CIV ENG 471-1
Transportation Systems Analysis
Department of Civil and Environmental Engineering
Northwestern University
Winter 2006

Urban Travel Forecasting

Course Notes

© 2006

by

David Boyce

Adjunct Professor
Department of Civil and Environmental Engineering
Northwestern University
Evanston, Illinois

2149 Grey Avenue
Evanston, Illinois 60201 USA
d-boyce@northwestern.edu
Phone: 847-570-9501
Fax: 775-587-2308

Minor revisions – November 2006

Preface

These Course Notes stem from teaching graduate courses in urban travel forecasting at the University of Pennsylvania, 1968-1976, the University of Illinois at Urbana-Champaign, 1977-1987, and the University of Illinois at Chicago, 1988-2002. The notes began to take their present form during courses taught in 2003 at the National University of Singapore, the University of Canterbury, Christchurch, New Zealand, and the University of Pennsylvania, Philadelphia. Subsequent additions and revisions of the notes were made in teaching courses at Northwestern University in 2005 and 2006.

I gratefully acknowledge my ongoing collaboration with Hillel Bar-Gera, Ben-Gurion University of the Negev, Israel, for his many contributions to my understanding of combined models of travel and route choice. I thank the following students for their assistance in preparing the current version of these Notes: Nan Liu, National University of Singapore, Jeff Casello, University of Pennsylvania, and Vincent Bernardin, Jr., Northwestern University. Biljana Dekic prepared the GIS maps in Section 3. Qian Xiong and Frederik Noeth, University of Illinois at Chicago, contributed related materials. I also thank the graduate students in these courses, too numerous to mention, for their comments, criticisms and patience with my efforts.

I am grateful for the opportunity to teach my urban travel forecasting course during the period 2003-2006, which was facilitated by the following colleagues: Professor Der-Horng Lee, National University of Singapore; Professor Alan Nicholson, University of Canterbury; Professor Vukan Vuchic, University of Pennsylvania; and Professor Joseph Schofer, Northwestern University.

These Course Notes may be used for research and teaching purposes with proper attribution of the source. Comments and criticisms are most welcome, as the revision and expansion of these notes is an ongoing activity.

David Boyce
April 2006

Table of Contents

1. Urban Travel Forecasting – The General Problem

We begin with a rather general statement of the problem and requirements for an urban travel forecasting method for a large, congested urban area. In an effort to address the core issues facing transportation planners today and in the future, we largely ignore past practice and the constraints imposed by computers and incomplete understanding of the problem. Although our intent is to be reasonably comprehensive, some aspects will be ignored or omitted.

1.1 Requirements and Context

We accept the following ground rules for our model in order to conform to current and recent practice:

1. The urban region is divided into small, relatively homogeneous areas called *zones*. The zone size depends on the density of development, so that activity levels per zone are relatively similar.

2. Urban activities in zones are described in terms of the following types of variables:
   a. residential population and households;
   b. employment
   c. education (primary, secondary, higher) and day care;
   d. shopping and personal and business services;
   e. recreation.

3. Facilities for urban activities are described in terms of land and building area:
   a. housing;
   b. workplaces;
   c. schools and colleges;
   d. shopping and services;
   e. recreational facilities: parks, stadiums, theaters. etc.

4. Travel occurs on two types of transportation modes:
   a. *private* vehicle/roadway/traffic control system;
   b. *public* transit/roadway or railway/operations plan

   In addition, various types of trucks use the roadway system. These transportation systems are represented as *networks* consisting of *nodes*, *links*, link attributes, and in the case of transit, routes of scheduled services.

   The service characteristics of links, which are one type of link attribute, depend upon a number of <u>fixed</u> variables:
   a. physical roadway and vehicle characteristics: length, number and width of lanes, grades, etc., and transit station spacing and vehicle performance;
   b. control and operations plans: speed limits, signal settings, road tolls; service frequencies and operating speeds, transit fares;

Other variables related to travel flows (demand) also determine service characteristics:

c. vehicle flows in the case of private vehicles, and number of boardings and alightings at stops per unit time, in the case of public transit.

Taken together, these variables determine the *performance characteristics* of an individual link. In general, they are of the form:

link travel time = f(<u>vehicle and roadway characteristics</u>, <u>operations plan</u>, flows),

where the underlining denotes that these variables are fixed.

Such performance functions are often confused with *supply* functions. For a supply function, in contrast, the vehicle and roadway characteristics, and the operations plan are not fixed, but rather are decision variables of the operator or supplier. In travel forecasts, variations in supplier are typically not represented. For example, adjustments to signal timing and transit service frequencies are not generally made in response to the forecasted travel choices.

5. Travel between daily activities (residing, working, eating, shopping, schooling, recreating) may be described in terms of pairs of activities linked by traveling; for example:
   a. home-work;
   b. work-eat meal;
   c. work-shop;
   d. shop-home;

   Over the 24-hour weekday, travel described in terms of several activities make up a sequence of trips for various purposes, or *trip chain*. Naturally, the duration of the activity, and the time required for travel, determine the geographic range of activities visited. Therefore, travel may be described in terms of individual trips, called trip-based, or sequences of trips and activities, called activity-based, or sometimes trip chaining. Most of these notes address trip-based models, but trip chain models are described in Section 5.

