Notes on Transportation Planning

CE 3500

Stephen Boyles

Spring 2011

1 Introduction

Transportation planning provides the "big picture" view of transportation systems. When doing planning, we are concerned with a relatively large area all at once, such as a major metropolitan area — in Austin, Texas, for instance, the official planning model consists of the five counties surrounding the city. In Chicago, the planning model consists of sixteen counties. As a rule of thumb, the scope of a planning model should be large enough that most of the trips in your model both begin and end within the study area. For example, if constructing a planning model for Laramie, it would be wrong only to include the part of the city south of Grand Avenue and east of 3rd Street, because many of the drivers in this part of the city are either going to somewhere else (say, the University, or downtown) or coming from there. When we move to traffic operations in a few weeks, we won't need this requirement, and we'll take a look at smaller areas in greater detail.

The reason for this is that the major questions in transportation planning all relate to how and why people are traveling in the first place. Where are travelers coming from? Where are they going? What time of day are they traveling? What mode of transportation are they using (car, bus, bicycle, etc.)? What route are they taking? It is simply impossible to answer questions such as these if most of the trips start or end outside of your study area.

The main goal of transportation planning is to predict how travelers will use the transportation system: the number of drivers on each road, the number of passengers on each bus route, and so forth. These are called *link flows.* Predicting link flows allows a city or state government to evaluate different options. For instance, if a new bridge over the railroad tracks is built in Laramie, a planning model is used to predict how people will change the routes they take, and to measure the impact on traffic levels in different neighborhoods. On a larger scale, Wyoming is considering a toll on I-80, and a planning model would be used to see how many trucks divert onto other roadways. Sometimes, a "do nothing" model is used to raise support for transportation improvement projects, to show what would happen in the future if nothing is done while travel demand continues to grow. Frequently, a planning model is used to provide quantitative comparison of several different options.

If link flows are the output of a planning model, the main input is demographic data. That is, *given certain information about a population (number of people, income, amount of employment, etc.), we want to predict how many trips they will make, and how they will choose to travel.* Census records form an invaluable resource for this, often supplemented with travel surveys. Commonly, a medium-to-large random sample of the population is offered some money in exchange for keeping detailed diaries indicating all of the trips made within the next several weeks, including the time of day, reason for traveling, and other details.

To get link flows from demographic data, most regions use the so-called *four-step model* (Figure 1).

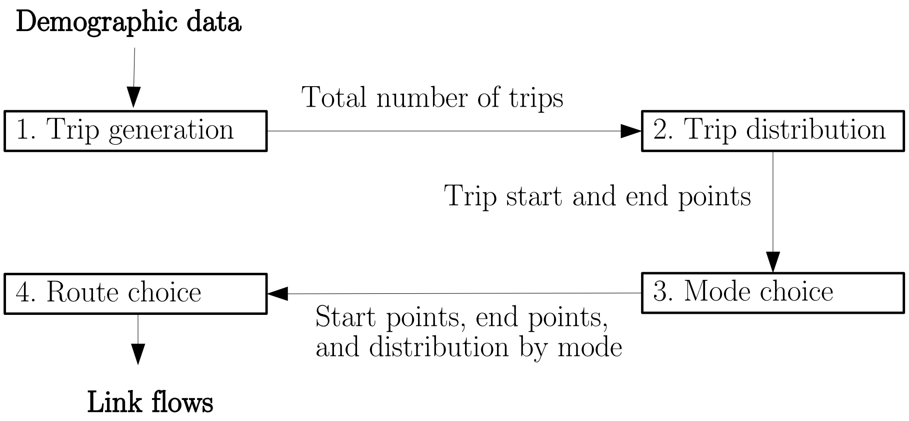


Figure 1: Schematic of the four-step process.

The first step is *trip generation*: based on demographic data, how many trips will people make? The second is *trip distribution*: once we know the total number of trips people make, what are the specific locations people will travel to? The third is *mode choice*: once we know the trip locations, will people choose to drive, take the bus, or use another mode? The fourth and final step is *route choice*: once we know the modes people will take to their trip destinations, what routes will they choose? Thus, at the end of the four steps, the transition from demographic data to link flows has been accomplished.[1]

Demographics are not uniform in a city; some areas are wealthier than others, some areas are residential while others are commercial, some parts are more crowded while other parts have a lower population density. For this reason, planners divide a city into multiple *zones*, and assume uniform conditions within each zone. Clearly this is only an approximation to reality, and the larger the number of zones, the more accurate the approximation. (At the extreme, each household would be its own zone and the uniformity assumption becomes irrelevant.) On the other hand, the more zones, the longer it takes to run each model, and at some point computational resources become limiting. For concrete examples, the Austin model contains about 500 zones, while the Chicago model contains about 1,800. Zones are often related to census tracts, to make it easy to get demographic information from census results.

The last piece of the puzzle is a representation of the transportation infrastructure itself. This is done using a mathematical *network*which consists of *links* and *nodes*. In transportation applications, a link usually represents a means of travel from one point to another: a road segment between two intersections, a bus route between two stops, and so on, as seen in Figure 2. The nodes, in turn, are the endpoints of the links. Quite often, nodes are adjacent to multiple links, so a node representing an intersection may adjoin multiple links representing road segments. Nodes and links may also be more abstract; for instance, nodes may represent zone “centroids” where trips begin and end, and links in a multimodal network might represent a transfer from one transport mode to another. The level of detail in a network varies from application to application. For multistate

[1] In more sophisticated models, the four steps may be repeated again, to ensure that the end results are consistent with the input data. We won’t worry about this in CE 3500.

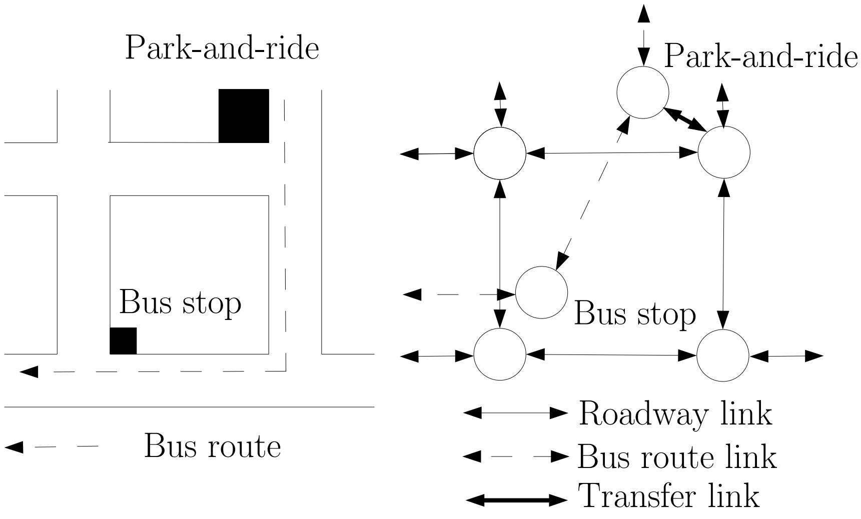


Figure 2: Nodes and links in transportation networks.

freight models, major highways may be the only links, and major cities the only nodes. For a city's planning model, all major and minor arterials may be included as well. For a more detailed model, individual intersections may be "exploded" so that different links represent each turning movement (Figure 3). A network is often compactly written as $G = (N, A)$ where N and A respectively represent the sets of nodes and links. A link is often written in terms of its endpoints, so (i, j) represents a link starting at node i and ending at node j. In the real world, these networks can be quite large; the Chicago network contains nearly 13,000 nodes and 39,000 links!

One final note before launching into the details of the four steps. In each of these four steps, a lot of seemingly unrealistic assumptions will be made. Especially in transportation planning, engineers often balk at the "hand waving" that's done. Fundamentally, this is because all of the questions in transportation planning are questions of human behavior, and unlike a fluid or pin joint, humans do not behave in simple ways that are easy to reproduce. However, this is less worrying than it may seem at first glance. Two things need to be kept in mind. First, predicting aggregate behavior of a large group of people is easier than predicting the specific behavior of one individual. More importantly, if we are to be precise, every model, in every field, is wrong: if it were exactly true, it wouldn't be a model, we'd have the universe in a box! The question is whether a model is close enough to reality that you are capturing the basic behavior well enough to make good decisions, and at this the four-step model has been repeatedly validated. The famous statistician George Box is said to have quipped that "all models are wrong, but some models are useful."

2 Trip Generation

The first of the four steps is trip generation, which takes demographic data for each zone as input, and generates a total number of trips for each zone as an output. Trip generation distinguishes

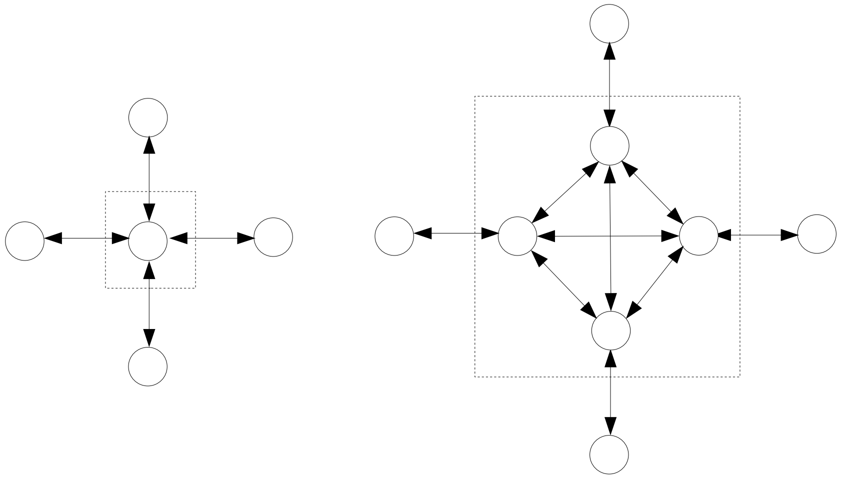


Figure 3: Two representations of the same intersection.

between *productions*, the number of trips made by households, and *attractions*, the places where these households travel. Every trip involves both a production and an attraction: for example, when I travel to work, there is a production at my apartment, and an attraction at UW. When I shop for groceries, there is a production at my apartment, and an attraction at Safeway. Basically, productions occur where people live, and attractions occur at the places people need to travel to. Because every trip involves one of each, the total number of productions must equal the total number of attractions. At the end of trip generation, we can say how many trips are produced by each zone, and how many trips are attracted to each zone.

It is also useful to divide trips into different categories based on the purpose (work, shopping, recreation, etc.), both from the standpoint of planning and from accurate modeling. This way, we can estimate the total number of work trips produced by a zone, the total number of shopping trips attracted to a zone, and so forth.

