zero, the flow is also zero because there are no vehicles on the highway.
- As density increases, flow increases.
- When the density reaches a maximum jam density \( k_j \), flow must be zero because vehicles will line up end to end.
- Flow will also increase to a maximum value \( q_{\text{max}} \), increases in density beyond that point result in reductions of flow.
- Speed is space mean speed.
- At density = 0, speed is freeflow \( \bar{v}_f \). The upper half of the flow curve is uncongested, the lower half is congested.
- The slope of the flow density curve gives speed. Rise/Run = Flow/Density = Vehicles per hour/Vehicles per km = km/hour.

189
Observation (Triangular or Truncated Triangular)
Actual traffic data is often much noisier than idealized models suggest. However, what we tend to see is that as density rises, speed is unchanged to a point (capacity) and then begins to drop if it is affected by downstream traffic (queue spillbacks). For a single link, the relationship between flow and density is thus more triangular than parabolic. When we aggregate multiple links together (e.g. a network), we see a more parabolic shape.

Microscopic and Macroscopic Models
Models describing traffic flow can be classed into two categories: microscopic and macroscopic. Ideally, macroscopic models are aggregates of the behavior seen in microscopic models.

Microscopic Models
Microscopic models predict the following behavior of cars (their change in speed and position) as a function of the behavior of the leading vehicle.

Macroscopic Models
Macroscopic traffic flow theory relates traffic flow, running speed, and density. Analogizing traffic to a stream, it has principally been developed for limited access roadways (Leutzbach 1988). The fundamental relationship “q=kv” (flow (q) equals density (k) multiplied by speed (v)) is illustrated by the fundamental diagram. Many empirical studies have quantified the component bivariate relationships (q vs. v, q vs. k, k vs. v), refining parameter estimates and functional forms (Gerlough and Huber 1975, Pensaud and Hurdle 1991; Ross 1991; Hall, Hurdle and Banks 1992; Banks 1992; Gilchrist and Hall 1992; Disbro and Frame 1992).

The most widely used model is the Greenshields model, which posited that the relationships between speed and density is linear. These were most appropriate before the advent of high-powered computers enabled the use of microscopic models. Macroscopic properties like flow and density are the product of individual (microscopic) decisions. Yet those microscopic decision-makers are affected by the environment around them, i.e. the macroscopic properties of traffic.
While traffic flow theorists represent traffic as if it were a fluid, queueing analysis essentially treats traffic as a set of discrete particles. These two representations are not necessarily inconsistent. The figures to the right show the same 4 phases in the fundamental diagram and the queueing input-output diagram.

**Time-Mean and Space-Mean Speeds**

**Problem:**
Given five observed velocities (60 km/hr, 35 km/hr, 45 km/hr, 20 km/hr, and 50 km/hr), what is the time-mean speed and space-mean speed?

**Solution:**
Time-Mean Speed:

\[
\bar{v}_t = \frac{1}{5} (60 + 35 + 45 + 20 + 50) = 42
\]

Space-Mean Speed:

\[
\bar{v}_n = \frac{N}{\sum_{n=1}^{N} \frac{1}{v_n}} = \frac{5}{\frac{1}{60} + \frac{1}{35} + \frac{1}{45} + \frac{1}{20} + \frac{1}{50}} = 36.37
\]

The time-mean speed is 42 km/hr and the space-mean speed is 36.37 km/hr.
Queueing and Traffic Flow

Queueing and Traffic Flow is the study of traffic flow behaviors surrounding queueing events. It combines elements from the previous two sections, Queueing and Traffic Flow.

The first side figure illustrates a traffic bottleneck that drops the roadway from two lanes to one. It allows us to illustrate the changes in capacity, the changes in lane flow, and the stability in the total section flow. Over an extended period of time, by laws of conservation, flow through the bottleneck (q) must equal flow through the upstream section (Q).

A bottleneck causes lane flow (Ql) to drop, but not section flow (Q), where:
When researchers observe a “backward-bending” flow-travel time curve, this is occasionally what they are seeing.

\[ Q = \sum_{t} Q_t \]

The lane flow drops upstream of a bottleneck when demand for a bottleneck exceeds the capacity of the bottleneck. Thus, the same level of flow can be observed at two different speeds on an upstream lane. The first is at freeflow speed in uncongested conditions and the second is at a lower speed when the downstream bottleneck is at capacity.

In other words, congestion (a queue) forms when \( Q > Q_{\text{max}} \) for any period of time. Ultimately, what enters the queue must eventually exit; otherwise the queue will grow to infinity. This is illustrated with the queueing input-output (IO) diagram shown in the second side figure. We can identify four distinct phases in traffic.