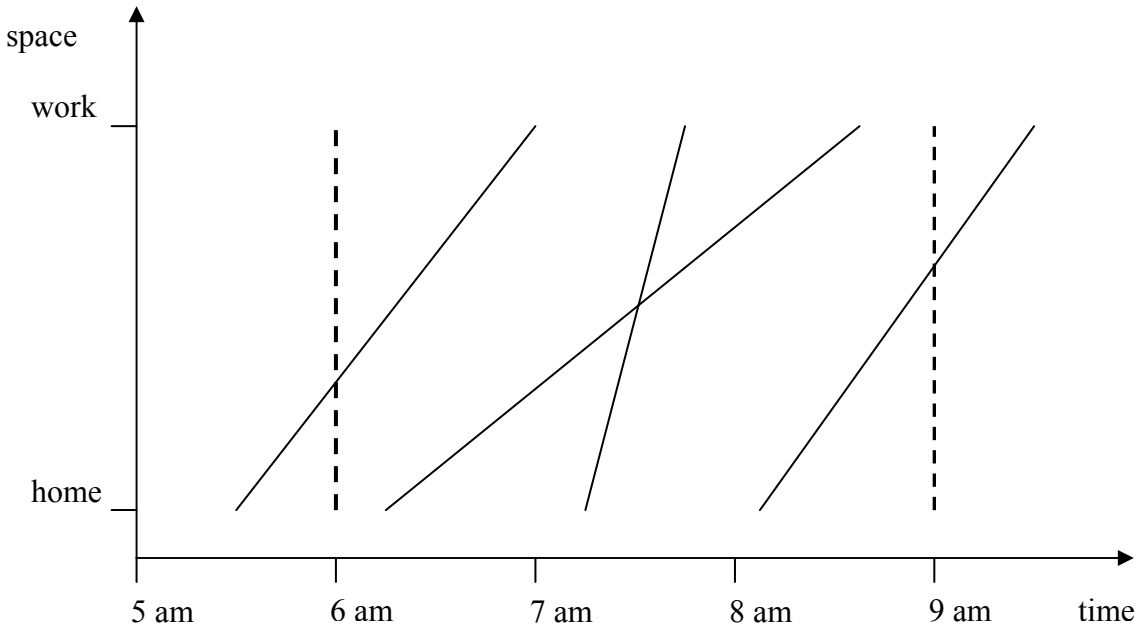
6. Whether travel occurs by private car, either alone or with others, or by public transit, depends on the availability of each mode, and their relative service times and monetary costs, as well as less tangible factors like comfort and convenience.

7. The timing of travel during the day also depends on constraints imposed by activity schedules, and the travel conditions on the private and public networks.

8. Travel may be assumed to occur during a given period of the 24-hour day, such as the morning peak commuting period. To convert actual trips, with their specific departure and arrival times, into *flows* (persons/unit time), we need a conversion procedure, such as one of the following:
   a. all travelers departing from home and going to work during 6-9 am are counted as flows in persons/hour;
   b. all travelers arriving at work from home during 6-9 am are counted as flows in persons/hour;

c. all travelers in the network between 6 and 9 am are counted as whole or partial flows in persons/hour.

The following space-time diagram seeks to illustrate the problem. As will become clear, either explicitly or implicitly when we examine specific models, the variables of interest are flows, not trips or trip chains.



In this example, 3 travelers depart during 6-9 am; 3 travelers arrive during 6-9 am, but they are different travelers. About 3.4 travelers are in the network from 6-9 am. The different slopes of the trajectories denote that home and work are separated differently for each traveler, and also that they have different speeds in the network.

Flows can be compiled from household travel surveys based on these various definitions. These issues arise from seeking to represent a dynamic phenomenon as a static one.

From the viewpoint of transportation system planning, facility design, operations planning and conforming with air quality regulations, we require the forecasts of the following variables for a given transportation system/activity pattern scenario:

1. flows of private cars, trucks and transit vehicles on the road network for the most congested periods of the 24-hour weekday (the morning and evening peak commuting periods) and for other longer periods during which travel conditions are reasonably stable;

2. flows of persons on the public transit network by service type (e.g., bus, rapid transit, commuter rail) by time period of the 24-hour weekday;

3. flows of persons from origin zone to destination zone by private and public modes and time period, and by route, if available and pertinent.

A capability to examine changes in these flows in response to changes in network capacity and service attributes, monetary costs (e.g., fuel, tolls, fares, parking fees), and changes in zonal activity levels is also implied.