Both productions and attractions are estimated using linear regression; Section 2.3 provides a brief overview of linear regression and shows how to do a regression using Excel. (If you already know how to do linear regression using another software program, you are free to keep using that program.) For this reason, Sections 2.1 and 2.2 present regression results without detailed derivations in order to keep the focus on the transportation planning applications.

2.1 Productions

To calculate the number of trips produced by each zone, we need two ingredients. First, we need to know how to calculate the number of productions as a function of demographic data. Second, we need to know that zone's demographic data, so we can apply the formula correctly. Planners use travel surveys to accomplish the first of these, and census data to accomplish the second. For

Table 1: Travel survey responses from Neptune City

Household	Income	Cars	Size	Workers	Work trips	Shopping trips
1	61,000	3	1	1	1.1	0.9
2	36,000	1	6	5	4.2	2.3
3	75,000	2	5	1	0.3	2.9
4	60,000	3	4	1	1.0	2.4
5	54,000	1	5	3	2.4	2.1
6	30,000	1	1	1	0.6	0.1
7	62,000	1	5	2	1.9	1.6
8	44,000	1	5	4	3.2	1.7
9	44,000	0	4	3	2.9	1.2
10	54,000	2	1	1	0.9	1.5
11	39,000	1	2	1	0.6	0.9
12	55,000	3	6	3	2.1	2.1
13	35,000	1	6	2	1.8	1.6
14	71,000	1	2	1	0.8	1.5
15	40,000	2	4	2	2.2	1.3
16	58,000	2	3	2	1.3	1.2
17	48,000	1	5	4	3.2	1.9
18	45,000	0	3	1	1.0	1.0
19	48,000	2	1	1	0.7	1.1
20	55,000	2	3	1	0.5	2.0

example, let's say we've conducted a travel survey in the fictitious town of Neptune City and obtained twenty responses (in reality, a much larger sample is used) as shown in Table 1. Each household in the survey has reported their total annual income, the number of cars they own, the number of people in the household, the number of *employed* people in the household, and the average number of work trips and shopping trips made each day during the survey period of two weeks.

We want to create an equation that relates a household's income I, vehicle ownership v, size n, and employment level e to the number of work and shopping trips produced by that household (P_w and P_s, respectively). For simplicity, we will assume linear equations of the form

$$P_w = \beta_0^w + \beta_I^w I + \beta_v^w v + \beta_n^w n + \beta_e^w e \tag{1}$$

$$P_s = \beta_0^s + \beta_I^s I + \beta_v^s v + \beta_n^s n + \beta_e^s e \tag{2}$$

Our goal is to find the β values which best fit the survey data, and linear regression is the appropriate tool for this task. Performing the regression, we obtain the following equations:

$$P_w = 0.30 - (4.8 \times 10^{-6})I - 0.041v - 0.0024n + 0.82e \tag{3}$$

$$P_s = -0.80 + (2.4 \times 10^{-5})I + 0.14v + 0.25n + 0.028e \tag{4}$$

What do these mean? Let's say we have a household with an annual income of $I = 50000$, which owns three vehicles ($v = 3$) and consists of five members ($n = 5$), two of whom are employed

Table 2: Neptune City zone information

Zone	Households	Income	Cars	Size	Workers	Office (ft$^2 \times 10^6$)	Retail (ft$^2 \times 10^6$)
1	23,000	30,000	1.4	2.1	1.4	2	5
2	35,000	25,000	1.8	2.2	1.6	3	15
3	85,000	55,000	2.5	2.3	1.5	10	10
4	15,000	85,000	1.1	1.5	1.3	25	20

Table 3: Neptune City daily zone productions

Zone	Average work trips	Average shopping trips
1	29,000	15,000
2	50,000	22,000
3	100,000	125,000
4	14,000	27,000

($e = 2$). Then substituting these values, we would expect this household to make an average of 1.6 work trips and 0.86 shopping trips each day. Examining the regression equations, we can see the impact of each of these variables. For instance, for shopping trips, $\beta^s_I > 0$ and $\beta^s_n > 0$, indicating that households with a higher income make more shopping trips on average, as do households with more people. Both of these findings are intuitive, but the regression equation quantifies the effect: on average, each additional person in a household results in 0.25 additional shopping trips each day, and an additional \$10,000 in income would increase daily shopping trips by 0.24. For work trips, some of the β^w values are negative, indicating that an increase in the corresponding variable (such as income or vehicle ownership) would result in a *decrease* in the number of work trips made. Perhaps most puzzling is $\beta^w_n < 0$, which at first glance would suggest that larger households make *fewer* work trips, which seems counterintuitive. Can you explain what's going on here?

Creating the regression equations accomplishes the first step. The second is to use each zone's demographic data to estimate the number of trips produced. Neptune City is divided into four zones, as shown in Table 2. For each zone, the table reports the total number of households in that zone, the average income, vehicle ownership, household size, and household employment data, as well as the total office space and retail space (reported in millions of square feet). Using the regression equations, an average household in Zone 1 would make $0.30 - (4.8 \times 10^{-6})(30000) - 0.041(1.4) - 0.0024(2.1) + 0.82(1.4) = 1.3$ work trips and $-0.80 + (2.4 \times 10^{-5})(30000) + 0.14(1.4) + 0.25(2.1) + 0.028(1.4) = 0.65$ shopping trips each day. Multiplying by the number of households (23,000), we expect this zone to produce 29,000 work trips and 15,000 shopping trips on an average day, rounding to two significant figures.[2] (Here's where the "miracle of aggregation" comes in! Even though a single household may make more or less trips on any given day, when we look at all 23,000 households, a lot of these daily fluctuations cancel out.) Substituting values for the other zones, we obtain the total productions in Table 3.

If the city were different, or if the travel survey were different, this procedure could still be fol-

[2] All numbers in this example are rounded to two significant figures *when presented in these notes.* When doing the computations, it's important to keep as many digits as possible. Only round when you are reporting the numbers in final form.

Table 4: Neptune City productions and attractions

Zone	A_w (Raw)	A_s (Raw)	P_w	P_s	A_w (Scaled)	A_s (Scaled)
1	11,000	17,000	29,000	15,000	12,000	17,000
2	15,000	57,000	50,000	22,000	16,000	57,000
3	43,000	37,000	100,000	125,000	48,000	37,000
4	103,000	77,000	14,000	27,000	116,000	78,000
Total	170,000	188,000	192,000	189,000	192,000	189,000

lowed with minor modifications. Perhaps the travel survey collected different information (e.g., the number of children in a household, the ages of the household members, etc.), in which case the regression equations would involve different variables. Even with the same variables, different travel surveys will almost certainly produce different regression equations (that is, different β values — although we hope they're close!). Finally, in a different city, the number of zones and the zonal demographic data is different, which would result in different numbers of productions even with the same regression equation.

2.2 Attractions

Determining the number of trips attracted to each zone is based on the same principle as estimating productions: obtain a regression equation relating zonal characteristics to number of attractions. Collecting this type of data is more difficult than for productions, where travel surveys were enough. To get attraction information, we would either have to cross-reference all of the travel survey records to identify the destination zone (which is time consuming and error-prone) or collect detailed information on all the types of businesses, schools, and so forth in each zone (this is not in the census, so very labor-intensive). **For these reasons, in CE 3500 I will give you the regression equations for attractions directly.** In practice, more detailed methods are available. For example, the Institute of Transportation Engineers publishes a large handbook entitled *Trip Generation*, which provides estimated trip attractions for a wide variety of land use types.

Continuing with the Neptune City example, assume that someone has performed one of the labor-intensive procedures mentioned in the last paragraph, and derived the following estimations for the number of work and shopping attractions (A_w and A_s, respectively).

$$A_w = 2500 + I/3000 + S_o/250 \tag{5}$$

$$A_s = -3500 + I/100 + S_r/250 \tag{6}$$

where I is the average household income in that zone, and S_o and S_r are the total office space and retail space in that zone, measured in square feet. Using the data in Table 2, we can calculate the number of trips attracted to each zone. For instance, zone 1 attracts $2500 + 30000/3000 + (2 \times 10^6)/250 = 11000$ work trips and $-3500 + 30000/100 + (5 \times 10^6)/250 = 17000$ shopping trips each day. Table 4 shows these values in the columns labeled "A_w (Raw)" and "A_s (Raw)."

However, we need to apply one more step. Remember that every trip involves both a production and an attraction, so the total number of work productions must equal the total number of work attractions, and the total number of shopping productions must equal the total number of shopping attractions. There's no guarantee that the linear regression formulas will give numbers consistent with this requirement, and indeed Table 4 reveals this to be the case. The model predicts 170,000 work attractions, as compared to 192,000 work productions, with a similar discrepancy for shopping productions and attractions. In order to get consistent values, we apply a scaling factor to the attractions so they match the productions, while keeping the production values fixed. (This reflects our greater confidence in the production calculations, because the data availability is so much better.) Thus, in this example, we multiply each zone's work attractions by 192/170, and each zone's shopping attractions by 189/188, obtaining the values in the two rightmost columns of Table 4. As shown by the last row in the table, the productions and attractions now match.

2.3 Linear Regression

Linear regression is a statistical technique used to identify a general relationship between variables in a data set. It is very widely used in virtually every scientific field, and many software packages are capable of performing linear regression. SAS, SPSS, and Excel are common commercial software for doing this type of analysis. Many free/open-source options are available as well, including Gnumeric, OpenOffice Calc, and R, and these are often more accurate!

The main concept is to assume that one variable of interest y (the *dependent variable*) is related to other variables $x_1, \ldots, x_n$ (*independent variables*) through a linear equation of the form

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

where $\beta_0, \ldots, \beta_n$ are unknown parameters which need to be estimated from the data. This equation will only be exact if all of the data lie on the same line, which almost never occurs in practice because it is impossible to measure (or even observe) all of the relevant factors. Otherwise, it is not possible to choose β values so that the equation matches all of the data. Instead, linear regression finds a "best-fit" line which does the best possible job given this fact.[3] For example, Figure 4 plots work trips against number of employed household members, using the survey data from Table 1. The linear regression line is also shown on this figure; notice that it does not intersect all of the data points (or any of them, for that matter), but gives a decent approximation.