- Phase 1 is the uncongested phase when there is no influence of the increasing density on the speeds of the vehicles. The speed does not drop with the introduction of newer vehicles onto the freeway.
- Phase 2 finds the freeway cannot sustain the speed with injection of newer vehicles into the traffic stream. The density increases while speed falls, maintaining the flow.
- Phase 3 shows decreased speed and decreased flows. This is caused by very low speeds cause the queue discharge to drop (slightly) at an active bottleneck, or a queue from a downstream bottleneck may be constraining the flow.
- Phase 4 is the recovery phase. During this phase, the density of traffic starts decreasing and speed starts increasing.

As reported by many researchers (Banks, 1991; Hall and Agyemang-Duah, 1991; Persaud and Hurdle, 1991; Cassidy and Bertini, 1999), the reduced bottleneck capacity after breakdown ranges from 0% to 8%. A majority of papers conclude that the section capacity reduction a queue presents is negligible. (As noted above, lane flow upstream of a bottleneck does drop, but this is due to the downstream bottleneck). This is a second source of the “backward-bending” phenomenon.

Traditional queues have “servers”. For example, the check-out line at the grocery store can serve one customer every 5 minutes, or one item every 10 seconds, etc. A conveyor belt can serve so many packages per hour. Capacity is often referred to as belonging to the road. However when we talk about road capacity, it is really a misnomer, as capacity is located in the driver, more precisely
in the driver’s willingness and ability to drive behind the driver ahead. If drivers were willing and able to drive behind the vehicle ahead with no gap (spacing between vehicles), and that driver was driving behind the driver ahead of him with no gap, and so on, at high speeds, many more vehicles per hour could use the road. However, while some compression of vehicles occurs in heavy traffic, this situation is unstable because a driver will tap the brakes or even let-up on the accelerator for any number of reasons (to change lanes, to respond to someone else trying to change lanes, because he sees an object in the road, to limit the forces when rounding a corner, etc). This by definition lowers his speed, which in turn will lower the flow. Risk-averse drivers behind him will slow down even more (the driver has established some unpredictability in behavior, it is reasonable for other drivers to establish an even larger gap to accommodate the unpredictable driver’s behavior, especially given there is a reaction time between the lead driver’s actions, and the following driver’s perception (lead vehicle is slowing), decision (must brake), action (tap the brakes), the vehicle’s response to the action (tighten brakes on wheel)). In this way, maximum flow possible, our capacity ($q_{max}$), is a function of the drivers. The road shapes the driver’s willingness to take risks. Drivers will slow down around curves, vehicles may have difficulty accelerating uphill, or even from a slower speed (and even if they don’t, driver’s may provide insufficient fuel before they realize they are going too slow by not giving the vehicle enough gas), merges take time to avoid collision, etc. Clearly, different drivers and different vehicles (for instance racecars or taxis) could increase the maximum flow through the bottleneck ($q_{max}$). It is better to think of capacity as a maximum sustainable flow (over an extended period of time), given typical drivers’ willingness to follow (subject to highway geometrics and environmental conditions) and their vehicles’ ability to respond to decisions. A series of aggressive drivers may exceed this ‘capacity’ for a short period of time, but eventually more cautious drivers will even out the function.

The IO diagram lets us understand delay in a way that the fundamental diagrams of traffic flow don’t easily allow.

The first point to note about the IO diagram is that delay differs for each driver. The average delay can be measured easily (the total area in the triangle is the total delay, the average delay is just that triangle divided by the number of vehicles). The variation (or standard deviation) can also be measured with some more statistics. As the total number of vehicles increases, the average delay increases.

The second point to note about the IO diagram is that the total number of queued vehicles (the length of the queue) can also be easily measured. This also changes continuously; the length of the queue rises and falls with changes in the arrival rate (the tail of the queue edge of the shockwave, where travel speed changes suddenly). Shockwaves indicate a change in state, or speed that suddenly occurs. One such shockwave is found where vehicles reach the back of queues. The arrival rate, coupled with the queue clearance rate, will tell you where the back of the queue is. This is interesting and is where the traveler first suffers delay – but it is not the source of the delay.