1.2 General Formulation of the Urban Travel Forecasting Problem

Consider the following definitions and notation:

d_{pqmr} person flow (trips/hour) from origin zone p to destination zone q on route r of mode m
R_p number of persons residing in zone p
W_q number of employees working in zone q
E_q number of persons attending school or college in zone q
S_q level of shopping and service activity in zone q
c_{pqmr} generalized cost of travel from zone p to zone q on route r of mode m
$\delta^a_{pqmr} = 1$, if link a belongs to route r of mode m from zone p to zone q;
0, otherwise
f_a flow of vehicles per hour on link a
$c_a(f_a)$ generalized cost of travel per vehicle on link a at flow f_a
L_m set of links in the network of mode m
M set of modes
R_{pqm} set of routes connecting zone p to zone q by mode m
η_m occupancy of mode m in persons per vehicle

The general problem we wish to solve may be stated as follows:

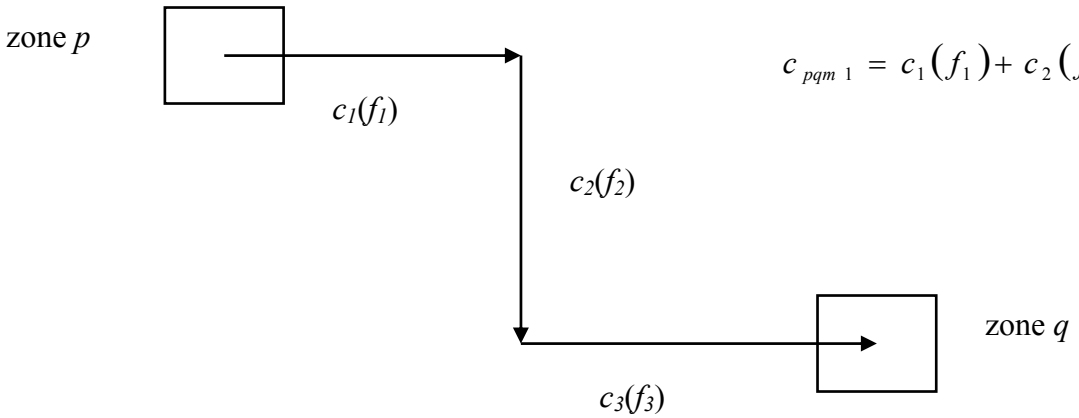
$$d_{pqmr} = f(R_p, W_q, E_q, S_q, ..., c_{pqmr}), \quad \text{for all } p, q, m, r$$

$$c_{pqmr} \equiv \sum_{a \in L_m} c_a(f_a) \cdot \delta^a_{pqmr}, \quad \text{for all } p, q, m, r$$

$$f_a \equiv \frac{1}{\eta_m} \sum_{pqm} \sum_{r \in R_{pqm}} d_{pqmr} \cdot \delta^a_{pqmr}, \quad \text{for all } a \in L_m$$

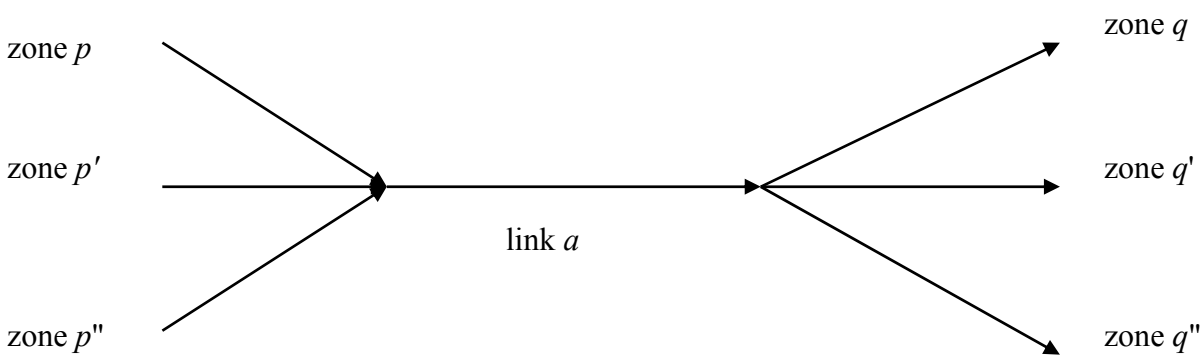
Note that the last two expressions are definitions, which are illustrated by the following figures.

Definition of Interzonal Travel Cost



$$c_{pqm\,1} = c_1(f_1) + c_2(f_2) + c_3(f_3)$$

Definition of Link Flow



$$f_a = d_{pq} + d_{p'q} + d_{p''q} + d_{pq'} + \cdots + d_{pq''} + d_{p'q''} + d_{p''q''}$$

Replacing these variables by their definitions, we obtain:

$$d_{pqmr} = f\left[R_p, E_q, S_q, \ldots, \sum_{a \in L_m} c_a \left(\frac{1}{\eta_m} \sum_{pqm} \sum_{r \in R_{pqm}} d_{pqmr} \cdot \delta^a_{pqmr} \right) \cdot \delta^a_{pqmr}\right]$$