To do this using Excel, enter all of the data contiguously, as in Figure 5. Then, in a blank space in your spreadsheet, select a horizontal range of cells. The number of cells to select is the number of independent variables (one for each of the β_i), plus one more for the constant β_0. For this example, there are four independent variables (income, vehicle ownership, household size, and employment level), so you select five cells horizontally. You will use the `LINEST` formula to do linear regression, which takes four arguments: (1) the range of data corresponding to the dependent variable (2) the range of data corresponding to the independent variables, (3) whether or not to include a constant β_0 (always enter `TRUE`), and (4) whether to return additional statistical data (it's OK to

[3] Readers interested in the formulas and mathematical details are referred to an elementary statistics textbook or website such as `http://mathworld.wolfram.com/LeastSquaresFitting.html`

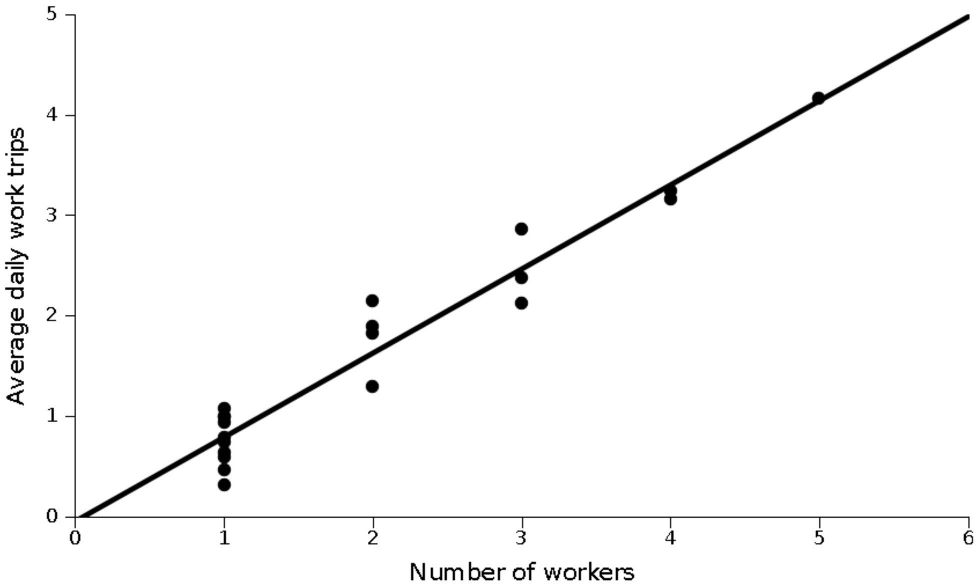


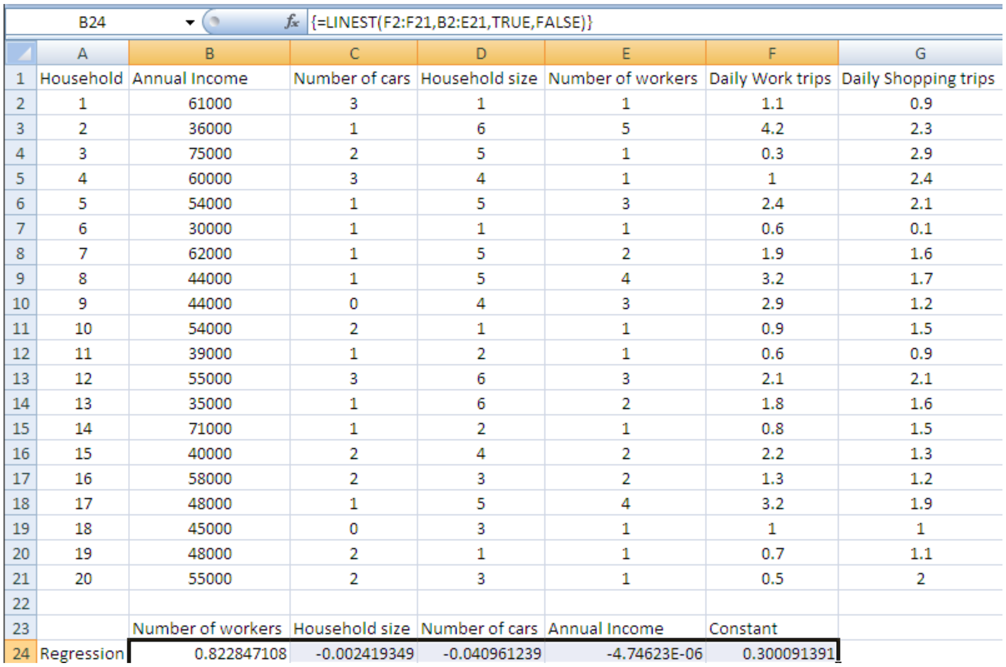
Figure 4: Illustration of linear regression.

enter `FALSE`). For this example, the formula is `=LINEST(F2:F21,B2:E21,TRUE,FALSE)`. **Once you type the formula, press CTRL+SHIFT+ENTER. Simply pressing ENTER will not perform the complete regression.** You should see values appear in the range you selected. Unfortunately, Excel is a bit confusing in the order in which it gives the β values.

1. The *rightmost* column is always the constant β_0.
2. The remaining columns correspond to the independent variables *in the reverse order*, that is, $\beta_n, \beta_{n-1}, \ldots, \beta_1$.

The labels in Figure 5 show this correspondence.

As with any other statistical procedure, certain assumptions are required for linear regression to be valid. In particular, we have to assume that any one household's deviation from the overall trend is unrelated to any other household (that is, the fact that I make more trips on average is completely unrelated to anything my neighbor does), and that the impact of the dependent variables is linear within the range of the data set (that is, an additional $1,000 in income would affect my tripmaking the same way, regardless of whether my old income was $10,000 or $100,000). Both of these assumptions have been found to be reasonable in transportation planning, although researchers have also developed more sophisticated models which do not require these assumptions.



B24 {=LINEST(F2:F21,B2:E21,TRUE,FALSE)}

	A	B	C	D	E	F	G
1	Household	Annual Income	Number of cars	Household size	Number of workers	Daily Work trips	Daily Shopping trips
2	1	61000	3	1	1	1.1	0.9
3	2	36000	1	6	5	4.2	2.3
4	3	75000	2	5	1	0.3	2.9
5	4	60000	3	4	1	1	2.4
6	5	54000	1	5	3	2.4	2.1
7	6	30000	1	1	1	0.6	0.1
8	7	62000	1	5	2	1.9	1.6
9	8	44000	1	5	4	3.2	1.7
10	9	44000	0	4	3	2.9	1.2
11	10	54000	2	1	1	0.9	1.5
12	11	39000	1	2	1	0.6	0.9
13	12	55000	3	6	3	2.1	2.1
14	13	35000	1	6	2	1.8	1.6
15	14	71000	1	2	1	0.8	1.5
16	15	40000	2	4	2	2.2	1.3
17	16	58000	2	3	2	1.3	1.2
18	17	48000	1	5	4	3.2	1.9
19	18	45000	0	3	1	1	1
20	19	48000	2	1	1	0.7	1.1
21	20	55000	2	3	1	0.5	2
22							
23		Number of workers	Household size	Number of cars	Annual Income	Constant	
24	Regression	0.822847108	-0.002419349	-0.040961239	-4.74623E-06	0.300091391	

Figure 5: Conducting linear regression in Excel.

3 Trip Distribution

The second step is trip distribution, which takes the total number of zonal trips as input, and returns the origins and destinations of these trips as output. The end result is an *origin-destination matrix* or *OD matrix*, which shows the total number of trips departing each origin to each destination. One way of thinking about this is converting productions and attractions into specific origins and destinations. In the previous section, we estimated that zone 3 in Neptune City would produce 100,000 work trips — but where are these work trips going? How many are going to zone 1, how many are going to zone 2, and so forth. Similarly, we estimated that zone 3 would attract 77,000 work trips. How many of these trips are coming from each zone? Answering these questions is the goal of trip distribution.

Usually, several OD matrices are estimated for different times of day: it is common to have an OD matrix for the morning peak period, for the evening peak period, and for the off-peak period. This reflects the fact that **one "trip" may involve more than one journey.** For instance, I may leave for work in the morning, so I travel from my apartment (the origin) to my office (the destination). But when I return home in the evening, there is a second journey, from my office (the origin for this journey) to my apartment (the new destination). These two journeys have a different origin and destination, so the morning peak OD matrix and evening peak OD matrix will be different. To reflect this, let S_i^t and E_i^t represent the number of trips starting and ending at zone i during time period t (we'll use the abbreviations AM, PM, and OP for the morning peak, evening peak, and off-peak). So S_1^{AM} is the number of trips starting at zone 1 during the morning peak, E_3^{PM} is the number of trips ending at zone 3 during the evening peak, and so forth.

In this class, we'll make the following assumptions. A work trip results in one journey from the production to the attraction during the morning peak, and one journey from the attraction back to the production in the evening peak. A shopping trip results in both a journey from the production to the attraction, and the return from the attraction to the production, during the off-peak period. That is, $S_i^{AM} = E_i^{PM} = P_i^w$, $E_i^{AM} = S_i^{PM} = A_i^w$, and $S_i^{OP} = E_i^{OP} = P_i^s + A_i^s$. In reality, the situation is not so simple, and planners use more sophisticated techniques to distribute trips throughout the day.

Let's say we're given two zones i and j, we know the number of starting and ending trips S_i and E_j, and want to know how many trips are going between these two zones. Using our intuition, we make several reasonable assumptions: (1) the more trips starting at zone i, the more trips from i to j; (2) the more trips ending at zone j, the more trips from i to j; (3) the farther away the two zones, the fewer the trips between them. So, given the average length L_{ij} of a trip between zone i and j, we calculate a *friction factor* $\phi(L_{ij})$ using a decreasing function ϕ. (For instance, $\phi(L_{ij}) = 1/L_{ij}$ or $\phi(L_{ij}) = e^{-L_{ij}}$. As a starting point, we'll say that the number of trips d_{ij} between i and j is proportional to P_i, A_j, and $\phi(L_{ij})$:

$$d_{ij} \propto S_i E_j \phi(L_{ij})$$

This is the simplest model which satisfies these assumptions, and is called a *gravity model* based on an analogy with Newton's law of gravitation: the heavier the mass of two bodies (i.e., the magnitude of the productions and attractions), the greater the force betwen them; the farther apart they are, the lesser the force. In practice, we have to add an adjustment factor μ_j for each zone so that the table "balances" properly (the reason for this become clear in the example that follows):

$$d_{ij} = C_i \mu_j S_i E_j \phi(L_{ij})$$

where C_i is the proportionality constant. These constants are chosen so that the total number of trips from each origin is correct. That is, we need

$$\sum_j C_i \mu_j S_i E_j \phi(L_{ij}) = S_i$$

so

$$C_i = \frac{1}{\sum_j \mu_j E_j \phi(L_{ij})}$$

and

$$d_{ij} = \frac{\mu_j S_i E_j \phi(L_{ij})}{\sum_j \mu_j E_j \phi(L_{ij})} \tag{7}$$

The general process for estimating each of the three OD matrices is as follows:

1. Calculate the number of trips starting and ending at each zone (**S** and **E**) during the current time period.

2. Set the initial adjustment factors $\mu_j = 1$ for all zones j.

3. Create an OD matrix using the gravity model equation (7).

Table 5: Interzonal distances in Neptune City.