Only if the arrival rate exactly equals the departure rate would we expect to see a fixed length of the queue. If the queue results from a management practice (such as ramp metering), we can control the departure rate to match the arrival rate, and ensure that the queue remains on the ramp and doesn’t spill over to neighboring arterials. However that observation suggests that congested
“steady state” is likely to be a rare phenomenon, since in general the arrival rate does not equal the service rate.
Queueing analysts often make a simplifying assumption that vehicles stack vertically (queues take place at a point).
This is of course wrong, but not too wrong. The resulting travel time is almost the same as if the queue were measured over space. The difference is that the time required to cover distance is included when we make the better assumption, even under free flow conditions it takes time to travel from the point where a vehicle entered the back of the queue to the point where it exits the front. We can make that correction, but when queueing is taking place, that time is often small compared to the time delayed by the queue. We also assume a first-in, first-out logic, though again this can be relaxed without distracting from the main point. Another assumption we will make for exposition is that this is a deterministic process; vehicles arrive in a regular fashion and depart in a regular fashion. However sometimes vehicles bunch up (drivers are not uniform), which leads to stochastic arrivals and departures. This stochastic queueing can be introduced, and will in general increase the measured delay.
Hypercongestion

It has been observed that the same flow can be achieved on many links at two different speeds. Some call this the “backward-bending” phenomenon (Hau 1998, Crozet and Marlot 2001). The queueing analysis framework also has implications for “hypercongestion” or “backward-bending” flow-travel time curves, such as shown in the figure to the right. Recall we identified two sources for “backward-bending” speed-flow relationships. The first has to do with the point of observation. Observing the lane flow upstream of a bottleneck gives the impression of a backward bending relationship, but this disappears at the bottleneck itself. Under any given demand pattern, flow and speed are a unique pair. When demand is below the downstream active bottleneck’s capacity, a flow on an upstream link can be achieved at high speed. When demand is above the downstream active bottleneck’s capacity, the same flow on the upstream link can only be achieved at a low speed because of queueing. The second has to do with a capacity drop at the bottleneck itself under congested conditions. However, much research reports that this drop is slight to non-existent. As in the bottleneck, we define lowercase \( q \) to be flow (vehicles per hour) departing the front of the bottleneck and uppercase \( Q \) to be flow arriving at the back of the bottleneck. We also define \( k \) to be density (vehicles per kilometer), \( v \) to be speed (kilometers per hour), and \( s \) to be service rate (seconds per vehicle). The fundamental diagrams of traffic flow (\( q-k-v \) curves) represent a model of traffic flow as stylized in the traditional textbook representation of the fundamental diagram of traffic flow. The reasons why \( q \) should drop as \( k \) increases beyond a certain point at an isolated bottleneck are unclear. In other words, why should flow past a point drop just because the number of vehicles behind that point increases? Why should leading traffic be influenced by the behavior of following traffic?

If traffic behaves as a queue through a bottleneck (illustrated above), we should consider reasons why traffic flow departing the queue would not stay at its maximum. One reason is if the vehicles in the queue could not travel fast enough so that the front of the following car could not reach the point of the front of the leading car in the time allotted the service rate. If a lane serves 1800 vehicles per hour, it serves 1 vehicle every 2 seconds, we say there is a 2 second service rate. If there were a 10 meter spacing between the vehicles (including the vehicle length plus a physical gap), this implies that the service rate would be worse than 1 vehicle every 2 seconds only if it took longer than 2 seconds to travel 10 meters (i.e. speed < 18 km/hour). This is very slow traffic and may properly be called hypercongestion. Why would traffic get that slow if free flow speed is 120 km / hour? In general traffic departing the front of the queue won’t be traveling that slowly, as the shockwave that reduces speed (the wave of red brake lights) has moved to the back of the queue.
A second reason the bottleneck flow might drop is if the departure flow is affected by external sources, namely a bottleneck further downstream spilling backwards. If the queue at a downstream bottleneck becomes sufficiently long, it will reduce the number of vehicles that can depart the upstream bottleneck. This is because the first bottleneck is no longer controlling, the downstream bottleneck is.

A third reason that bottleneck flow might drop is if the bottleneck is not being fully served (slots for cars are going unfilled). If vehicles are separated by large time headways, then the bottleneck might lose capacity. Thus, if vehicles choose large gaps, the bottleneck flow might drop. However, in congested situations, vehicles tend to follow more closely, not less closely.

For an isolated bottleneck, the departure flow remains (essentially) a constant and arrival flow varies. As more vehicles arrive at the back of the queue, the expected wait at the queue increases. All vehicles will eventually be served. In general, there is no practical constraint on the number of vehicles arriving at the back of the queue, but there is a maximum output flow in vehicles per unit time. Examining traffic upstream of the bottleneck is interesting, but does not get to the root of the problem – the bottleneck itself. This view of hyper congestion is thus not inconsistent with the conclusion drawn by Small and Chu (1997) that the hyper congested region is unsuitable for use as a supply curve in congestion pricing analyses.