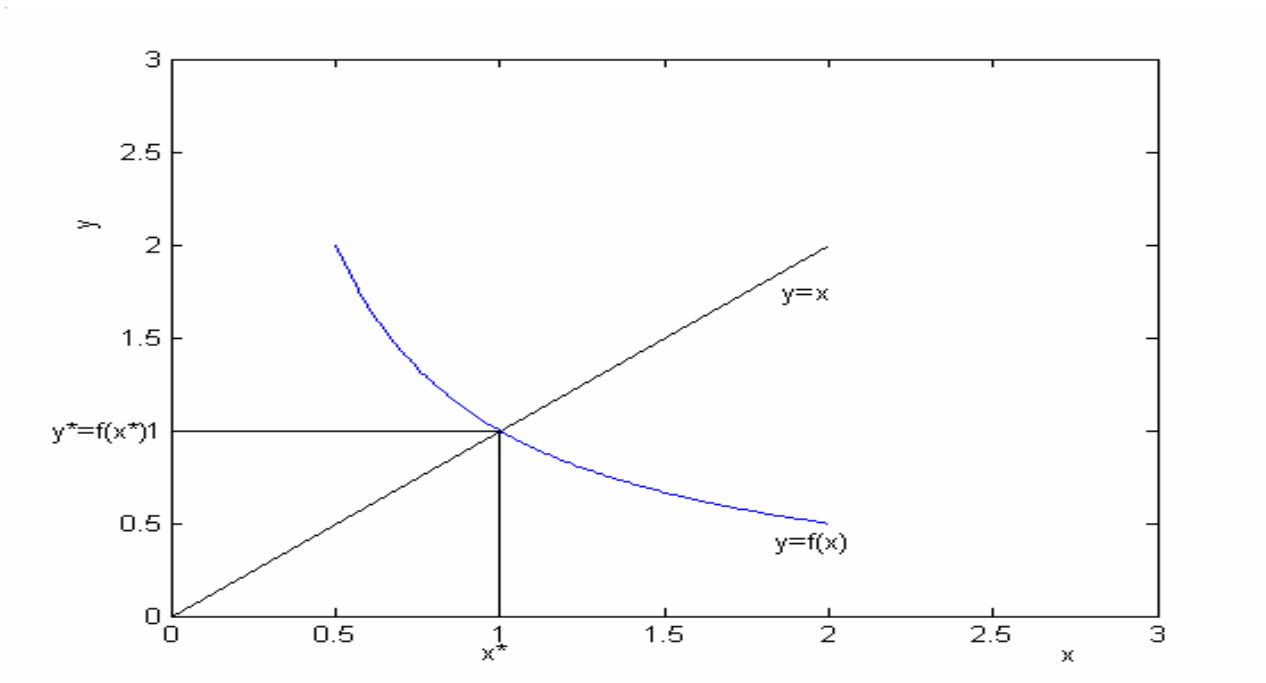
Therefore, d_{pqmr} is a function of itself through its accumulated effect on link flows, and therefore on link costs. Hence, the problem we seek to solve has the structure of a *fixed point problem*.

1.3 Fixed Point Problem

The fixed point problem is to find for the function $f(x)$ a value of

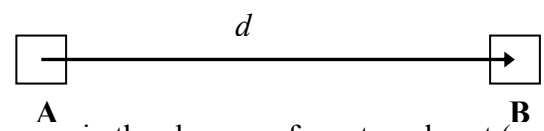
$$x \in X \text{ such that } x^* = f(x^*)$$

where X is the constraint space on variable x. An example is shown below.



1.4 Example of a Fixed Point Formulation of a Travel Model

The objective of this subsection is to illustrate the above ideas with a small example. Consider a travel forecasting problem with two nodes, A and B, connected by a single link from A to B. Travel from A to B in vehicles per hour is denoted by *d*, and the cost of travel is denoted by *c*. Assume the following demand function defines *d*:



$$d = d_o g(c)$$

where d_o is the total travel demand that would occur in the absence of any travel cost ($c = 0$), and $g(c)$ denotes a positive and decreasing function of travel cost. Examples of positive decreasing functions are:

$$g_1(c) = a - b \cdot c$$
$$g_2(c) = \exp(-\beta \cdot c)$$
$$g_3(c) = c^{-\alpha} = \exp(-\alpha \cdot \ln(c))$$

where a, b, β, α are parameters with positive values. Note that these functions are all positive and decreasing over values of $c \geq 0$.

Next, we need to define a functional relationship $h(\cdot)$ between link flow and link travel cost; that is, we need a performance function.

Since there is only one link, the flow *f* must equal *d* in this static model. Assume that

$$c = h(f) = h(d)$$

is a positive, increasing function of *f*. Examples of positive, increasing functions are:

$$h_1(f) = r + s \cdot f^n$$
$$h_2(f) = c_o\left(1 + s \cdot f^4\right)$$

where *r* and *s* are positive parameters, and *f* is raised to the power of *n*. The second function is widely known as the BPR function, which was originally proposed by engineers at the U.S. Bureau of Public Roads about 1960 with $s = 0.15 / capacity^4$, based on very little data.