	1	2	3	4
1	5	15	15	25
2	15	5	25	15
3	15	25	5	15
4	25	15	15	5

Table 6: Friction factors in Neptune City.

	1	2	3	4
1	0.20	0.067	0.067	0.040
2	0.067	0.20	0.040	0.067
3	0.067	0.040	0.20	0.067
4	0.040	0.067	0.067	0.20

4. If the table is balanced, we are done. Otherwise, create new adjustment factors and repeat the last step.

The first step is explained above and the second step is self-explanatory. As an example, let's create the AM peak OD matrix for Neptune City. The first step is to calculate $\mathbf{S}^{\mathbf{AM}}$ and $\mathbf{E}^{\mathbf{AM}}$ using the productions and attractions data from Table 4; we get $\mathbf{S}^{\mathbf{AM}} = \begin{bmatrix} 29,000 & 50,000 & 100,000 & 14,000 \end{bmatrix}$ and $\mathbf{E}^{\mathbf{AM}} = \begin{bmatrix} 12,000 & 16,000 & 48,000 & 116,000 \end{bmatrix}$. We also initialize $\boldsymbol{\mu} = \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix}$ as the second step.

We also need to calculate the friction factors; if Table 5 shows the distance between each pair of zones, and the friction factor is given by the simple relation $\phi(L_{ij}) = 1/L_{ij}$, then we can tabulate the friction factors as in Table 6

In the third step, we create an OD matrix one row (origin) at a time using the formula (7) and the vectors $\mathbf{S}$, $\mathbf{E}$, and $\boldsymbol{\mu}$, producing the following table:

	1	2	3	4	Total
1	6100	2800	8200	12,000	29,000
2	2900	12,000	7000	28,000	50,000
3	4200	3500	51,000	41,000	100,000
4	230	540	1600	11,000	14,000
Total	13,000	19,000	68,000	92,000	

(Using spreadsheet software will greatly simplify the number of calculations which have to be done!)

You'll notice that the sum of each row is correct (and forms the vector $\mathbf{S}$), but the sum of each column is not correct: the column sums should match $\mathbf{E} = \begin{bmatrix} 12,000 & 16,000 & 48,000 & 116,000 \end{bmatrix}$. Here's where the adjustment factors μ_j come into play. We will set each μ_j so that the destinations which do not receive enough trips become more attractive, and so that the destinations which receive too many trips become less attractive. Specifically, we multiply each component of $\begin{bmatrix}\mu_j\end{bmatrix}$ by $E_j/\sum_i d_{ij}$, so in this case $\boldsymbol{\mu} = \begin{bmatrix}0.89 & 0.88 & 0.71 & 1.25\end{bmatrix}$ (that is, in reality zone 1 only attracts 89% of the trips assigned in the first matrix, zone 2 only attracts 88%, and so on).

Using the new $\boldsymbol{\mu}$ values, we apply (7) again to construct a new OD matrix:

	1	2	3	4	Total
1	5400	2500	5900	15,000	29,000
2	2400	9800	4600	33,000	50,000
3	3900	3200	38,000	54,000	100,000
4	180	400	960	12,000	14,000
Total	12,000	16,000	50,000	115,000	

This is closer, but not quite there. So we update $\boldsymbol{\mu}$ again, multiplying each component by $E_j/\sum_i d_{ij} = \begin{bmatrix}0.99 & 1.02 & 0.97 & 1.01\end{bmatrix}$ to obtain $\boldsymbol{\mu} = \begin{bmatrix}0.88 & 0.90 & 0.69 & 1.27\end{bmatrix}$ For instance, to get the first component we multiply 0.89 (the value of μ from the last iteration) by 0.99, to get the second component we multiply 0.88 by 1.02, and so forth. Again applying (7), we obtain

	1	2	3	4	Total
1	5400	2600	5700	15,000	29,000
2	2400	10,000	4400	33,000	50,000
3	3900	3400	37,000	55,000	100,000
4	180	420	920	12,000	14,000
Total	12,000	16,000	48,000	116,000	

and now the trip ends correctly match. Thus, we now know the number of people traveling from every zone to every other zone during the morning peak period. Clearly zone 4 receives the bulk of the trips, because it contains the most office space. Note that we predict a certain level of "intrazonal" trips as well: 5000 trips will both start and end in zone 1, 9500 will start and end in zone 2, and so forth. These represent people whose workplaces are located within the same zone as their residence.

Repeating the same procedure, one can find the offpeak and PM peak OD matrices to be

	1	2	3	4	Total
1	11,000	9100	6800	6700	34,000
2	9100	68,000	10,000	28,000	114,000
3	6800	10,000	37,000	20,000	75,000
4	6700	27,500	20,000	101,000	156,000
Total	34,000	114,000	75,000	156,000	

and

	1	2	3	4	Total
1	5400	2400	3900	180	12,000
2	2600	10,000	3400	420	16,000
3	5700	4400	37,000	920	48,000
4	15,000	33,000	55,000	12,000	116,000
Total	29,000	50,000	100,000	14,000	

and respectively. Note that the PM peak OD matrix is the transpose of the AM peak OD matrix... can you explain why?

4 Mode Choice

The third step uses the OD matrix from trip distribution as a starting point, and divides the trips according to the mode of transportation. Usually modes correspond to different vehicle types: driving a car, riding the bus, riding a train, riding a bicycle, and so forth. Walking can also be considered a mode, and different vehicle modes can be subdivided further: driving alone vs. carpooling; taking an express bus vs. a local one; taking a subway train vs. light rail, etc.

The foundations of mode choice are *utility theory* and *discrete choice*, two conceptual frameworks developed by economists during the last century. The central ideas are (1) a person traveling from one zone to another must choose between one of several competing modes; (2) each choice results in a certain level of *utility* for the traveler (reflecting his or her satisfaction or happiness with traveling via that mode); and (3) this utility is closely related to characteristics of the mode (travel time, out-of-pocket cost) and to characteristics of the traveler (income, vehicle ownership, and so forth).

Utility is an abstract concept, and can represent virtually anything that would cause a person to choose one option over another. Different components of the utility function may include the travel time associated with a given mode (possibly separated into in-vehicle and out-of-vehicle travel time, if someone has to wait for a bus or walk from a parking lot), the monetary cost (transit fare, gas expenses, roadway toll, etc.), the reliability of the mode, the number of vehicles available to a household, their income, and so on. A linear form is usually used: $U_m = \beta_m + \sum_i \beta_{m,i} x_{m,i} + \epsilon$ where U_m is the utility of mode m, $x_{m,i}$ is the i-th attribute of mode m (travel time, etc.), and $\beta_{m,i}$ is a coefficient showing how important that attribute is. There are other, less tangible factors associated with each mode (for instance, some people prefer to drive alone because of the control they have, while others prefer to take transit so they don't have to worry about driving, or so they can read or do work while commuting), so even if driving and taking the bus were to take exactly the same amount of time and cost exactly the same amount, you would still see preferences for one or the other. These factors are captured in an attribute-specific constant β_m. ϵ represents an "unobserved" or random component of utility. No matter how thorough we are, there will still be factors which are not included in the model. People can also be inconsistent in their choices from day to day. Both of these ideas are encapsulated in the ϵ term.

The ϵ can be something of a problem: we can't know its value because, by definition, it represents things which we are not modeling. So, it's useful to rewrite just the known part of the utility function: $V_m = \beta_m + \sum_i \beta_{m,i} x_{m,i}$ so $U_m = V_m + \epsilon$, because V is something we can actually work with. Don't worry, though, ϵ will end up playing a role in the end. We can call U the total utility and V the known part of the utility.

The β values are estimated using the maximum likelihood technique, which bears some similarity to linear regression in that it takes survey data as input, and returns the β values which fit the surveys as closely as possible. We won't be covering maximum likelihood in this class, so the β values will be given to you. As a concrete example, if I is household income in thousands of dollars, t is travel time in minutes, and c is cost in dollars, we might have the following equations:

$$V_{car} = 1 + 0.003I - 0.04t_{car} - 0.24c_{car} \tag{8}$$

$$V_{bus} = -3 - 0.001I - 0.04t_{bus} - 0.24c_{bus} \tag{9}$$

Notice the signs of the β coefficients: as income increases, the utility for driving a car increases and for taking the bus decreases (perhaps representing social pressure, that "the bus is for poor people"[4]). As travel time or out-of-pocket cost increases, the utility for both modes decreases, which makes sense: the longer my commute, or the more expensive, the less happy I am and the lower my utility.

Now, given the known part of the utilities V_{car} and V_{bus} of two different modes, we need to say what proportion of travelers will choose each mode (P_{car} and P_{bus}). At the very least, our formulas for P_{car} and P_{bus} should satisfy the following properties:

1. Every traveler should choose one of the two modes, that is, $P_{car} + P_{bus} = 1$.
2. The greater the utility of either mode, the more people will use it; that is, P_{car} is increasing in V_{car}, and P_{bus} is increasing in V_{bus}.
3. No matter what, at least some people will be using both modes, that is, $P_{car} > 0$ and $P_{bus} > 0$. (This reflects the unknkown ϵ term.)

One simple equation which satisfies all of these properties is the logistic curve

$$P_{car} = \frac{e^{V_{car}}}{e^{V_{car}} + e^{V_{bus}}} \qquad P_{car} = \frac{e^{V_{bus}}}{e^{V_{car}} + e^{V_{bus}}} \tag{10}$$

Clearly the two add up to one; clearly as the utility of each mode increases, it will grab a larger share of travelers; and because e^x is always strictly greater than zero, both proportions will be positive.

So, let's figure out the mode split for Neptune City during the morning peak hour. We need to be given the zone-to-zone travel times and travel costs by car and by bus; these are shown in Tables 7 and 8.[5]

[4] Whether such a perception *should* exist or not is irrelevant for mode choice; the fact is that it does exist, and does impact mode choice behavior.

[5] Do the travel costs by car seem high to you? By the time you account for insurance, depreciation, maintenance, accident risk, and fuel consumption, the cost is much higher than you would think. The official federal government figure is around 55 cents per mile.