Now, let us combine these functions to form a simultaneous model of travel demand and travel cost:

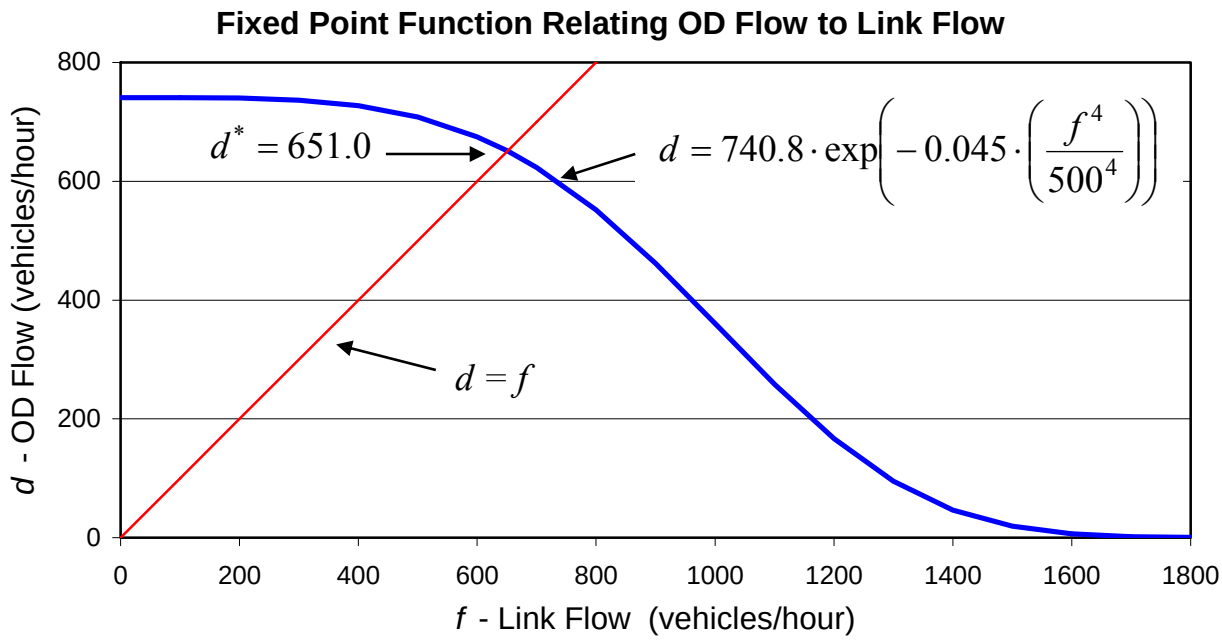
$$d = d_o g(h(f)) = d_o g(h(d))$$

Thus, we see that travel demand from A to B is a function that depends on that demand itself plus several parameters. For example,

$$d = d_o g(h(f)) = d_o \cdot \exp\left(-\beta\left(c_o\left(1 + s \cdot f^n\right)\right)\right)$$
$$= d_o \cdot \exp(-\beta \cdot c_o) \cdot \exp\left(-\beta \cdot c_o \cdot s \cdot d^n\right) = d'_o \cdot \exp\left(-\beta' \cdot d^n\right)$$
$$\text{if } d'_o \equiv d_o \cdot \exp(-\beta \cdot c_o) \text{ and } \beta' \equiv \beta \cdot c_o \cdot s$$

Such a formulation is called a fixed point problem. This class of problems is known to have a solution (existence) under rather general conditions. Under somewhat more restrictive conditions, there exists only one solution (uniqueness). Such properties are important for solving large-scale versions of the problem.

Next, consider a graphical representation of the example. Let $d_o = 1000$ vehicles/hour, $\beta = 0.01$, $s = 0.15/500^4$, $c_o = 30$ min. Plot $f(d) = d'_o \cdot \exp\left(-\beta' \cdot d^4\right)$ vs. *d*, and also plot $f(d) = d$. The intersection of these two equations is the fixed point d^*, which is the solution to the problem. It may also be regarded as an equilibrium point.



1.5 Sequential Travel Forecasting Procedure

Because of the complexity of solving the general travel forecasting problem for its equilibrium, the problem was simplified by converting it into a sequence of *steps* or *stages*. This simplification occurred in the very first transportation planning studies, and continues to this day.

This *decomposition* of such a problem is appropriate if the equilibrium solution is maintained. Unfortunately, this requirement was not even understood, much less observed in practice. Recently, efforts to solve the four-step procedure with *feedback* have recognized this requirement, but generally not fulfilled it completely and rigorously. The four-step procedure is described below as background for the consideration of the general problem in these notes. The equilibrium problem described above was conceptually solved as a sequence of four procedures. The notation is the same as defined earlier.

1. *Trip Generation* Determine the flow of travelers per day leaving or entering each zone as a function of zonal activity levels; also determine the auto availability for these trips.

$$O_p = f(R_p, \ldots) \qquad \text{for all } p$$
$$D_q = f(W_q, E_q, S_q, \ldots) \qquad \text{for all } q$$

2. *Trip Distribution* Determine the flow of travelers per day d_{pq} from origin zone p to destination zone q

$$d_{pq} = f\left(O_p, D_q, c_{pq}\right)$$

where c_{pq} is the generalized travel cost from origin zone *p* to destination zone *q*, determined in some convenient way, such as finding shortest routes over the roadway network for zero flow conditions. Assume d_{pq} is directly proportion to O_p and D_q, and inversely proportional to c_{pq}.

3. *Modal Split* Allocate flows of travelers to modes.

$$d_{pqm} = f\left(d_{pq}, c_{pqm}\right)$$

where *m* belongs to the set of modes serving zone pair *pq*.