Table 7: Zone-to-zone travel times, minutes

Auto	1	2	3	4
1	5	13	17	33
2	13	5	25	15
3	17	25	5	14
4	33	15	14	5

Bus	1	2	3	4
1	13	20	35	45
2	20	7	55	23
3	35	55	10	30
4	45	23	30	5

Table 8: Zone-to-zone travel costs, dollars

Car	1	2	3	4
1	2.75	8.25	8.25	13.75
2	8.25	2.75	13.75	8.25
3	8.25	13.75	2.75	8.25
4	13.75	8.25	8.25	2.75

Bus	1	2	3	4
1	0.75	0.75	0.75	1.25
2	0.75	0.75	1.25	0.75
3	0.75	1.25	0.75	0.75
4	1.25	0.75	0.75	0.75

With this information, we can calculate the utilities using the equations (8) and (9). For each origin and destination, we substitute the origin zone's average income (from Table 2), and the zone-to-zone travel times and costs, resulting in the known utilities shown in Table 9. Once we know these, we can substitute into equations (10), to get the proportions of people that will use each mode (Table 10). Finally, we can multiply the proportion by the total number of people traveling between each pair of zones to get the mode split (Table 11).

A few "idiot checks" worth mentioning: the utilities for driving are higher than for the bus, so the mode split should be tilted in this direction. The proportions for bus and driving add up to one; and the total number of people traveling between each OD pair is equal to the sum of the people driving and taking the bus.

5 Route Choice

5.1 Shortest Path Assumption and User Equilibrium

The fourth and final step is route choice: taking the OD matrix from mode choice, and assigning those trips onto specific routes. For example, if I know that 1,000 people per day travel between Laramie and Ft. Collins, route choice would tell me how many would use US-287, and how many

Table 9: Zone-to-zone known parts of utility, by mode

Auto	1	2	3	4
1	0.23	−1.41	−1.57	−3.53
2	−1.42	0.22	−3.22	−1.50
3	−1.50	−3.13	0.30	−1.38
4	−3.37	−1.32	−1.28	0.40

Bus	1	2	3	4
1	−3.73	−4.01	−4.61	−5.13
2	−4.00	−3.48	−5.52	−4.12
3	−4.64	−5.56	−3.64	−4.43
4	−5.18	−4.18	−4.47	−3.47

Table 10: Zone-to-zone mode proportions

Auto	1	2	3	4
1	0.98	0.93	0.95	0.83
2	0.93	0.98	0.91	0.93
3	0.96	0.92	0.98	0.96
4	0.86	0.95	0.96	0.98

Bus	1	2	3	4
1	0.02	0.07	0.05	0.17
2	0.07	0.02	0.09	0.07
3	0.04	0.08	0.02	0.04
4	0.14	0.05	0.04	0.02

Table 11: Zone-to-zone OD matrices, by mode

Auto	1	2	3	4
1	5900	2600	7800	9700
2	2600	12,000	6300	26,000
3	4000	3200	50,000	39,000
4	200	510	1500	11,000

Bus	1	2	3	4
1	110	190	370	2000
2	200	290	630	1900
3	170	280	970	1800
4	33	29	63	240

would use I-80/I-25 through Cheyenne. If we wanted, we could repeat the "utility function" approach that we used for mode choice, and when there are only a few options this would work well. However, in a major metropolitan area there are literally thousands of potential routes that connect most origins and destinations. (Imagine a grid network, and all the combinations of turns that can get you from one point to another.) For this reason, transportation planners use a simpler approach: **assume that all travelers choose the route that lets them reach the destination as quickly as possible.** This is sometimes called the *shortest path assumption.*

For example, if you have to choose between two routes, one of which takes ten minutes and the other fifteen, you would always opt for the first one. If you are the only one traveling, this is all well and good. The situation becomes more complicated, however, if there are others traveling. If there are ten thousand people making the same choice, and all ten thousand pick the first route, congestion will form and the travel time will rapidly increase. If this were to happen, some people would switch from the first route to the second route, because the first would no longer be faster.

Congestion effects are represented using a *delay function* $t(x)$, which gives the time required to travel on a roadway segment as a function of the number of people wanting to use it x. Examples of different delay functions are shown in Figure 6. They are usually increasing (or at the least nondecreasing — the more people on a road, the more congested it will be) and convex (the more congested the roadway, the greater the impact of one more vehicle). The most common delay function was developed by the Bureau of Public Roads (BPR) in 1964, and uses four parameters: the "free flow" travel time t_0, the roadway capacity c, and two shape parameters α and β which are fit to observed data:

$$t(x) = t_0 \left(1 + \alpha \left(\frac{x}{c} \right)^\beta \right) \tag{11}$$

The free flow travel time is the time required to drive assuming that you are on the only vehicle on the road (as you can verify, $t(0) = t_0$). The capacity represents the maximum number of people who can travel on the roadway during the study period. It's worth noting that (1) the travel time begins to increase even when $x < c$, and (2) nothing prevents x from exceeding the capacity. The first observation reflects the fact that average speeds begin to slow even before capacity is reached,

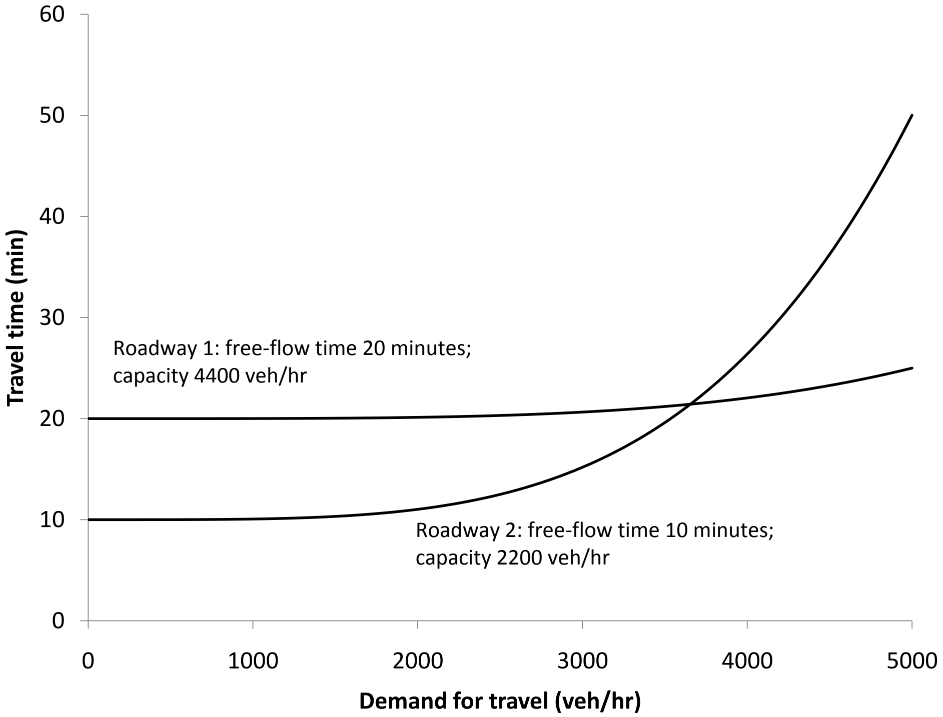


Figure 6: Two delay functions for different roadways.

as it becomes more difficult to pass slow drivers and drive at your desired speed. The second reflects the fact that x is the *demand* for travel on the roadway, not the actual number of people who can pass through during the study period. If $x > c$, then more people want to use the road than capacity allows, and congestion will definitely form. Typically, $\alpha = 0.15$ and $\beta = 4$ are good choices for these parameters. Figure 6 shows two different BPR functions with different free-flow times and different capacities. Note that the travel time increases very rapidly when $x > c$.

So, given two routes with delay functions $t_1(x_1)$ and $t_2(x_2)$, and a total of x travelers who have to pick between these two routes, the shortest path assumption forces one of the following three scenarios to be true:

- Route 1 is faster even if all x people choose it (that is, $t_1(x) < t_2(0)$). Then $x_1 = x$ and $x_2 = 0$.
- Route 2 is faster even if everyone chooses it (that is, $t_2(x) < t_1(0)$). Then $x_2 = x$ and $x_1 = 0$.
- Most commonly, neither route dominates the other. Then people will choose routes so the travel times are *equal.* That is, x_1 and x_2 satisfy $t_1(x_1) = t_2(x_2)$, with $x_1 + x_2 = x$.

Because the third case is most common, this basic route choice model is called *user equilibrium*: the two routes are in equilibrium with each other. Why must the travel times be equal? If the first route was faster than the second, people would switch from the second to the first. This would decrease the travel time on the second route, and increase the travel time on the first, and people would continue switching until they were equal. The reverse is true as well: if the second route were faster, people would switch from the first route to the second, decreasing the travel time on the first route and increasing the travel time on the second. The *only* outcome where nobody has any reason to change their decision, is if the travel times are equal on both routes.

This is important enough to state again: at equilibrium, **every used route connecting an origin and destination has *equal* and *minimal* travel time**. Unused routes may have a higher travel time, and used routes connecting different origins and destinations may have different travel times.

Let's take a concrete example: Figure 7 shows 6000 travelers traveling from zone 1 to zone 2 during one hour, and choosing between the two routes mentioned above: route 1, with free-flow time 20 minutes and capacity 4400 veh/hr, and route 2, with free-flow time 10 minutes and capacity 2200 veh/hr. That means we have

$$t_1(x_1) = 20\left(1 + 0.15\left(\frac{x_1}{4400}\right)^4\right) \tag{12}$$

$$t_2(x_2) = 10\left(1 + 0.15\left(\frac{x_2}{2200}\right)^4\right) \tag{13}$$

We need to choose x_1 and x_2 so that $t_1(x_1) = t_2(x_2)$ (equilibrium) and $x_1 + x_2 = 7000$. Substituting $x_2 = 7000 - x_1$ into the second delay function, the equilibrium equation becomes

$$20\left(1 + 0.15\left(\frac{x_1}{4400}\right)^4\right) = 10\left(1 + 0.15\left(\frac{7000 - x_1}{2200}\right)^4\right) \tag{14}$$

Using an equation solver (or guess-and-check, or Newton's method) we find that equilibrium occurs for $x_1 = 3560$, so $x_2 = 7000 - 3560 = 2440$, and $t_1(x_1) = t_2(x_2) = 20.3$ minutes. Alternately, we can use a graphical approach. Figure 8 plots the travel time on *both* routes as a function of the flow on route 1 (because if we know the flow on route 1, we also know the flow on route 2). The point where they intersect is the equilibrium: $x_1 = 3560$, $t_1 = t_2 = 20.3$.

In more complicated networks, writing all of the equilibrium equations and solving them simultaneously is much too difficult. Instead, a more systematic approach is taken, described in the next subsection.

5.2 General Framework for Larger Problems

In a large network, there is no solution methods which gives you the right answer immediately. That is, there isn't any "step one, step two, step three, and then we're done" recipe for solving large-scale equilibrium problems[6]. Instead, an iterative approach is used where we start with some assignment of drivers to routes and links, and move closer and closer to the equilibrium solution

[6] If you can think of one, please let me know. You'll win the Nobel prize in economics. I'm not exaggerating.

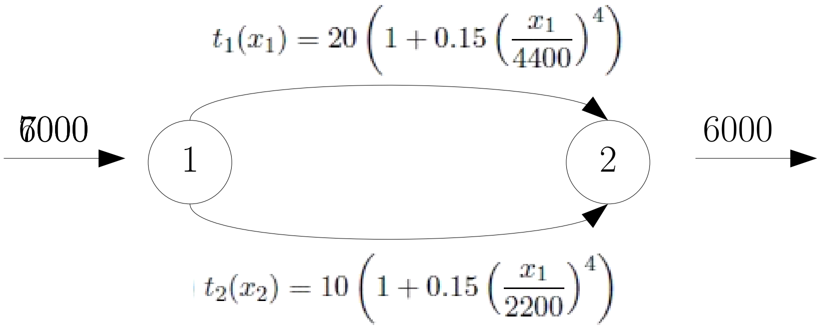

$t_1(x_1) = 20\left(1 + 0.15\left(\frac{x_1}{4400}\right)^4\right)$
6000
1
2
6000
$t_2(x_2) = 10\left(1 + 0.15\left(\frac{x_1}{2200}\right)^4\right)$

Figure 7: Small example using the two links.

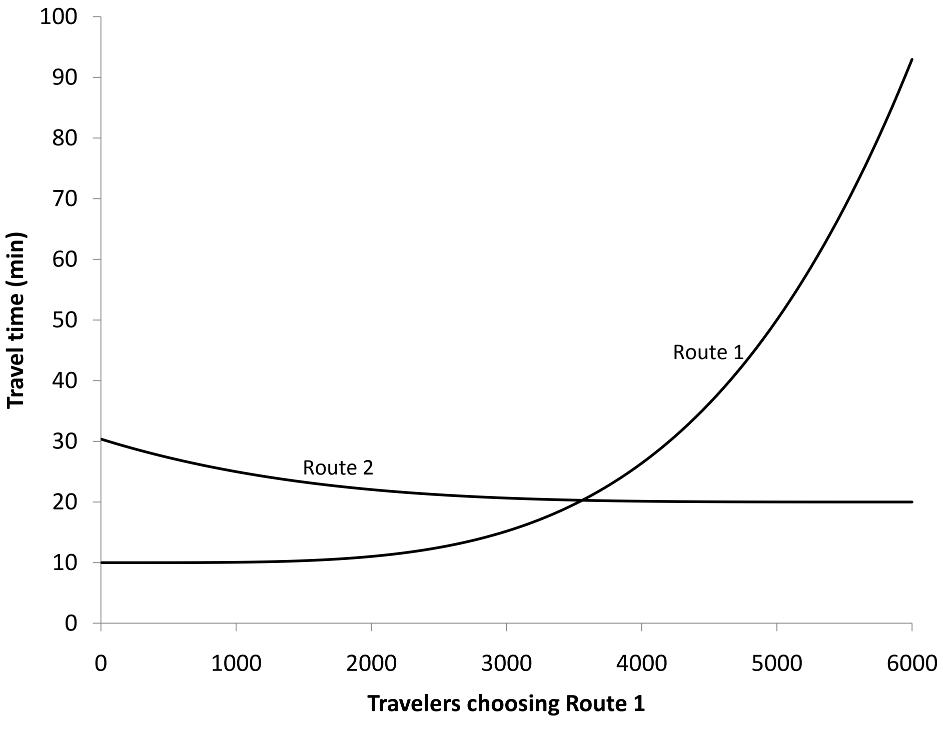


Figure 8: The equilibrium point lies at the intersection of the delay functions.

as you repeat a certain set of steps over and over, until you're "close enough" to quit and call it good.[7]

Broadly speaking, all equilibrium solution algorithms repeat the following three steps:

1. Based on current link volumes, calculate the travel times. If we're close to equilibrium, stop; otherwise, go to step 2.
2. Find the fastest path between each origin and each destination.
3. Shift travelers from slower paths to faster ones, and return to step one.

The first step is straightforward, and is nothing more than evaluating the delay functions on each link with the current flows. The second step isn't too difficult either — the problems we'll be solving in class are small enough that you can spot the shortest path by inspection, and in larger real-world networks there are relatively fast computer algorithms for finding shortest paths. The third step requires the most care; the danger here is shifting either too few travelers onto faster paths, or shifting too many. If we shift too few, then it will take a long time to get to the equilibrium solution. On the other hand, systematically shifting too many can be even more dangerous, because it creates the possibility of "infinite cycling" and never finding the true equilibrium.

Recall the example in Figure 7, where the equilibrium is for 3560 travelers to choose the top route, and 2440 to choose the bottom route, with an equal travel time of 20.3 minutes on both paths. Solving this example using the above process, initially (i.e., with nobody on the network) the travel times on top and on bottom are 20 minutes and 10 minutes, respectively. The fastest path is the bottom one one (step two), so let's assign all 6000 travelers onto the bottom path (step three). Returning to the first step, we recalculate the travel times as 20 minutes on the top link, and 93 minutes on the bottom. This is not at all an equilibrium, so we go back to the first step, and see that the bottom path is now faster, so we have to shift some people from the top to the bottom. If we wanted, we could shift travelers one at a time, that is, assigning 1 to the top route and 5999 to the bottom, seeing that we still haven't found equilibrium, so trying 2 and 5998, then 3 and 5997, and so forth, until finally reaching the equilibrium with 3560 and 2440. Clearly this is not efficient, and is an example of shifting too few travelers at a time.

At the other extreme, let's say we shift *everybody* onto the fastest path in the third step. That is, we go from assigning 0 to the top route and 6000 to the bottom, to assigning 6000 to the top and 0 to the bottom. Recalculating link travel times, the top route now has a travel time of 30.4 minutes, and the bottom a travel time of 10. Repeating the process, we try to fix this by shifting everybody back (0 on top, 6000 on bottom), but now we're just back in the original situation. If we kept up this process, we'd keep bouncing back and forth between these solutions. This is even worse than shifting too few, because we never reach the equilibrium no matter how long we work! You might think it's obvious to detect if something like this is happening. With this small example, it might be. Trying to train a computer to detect this, or trying to detect cycles with over 2 million OD pairs (as in Chicago), is much much harder.

[7] One iterative algorithm you probably saw in calculus was Newton's method for finding zeroes of a function. Repeat the same step over and over until the function is sufficiently close to zero.

5.3 Method of Successive Averages

Although the method of successive averages (MSA) is not competitive with other equilibrium solution algorithms, its simplicity and clarity in applying the three-step iterative process make it an ideal starting place. The first and second steps of MSA operate the same as in all other equilibrium algorithms, so this section and all following ones focus only on step three: once you've found the shortest paths, how do you decide how many travelers to shift onto these, and how many stay on their current paths? As shown above, there are problems if you shift too few travelers, and potentially even bigger problems if you shift too many. MSA adopts a reasonable middle ground: initially, we shift a lot of travelers, but as the algorithim progresses, we shift fewer and fewer until we settle down on the average. The hope is that this avoids both the problems of shifting too few (at first, we're taking big steps, so hopefully we get somewhere close to equilibrium quickly) and of shifting too many (eventually, we'll only be moving small amounts of flow so there is no worry of infinite cycling).

Specifically, on the i-th iteration, MSA shifts $1/i$ of the travelers onto the shortest paths. So, the first time through the three steps, *everybody* is assigned to shortest paths. The second time through, half of the people stay on their current paths and half shift to the new shortest paths. On the third iteration, a third of the people shift to new paths, and two thirds stay on their old paths, and so forth. A complete description of MSA is as follows; in these steps, $\mathbf{x^i}$ is the vector of link flows after the i-th iteration of MSA.

1. Set the iteration counter $i = 0$.
2. (Re)-calculate the link travel times.
3. Find the shortest path between each origin and destination.
4. Shift travelers onto shortest paths:
   (a) Find the link flows if everybody were traveling on the shortest paths found in step 1, store these in $\mathbf{x}^*$.
   (b) If this is the zero-th iteration, $\mathbf{x^0} = \mathbf{x}^*$. Otherwise, $\mathbf{x^i} = (1/i)\mathbf{x}^* + (1 - 1/i)\mathbf{x^{i-1}}$.
5. Decide if we are close enough to equilibrium to stop. If not, increase the iteration counter i by one and return to step 1.

Here's how MSA would work using that same example. Without any vehicles (free-flow conditions), route 1 is faster ($10 < 20$), so the initial link flows are $\mathbf{x^1} = \begin{bmatrix} x_1^1 & x_2^1 \end{bmatrix} = \begin{bmatrix} 6000 & 0 \end{bmatrix}$. With these flows, the travel time on route 1 is $t_1(6000) = 92.9$ and the travel time on route 2 is $t_2(0) = 20$. At this point, the fastest route is route 2, so $\mathbf{x}^* = \begin{bmatrix} 0 & 6000 \end{bmatrix}$ and $\mathbf{x^1} = 1/2 \begin{bmatrix} 0 & 6000 \end{bmatrix} + 1/2 \begin{bmatrix} 6000 & 0 \end{bmatrix} = \begin{bmatrix} 3000 & 3000 \end{bmatrix}$. Now the travel times are $t_1(3000) = 15.2$ and $t_2(3000) = 20.6$, so the fastest route is route 1, $\mathbf{x}^* = \begin{bmatrix} 6000 & 0 \end{bmatrix}$, and $\mathbf{x^1} = 1/3 \begin{bmatrix} 6000 & 0 \end{bmatrix} + 2/3 \begin{bmatrix} 3000 & 3000 \end{bmatrix} = \begin{bmatrix} 4000 & 2000 \end{bmatrix}$. Table 12 shows the progress of subsequent iterations; at iteration 5, the difference in travel times between the two routes is only 0.5 minutes (30 seconds!), a difference of about 2%, so we stop the algorithm, and have the (approximate) equilibrium solution $x_1 = 3600$, $x_2 = 2400$.

Table 12: Method of successive averages demonstration.