4. *Traffic Assignment* Allocate flows of interzonal travelers by mode to routes

$$d_{pqmr} = f\left(d_{pqm}, c_{pqmr}\right)$$

where *r* belongs to the set of routes serving zone pair *pq* by mode *m*. Recall that

$$c_{pqmr} \equiv \sum_{a \in L_m} c_a(f_a) \cdot \delta^a_{pqmr}, \qquad \text{for all } p, q, m, r$$

where $c_a\left(f_a\right)$ = generalized cost of travel per vehicle on link *a* at flow f_a

$\delta^a_{pqmr} = 1$, if link *a* belongs to route *r* of mode *m* from zone *p* to zone *q;*

0, otherwise

Note the nature of the terminology used, as if travel behavior was a kind of mechanistic process. Our notation implies:

$$d_{pq} = \sum_m d_{pqm}$$

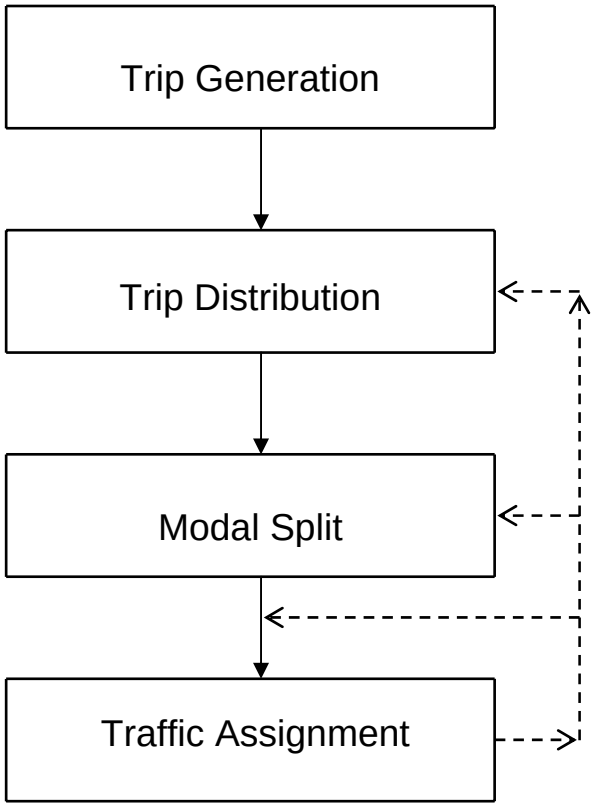
$$d_{pqm} = \sum_r d_{pqmr}$$

The relation of the four models (procedures) is shown in the following figure.

Often the questions of the source of the interzonal travel cost c_{pq} and its relation to c_{pqmr} and $c_a\left(f_a\right)$ is not addressed by this procedure. In other words, c_{pqmr} is assumed to be available in the trip distribution and modal split steps, but is determined only in the traffic assignment step. The dashed feedback arrow solves this problem conceptually, but leaves unanswered the question of whether the implied iterations converge to a stable solution. Until the *feedback* requirement imposed by the U. S. Intermodal Surface Transportation Efficiency Act of 1991 (ISTEA), such iterations were only rarely attempted in the United States.

Another issue is the time period of the travel activity. In the past, most forecasts were prepared for the 24-hour weekday. This was convenient since origins and destinations tend to balance

over 24 hours. The difficulty with the 24-hour concept is that travel costs vary over this period due to congestion during peak travel periods, which is the very congestion we are concerned with predicting. One approach to this problem is to compute a peaking factor, i.e. the ratio of peak-hour flow to 24-hour flow, typically, about 0.1. Such factors were used to convert hourly capacity into 24-hour capacity,



Sequential Travel Forecasting Procedure

or 24-hour link flows into hourly flows. The difficulty, of course, is that the factor depends on the level of congestion. An alternative approach is to factor 24-hour trips to hourly trip flows using observed data on trip departures and perform the analysis for selected time-of-day periods, say the peak morning and evening travel periods.

All such models are *static* flow models, because the travel times and flows are considered to be constant during the period of analysis. They are also *macroscopic*, since there is no concept of individual trips or vehicles moving through the network in these models. The peak period should be long enough, therefore, that it is reasonable to consider that trips can move from origin to destination. Usually, at least one hour is used. Other than the conceptual issue that travel times may exceed the period length, it doesn't matter. For the same reason, flows are said to propagate instantaneously from origin to destination. In other words, in the model the flows from origins to destinations either exist or they don’t exist.

References

J. de D. Ortúzar and L. G. Willumsen (2001) *Modelling Transport*, 3rd edition, Wiley; pp. 1-9 for discussion of demand and supply functions.

M. Manheim (1979) *Fundamentals of Transportation Systems Analysis*, Vol. 1, MIT Press, Cambridge, MA; especially Ch. 7.

Exercise

1. Consider the following generalized cost equation. Note that the distance term does not depend on link flow.

$$c_a(f_a) = t_a(f_a) + \gamma \cdot d_a$$

where $c_a(f_a) =$ generalized cost of travel on link a at flow f_a (minutes)

$t_a(f_a) =$ travel time on link a at flow f_a (minutes)

$d_a =$ travel distance on link a (miles)

$\gamma = \left(\frac{\text{cents}}{\text{mile}}\right) \cdot \left(\frac{\text{min}}{\text{cent}}\right)$, a conversion factor from cents/miles to minutes/cents to minutes/mile

The second term based on distance traveled is assumed to represent the monetary cost of travel, whereas the first term is assumed to represent the disutility of time. For route choice, assume that travelers only consider travel time, and not monetary cost in making their choices. However, for mode and OD choice, they consider both time and monetary costs in computing their generalized cost.

Define a more detailed version of the "General Problem" which implements these assumptions. Keep in mind this is a conceptual model, so the point is the concepts, not the mathematics. If you wish, give your answer in words, rather than equations.

Appendix - Optimization Methods

The objective of these notes is to introduce informally the methods needed to formulate and solve travel forecasting models as constrained optimization problems and more general approaches. The methods begin with unconstrained optimization from differential calculus, and proceed to the method of Lagrange multipliers (optimization with equality constraints). This classic approach is then extended to inequality and non-negativity constraints based on the Karush-Kuhn-Tucker Theorem (KKT). The conditions for existence and uniqueness of solutions are then stated.

The application of KKT method of network equilibrium problems often leads to assumptions that are not satisfied by the application of interest. In this event, either restricted assumptions are made, or a more general method based directly on the equilibrium conditions of interest is applied. One of these approaches is the nonlinear complementarity problem; another is the variational inequality problem, and the most general is the fixed point problem. These methods are introduced next, again in an informal manner.

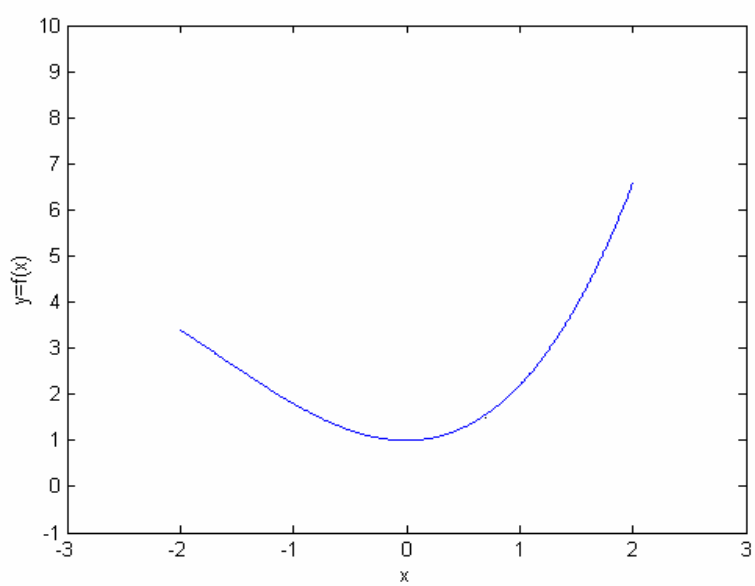
Further readings:

R.E. Miller, *Optimization*, Wiley, New York, 2000. (introductory and less formal)
M.S. Bazaraa, H.D. Sherali and C.M. Shetty, *Nonlinear Programming*, Wiley, New York, 1993. (more formal and advanced)
M. Patriksson, *The Traffic Assignment Problem, Models and Methods*, VSP, Utrecht, 1994, Section 3.1.

A.1 Preliminaries

1. Functions

Let $y = f(x)$ be a continuous function describing a relationship between scalars y and x.



If $\boldsymbol{x}$ is a vector of length n, then the function is a surface, possibly in n+1 dimensions.

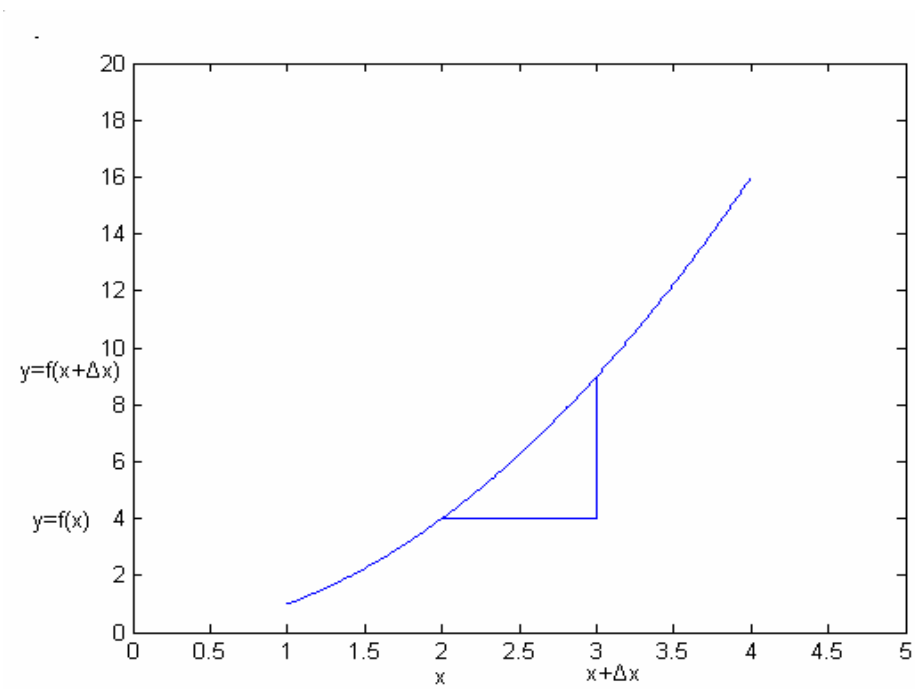
2. Derivatives

Consider

$$\frac{\Delta y}{\Delta x} = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