Iteration	x_1	x_2	t_1	t_2	x_1^*	x_2^*
0	0	0	10.00	20.00	6000	0
1	6000	0	92.99	20.00	0	6000
2	3000	3000	15.19	20.65	6000	0
3	4000	2000	26.39	20.13	0	6000
4	3000	3000	15.19	20.65	6000	0
5	**3600**	**2400**	**20.76**	**20.27**		

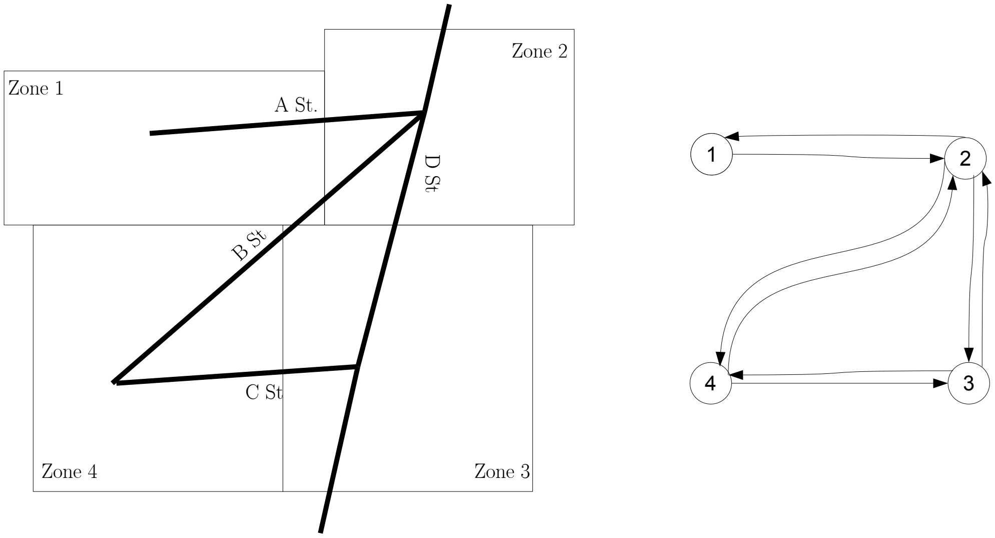


Figure 9: Zone map of Neptune City, and associated network representation.

Now, let's look at a more realistic example. Figure 9 shows a map of the major streets in Neptune City, along with the network representation. The left part of Table 11 tells us the number of driving trips from each origin to each destination. **Intrazonal trips (i.e., trips starting and ending at the same zone) are assumed to use minor roads, and are ignored during the route choice step.** Table 13 gives information on each roadway link (identified by the zone where the link starts and ends). Note that two links have been created for each roadway, representing travel in both directions. (If there was a one-way street, there would be a link only in one direction.)

Now, we can start MSA. With no vehicles at first, the travel times from the first step are simply the free-flow times. The second step is to find the fastest path from every zone to every other zone, using the free-flow times. These are shown in Table 14. To find the flow on each link, see which zones use it as a fastest path. For instance, link (1,2) is part of the fastest path from zone 1 to zone 2, from zone 1 to zone 3, and from zone 1 to zone 4; so, consulting Table 11, its initial volume is $2600 + 7800 + 9700 = 20,100$. Or, for link (2,3), it is part of the fastest paths from zone 1 to zone 3, and from zone 2 to zone 3, so its initial volume is $7800 + 6300 = 14,100$. These are shown as x^0 in Table 15, along with the travel times which correspond to these link volumes (using the

Table 13: Link data for Neptune City.

Street	Link ID	Free-flow time	Capacity
A St. (eastbound)	(1,2)	10	18,000
A St. (westbound)	(2,1)	10	18,000
B St. (eastbound)	(4,2)	20	14,000
B St. (westbound)	(2,4)	20	14,000
C St. (eastbound)	(4,3)	15	44,000
C St. (westbound)	(3,4)	15	44,000
D St. (northbound)	(3,2)	15	8,000
D St. (southbound)	(2,3)	15	8,000

Table 14: Fastest paths with free-flow travel times.

	1	2	3	4
1	—	1-2	1-2-3	1-2-4
2	2-1	—	2-3	2-4
3	3-2-1	3-2	—	3-4
4	4-2-1	4-2	4-3	—

standard BPR function).

For the next step, we see if any of the fastest paths have changed with the new travel times. We see that for those traveling from zone 1 to zone 4, the new fastest path is 1-2-3-4, rather than 1-2-4; and for those traveling from zone 2 to zone 4, the new fastest path is 2-3-4 instead of 2-4. Both of these replaced link $(2,4)$ with the path $(2,3),(3,4)$, because the latter has a travel time of 53 minutes, as opposed to 149 minutes on $(2,4)$. Knowing this, we find the new x^* by again seeing which fastest paths use which link. The only links which are different are (2,4) (which is now not used by *any* fastest path, so its volume is 0), and links (2,3) and (3,4) (which now carry vehicles traveling from zone 1 to zone 4, and from zone 2 to zone 4, in addition to the previous load). So $(2,3)$, being used by travelers from zone 1 to zone 3, from zone 1 to zone 4, from zone 2 to zone 3, and from zone 2 to zone 4, now has volume $7800 + 6300 + 9700 + 26,000 = 50,000$. The new flows x^1 are the average of x^0 and x^*.

Again, we calculate the new travel times with respect to x^1, and see that the shortest paths between zones 1 and 4, and zones 2 and 4, have changed back: 1-2-4 is now faster than 1-2-3-4, and 2-4 is faster than 2-3-4 (see boldfaced values in Table 11). Thus, x^* is the same as x^0 (because the current fastest paths are the same as the fastest paths at free flow), and the new flows x^2 are the weighted average of x^1 and x^*, with a weight of 2/3 on x^1 and 1/3 on x^*. This process continues for additional iterations, as shown in the table. After six iterations, the difference between paths 2-4 and 2-3-4 (the boldfaced values) is under two percent, and we terminate.

Table 15: MSA for Neptune City.

Iteration	(1,2)	(2,1)	(2,4)	(4,2)	(2,3)	(3,2)	(3,4)	(4,3)	(2,3)+(3,4)
$\mathbf{x^0}$	20122	6869	35852	711	14081	7179	38914	1515	
$\mathbf{t}(\mathbf{x^0})$	12.34	10.03	**149.02**	20.00	36.59	16.46	16.38	15.00	**52.97**
$\mathbf{x}^*$	20122	6869	0	711	49933	7179	74765	1515	
$\mathbf{x^1}$	20122	6869	17926	711	32007	7179	56839	1515	
$\mathbf{t}(\mathbf{x^1})$	12.34	10.03	**28.06**	20.00	591.48	16.46	21.27	15.00	**612.75**
$\mathbf{x}^*$	20096	6045	42937	528	9348	6755	52181	866	
$\mathbf{x^2}$	20122	6869	23901	711	26031	7179	50864	1515	
$\mathbf{t}(\mathbf{x^2})$	12.34	10.03	**45.49**	20.00	267.24	16.46	19.02	15.00	**286.26**
$\mathbf{x}^*$	20096	6045	42937	528	9348	6755	52181	866	
$\mathbf{x^3}$	20122	6869	26889	711	23044	7179	47876	1515	
$\mathbf{t}(\mathbf{x^3})$	12.34	10.03	**60.82**	20.00	169.89	16.46	18.15	15.00	**188.05**
$\mathbf{x}^*$	20096	6045	42937	528	9348	6755	52181	866	
$\mathbf{x^4}$	20122	6869	28681	711	21251	7179	46084	1515	
$\mathbf{t}(\mathbf{x^4})$	12.34	10.03	**72.85**	20.00	127.04	16.46	17.71	15.00	**144.74**
$\mathbf{x}^*$	20096	6045	42937	528	9348	6755	52181	866	
$\mathbf{x^5}$	20122	6869	30730	711	19203	7179	44035	1515	
$\mathbf{t}(\mathbf{x^5})$	12.34	10.03	**89.64**	20.00	89.69	16.46	17.26	15.00	**106.95**
$\mathbf{x}^*$	20096	6045	42937	528	9348	6755	52181	866	
$\mathbf{x^6}$	20122	6869	31370	711	18562	7179	43395	1515	
$\mathbf{t}(\mathbf{x^6})$	12.34	10.03	**95.63**	20.00	80.22	16.46	17.13	15.00	**97.34**

5.4 Frank-Wolfe

One of the biggest drawbacks with MSA is that it has a fixed step size (or, more informally, a "dumb" step size). Iteration i moves exactly $1/i$ of the travelers onto the new shortest paths, no matter how close or far away we are from the equilibrium. Essentially, MSA decides its course of action before it even gets started, then sticks stubbornly to the plan of moving $1/i$ travelers each iteration. The Frank-Wolfe (FW) algorithm fixes this problem by using an *adaptive* step size. At each iteration, FW calculates exactly the right amount of flow to shift to get as close to equilibrium as possible.

So, at each iteration we calculate the new flows with the equation $\mathbf{x^i} = \lambda \mathbf{x}^* + (1-\lambda)\mathbf{x^{i-1}}$. With MSA we always chose $\lambda = 1/i$, but with FW λ is chosen adaptively. The extreme values $\lambda = 0$ and $\lambda = 1$ mean we keep everybody on the current path, or shift everybody to the shortest path, respectively. We want to pick λ in this range in such a way that $\mathbf{x^i}$ is as close to equilibrium is possible. The key to doing this is to move just enough people that both the old flows $\mathbf{x^{i-1}}$ and the target flows $\mathbf{x}^*$ are "balanced," that is, with the updated costs $\mathbf{t}(\mathbf{x^i})$, both $\mathbf{x^{i-1}}$ and $\mathbf{x}^*$ are equally good.

More precisely, we need $\sum_{a \in A} x_a^{i-1} t_a(x_a^i) = \sum_{a \in A} x_a^* t_a(x_a^i)$. (Study this equation carefully: $\mathbf{x^{i-1}}$ is the *old* vector of flows from the last iteration; $\mathbf{x^i}$ is the *new* vector of flows after we shift λ of drivers to the shortest paths $\mathbf{x}^*$.) What happens if the two sides of this equation are not equal? Let's say $\sum_{a \in A} x_a^{i-1} t_a(x_a^i) > \sum_{a \in A} x_a^* t_a(x_a^i)$. This essentially means that the old flows $\mathbf{x^{i-1}}$ have too high of a cost, and people want to shift towards $\mathbf{x}^*$, which has lower cost — in other words, we haven't shifted enough people to the shortest paths, and we need a bigger λ. On the other hand, if $\sum_{a \in A} x_a^{i-1} t_a(x_a^i) < \sum_{a \in A} x_a^* t_a(x_a^i)$, then we've shifted too many people, and the old flows $\mathbf{x^{i-1}}$ look better than the "target" flows $\mathbf{x}^*$. Only if the two sums are exactly equal do drivers have no incentive to move either back towards $\mathbf{x^{i-1}}$ (smaller λ) or closer to $\mathbf{x}^*$ (bigger λ).

Because the difference $\sum_{a \in A} x_a^{i-1} t_a(x_a^i) - \sum_{a \in A} x_a^* t_a(x_a^i)$ is strictly decreasing, it is easy to find the right λ value either using an equation solver (like that in Excel) or by guess-and-check in any spreadsheet program. The steps for FW can be described as follows:

1. Set the iteration counter $i = 0$.
2. (Re)-calculate the link travel times.
3. Find the shortest path between each origin and destination.
4. Shift travelers onto shortest paths:
   (a) Find the link flows if everybody were traveling on the shortest paths found in step 1, store these in $\mathbf{x}^*$.
   (b) If this is the zero-th iteration, $\mathbf{x^0} = \mathbf{x}^*$. Otherwise, $\mathbf{x^i} = \lambda \mathbf{x}^* + (1-\lambda)\mathbf{x^{i-1}}$ where λ is chosen so that

$$\sum_{a \in A} x_a^{i-1} t_a(\lambda \mathbf{x}^* + (1-\lambda)\mathbf{x^{i-1}}) = \sum_{a \in A} x_a^* t_a(\lambda \mathbf{x}^* + (1-\lambda)\mathbf{x^{i-1}})$$

5. Decide if we are close enough to equilibrium to stop. If not, increase the iteration counter i by one and return to step 1.

Returning to the Neptune City example, we re-do route choice using FW instead of MSA. The first iteration proceeds in the same way: identify the shortest paths and assign all trips to these (the row labeled $\mathbf{x^0}$ in Table 16). $\mathbf{t}(\mathbf{x^0})$ and $\mathbf{x}^*$ are again the same as in MSA. However, choosing $\mathbf{x^1}$ is done differently. Recall that in MSA we simply averaged $\mathbf{x^0}$ and $\mathbf{x}^*$. In FW, we seek to find a weighted average, with the weight chosen more cleverly to balance the costs of $\mathbf{x^0}$ and $\mathbf{x}^*$. This means we need to choose λ such that $\sum_{a \in A} x_a^0 t_a(\lambda \mathbf{x}^* + (1-\lambda)\mathbf{x^0}) = \sum_{a \in A} x_a^* t_a(\lambda \mathbf{x}^* + (1-\lambda)\mathbf{x^0})$. Using an equation solver (Excel or a graphing calculator), we find that $\lambda = 0.123$ accomplishes exactly this. To wit:

1. You may verify that $0.123\mathbf{x^0} + (1-0.123)\mathbf{x}^*$ provides the row $\mathbf{x^1}$ in Table 16.
2. Likewise, you can easily check that the row $\mathbf{t}(\mathbf{x^1})$ results from substituting the values $\mathbf{x^1}$ into the BPR functions with parameters given in Table 13.
3. Finally, we need to check that the FW "balancing" condition is satisfied. Multiply the element of $\mathbf{x^0}$ for each arc by the corresponding element of $\mathbf{t}(\mathbf{x^1})$, and add them together:

$$\begin{aligned} &20122 \times 12.34 + 6869 \times 10.03 + 35852 \times 96.33 + 711 \times 20.00 + \\ &\qquad + 14081 \times 79.21 + 7179 \times 16.46 + 38914 \times 17.11 + 1515 \times 15.00 = 5707268 \end{aligned}$$

   Repeating the same with $\mathbf{x}^*$ and $\mathbf{t}(\mathbf{x^1})$ gives

$$\begin{aligned} &20122 \times 12.34 + 6869 \times 10.03 + 0 \times 96.33 + 711 \times 20.00 + \\ &\qquad + 49933 \times 79.21 + 7179 \times 16.46 + 74765 \times 17.11 + 1515 \times 15.00 = 5707268 \end{aligned}$$

   These two are equal, so the flows are "balanced."

Notice that we have found the exact equilibrium after only one step! The travel times on paths 2-4 and 2-3-4 have exactly the same travel time, so no travelers have an incentive to switch routes. In larger networks, FW cannot find the equilibrium in one step [8], but FW is faster than MSA in almost all instances.

5.5 Stopping Criteria

A general issue is how one chooses to stop the iterative process, that is, how one knows when a solution is "good enough" or close enough to equilibrium. This is called a *convergence criterion.* Many convergence criteria have been proposed over the years; perhaps the most common is the

[8] In fact there are much faster methods for complicated networks which, unfortunately, are themselves more complicated and thus beyond the scope of this course.

Table 16: FW for Neptune City.

Iteration	(1,2)	(2,1)	(2,4)	(4,2)	(2,3)	(3,2)	(3,4)	(4,3)	(2,3)+(3,4)
$\mathbf{x^0}$	20122	6869	35852	711	14081	7179	38914	1515	
$\mathbf{t(x^0)}$	12.34	10.03	**149.02**	20.00	36.59	16.46	16.38	15.00	**52.97**
$\mathbf{x}^*$	20122	6869	0	711	49933	7179	74765	1515	
$\mathbf{x^1}$	20122	6869	31442	711	18490	7179	43323	1515	
$\mathbf{t(x^1)}$	12.34	10.03	**96.33**	20.00	79.21	16.46	17.11	15.00	**96.33**

relative gap, which is defined here. If we let u^{rs} represent the time spent on the fastest path between origin r and destination s, the relative gap γ is commonly defined as follows:

$$\gamma = \frac{\sum_{a \in \mathcal{A}} t_a x_a}{\sum_{(r,s) \in \mathcal{D}} u^{rs} d^{rs}} - 1 = \frac{\mathbf{t} \cdot \mathbf{x}}{\mathbf{u} \cdot \mathbf{d}} - 1 \tag{15}$$

where d^{rs} is the number of trips from r to s.

Notice that the numerator of the fraction is the total time spent by everybody traveling – this is called the total system travel time (TSTT) and we'll see it again shortly. The relative gap is always nonnegative, and it is equal to zero if and only if the flows x_a satisfy the principle of user equilibrium.[9] It is these properties which make the relative gap a useful convergence criterion: once it is close enough to zero, our solution is "close enough" to equilibrium. For most practical purposes, a relative gap of 10^{-6} is small enough. Unless you're using specialized software, it may take a long time to obtain a gap this small, so for homeworks and course projects in this class, repeating the above steps five times (that is, five iterations of calculating travel times, finding shortest paths, and shifting flows) is enough.

6 Conclusion

OK, so we've just spent two or three weeks on the four-step model. What's the point? The four-step model is usually used by a metropolitan planning organization (MPO), a government entity responsible for transportation planning in cities.[10] One of the requirements of an MPO is to generate a long-range transportation plan (LRTP) and a transportation improvement program (TIP), outlining a plan for meeting the transportation needs of their city given budget limitations. Many MPOs use the four-step model to identify which projects will be most cost-effective, through the following process:

1. Forecast demographic data into the future (5-year time horizon for TIP, 20+ years for LRTP).
2. Run the four-step model to get "do-nothing" conditions. (What will things look like in 5, 20, etc. years if absolutely nothing at all is done?)

[9] It is worthwhile studying this equation closely until you understand these properties perfectly clearly.

[10] A city with more than 50,000 residents is required to have an MPO in order to receive federal funding for transportation projects.

3. Identify several different improvements or policy options (more lanes, new roadways, expanded bus system, toll roads, etc.) which fit within the budget.
4. For each of these options, run the four-step model again, and measure the system conditions.
5. Identify the most cost-effective combination of improvement projects, schedule them in the TIP/LRTP, and begin the process of implementation.

"Measure the system conditions" in step 4 will be different depending on your MPO's goals and objectives, and the output of the four-step model can be plugged into air quality models, economic models, and a variety of other evaluation tools. Perhaps the most universal goal is to reduce congestion. An easy way to measure the total level of congestion is to calculate the *total system travel time $TSTT$*:

$$TSTT = \sum x_{ij} t_{ij}(x_{ij})$$

for every roadway segment (i, j) in the network. In the previous example, the total system travel time can be calculated as follows:

(1,2)	20,100 veh	×	12.3 min =	248,000 veh-min
(2,1)	6900 veh	×	10.0 min =	68,000 veh-min
(2,4)	31,400 veh	×	95.6 min =	3,000,000 veh-min
(4,2)	710 veh	×	20.0 min =	14,000 veh-min
(2,3)	18,600 veh	×	80.2 min =	1,489,000 veh-min
(3,2)	7200 veh	×	16.5 min =	118,000 veh-min
(3,4)	43,000 veh	×	17.1 min =	743,000 veh-min
(4,3)	1500 veh	×	15.0 min =	23,000 veh-min
			TOTAL:	5,705,000 veh-min

or, in more convenient units, 95,000 vehicle-hours. This represents the total amount of time spent driving by all interzonal travelers during the morning peak period.

One potential improvement might be to increase the capacity on link (2,4) from 14,000 vehicles to 18,000 vehicles (perhaps by building an additional lane, or by optimizing the signal timing). To see the impact of this, run the four-step model again with the higher capacity on this link. The roadway capacity only plays a role in route choice, so that is the only step that needs to be repeated. Without going through the details, with this change, the equilibrium route flow on (2,4) would increase to 34,700 vehicles, and the flow on (2,3) and (3,4) would decrease to 15,300 vehicles and 40,100 vehicles, respectively; and likewise, the travel times on (2,4), (2,3), and (3,4) would change to 61.3 minutes, 44.8 minutes, and 16.6 minutes, respectively. With these changes, the new total system travel time is

(1,2)	20,100 veh	×	12.3 min =	248,000 veh-min
(2,1)	6900 veh	×	10.0 min =	68,000 veh-min
(2,4)	34,700 veh	×	61.3 min =	2,126,000 veh-min
(4,2)	710 veh	×	20.0 min =	14,000 veh-min
(2,3)	15,300 veh	×	44.8 min =	683,000 veh-min
(3,2)	7200 veh	×	16.5 min =	118,000 veh-min
(3,4)	40,100 veh	×	16.6 min =	664,000 veh-min
(4,3)	1500 veh	×	15.0 min =	23,000 veh-min
			TOTAL:	3,945,000 veh-min

or 66,000 vehicle-hours. So, the effect of this change would be to decrease the total time spent traveling during the morning peak by approximately 29,000 vehicle-hours, a substantial savings. A similar analysis could be done to calculate the benefits during the evening peak and off-peak hours as well; adding these up would give the total benefits of the project. These benefits can them be compared with the cost of the project, and with the costs and benefits of other options, in order to pick the best improvements available.

A few miscellaneous comments on the four-step model to serve as closure for this section:

- You may have noticed that the travel times or distances used for trip distribution and mode choice may not match those which come out of route choice. If they're pretty far out of alignment, it's good practice to repeat the four-step model again with the more realistic travel times, and iterate until they agree. (However, even though it's good practice, a lot of cities fail to do this.)

- Another good practice (rarely followed) is to consider multiple future scenarios. Because it's so hard to predict travel patterns, demographics, etc. into the future, it is a good idea to generate several different scenarios. For example, four scenarios might be "growth continues the same as in the last 20 years," "there is an economic boom and growth is faster than usual," "there is an extended recession and growth is very slow," and "new technologies reduce the need for people to travel to work." Different options can then be compared under all of the scenarios, identifying options that will work well under many different possible futures.