This equation describes Δy, the change in y, in relation to Δx, the change in x; then the derivative is defined as:

$$\left(\frac{dy}{dx}\right)_{x_0} = f'(x_0) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$$



3. Tangents and slopes

The slope of the tangent (a line just touching $f(x)$ and not intersecting it) to the function at x_0 is equal to the derivative at x_0.

4. Taylor series approximation

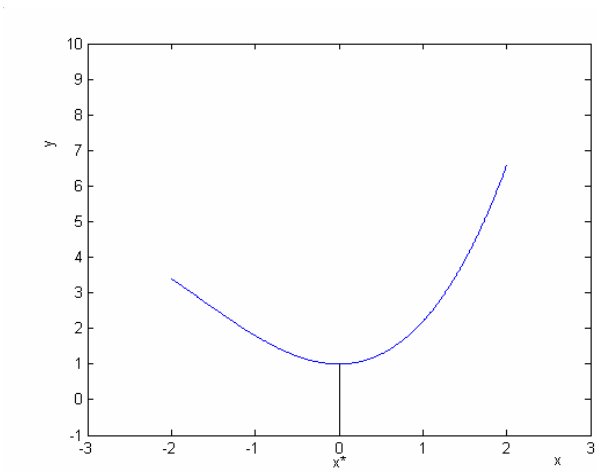
An approximation to $f(x)$ at x_0 can be found to any degree of accuracy, if the derivatives are known, as follows:

$$f(x_0 + dx) \approx f(x_0) + f'(x_0)dx + \left(\frac{1}{2!}\right) f''(x_0)(dx)^2 + \cdots$$

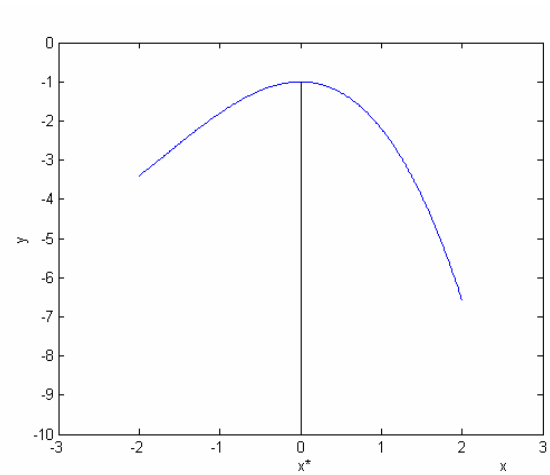
A.2 Minima and Convexity

1. Minima (maxima)

The minimum (maximum) of a function occurs at the point(s) x^* where the first derivative equals zero.



Convex function with minimum at 0



Concave function with maximum at 0

A zero-valued derivative is a *necessary* condition for a minimum or maximum point if $(-\infty < x < +\infty)$. If $(\underline{x} < x < \overline{x})$, then the minima or maxima may occur at the interval boundaries, where $\underline{x}$ denotes the minimum value of x, and $\overline{x}$ denotes the maximum value.

To determine if the point is a minimum or a maximum, we need to consider the second derivative,

$$\frac{d(dy/dx)}{dx} = \frac{df'(x)}{dx} = f''(x).$$

The second necessary condition for a minimum (maximum) is that $f''(x^*) \geq 0$ $(f''(x^*) \leq 0)$. In addition, there are similar sufficient conditions.

2. Convexity (concavity)

A function is *strictly convex* (*strictly concave*) if a straight line joining two points lies above (below) the function. If the line coincides with the function, it is *only convex* (*concave*). Hence, a linear function is both convex and concave. Likewise, consider a line tangent to $f(x)$ at $\hat{x}$; if $f(x)$ lies above (below) the tangent, the function is convex (concave).

3. Partial derivatives

For multi-dimensional variables ($\boldsymbol{x}$), the partial derivative $\frac{\partial f(\mathbf{x})}{\partial x_i}$ with respect to x_i is the derivative of $f(x)$ treating all variables other than x_i as constants, where T denotes transpose. The gradient of a function is a vector containing the first partial derivatives of the function:

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1} \cdots \frac{\partial f(\mathbf{x})}{\partial x_n} \right)^{\mathrm{T}}, \quad \text{where } \mathbf{x} = (x_1 \cdots x_n).$$

The Hessian of a function is the symmetric matrix of second partial derivatives

$$H = \begin{pmatrix} \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_1} \cdots \frac{\partial^2 f(\boldsymbol{x})}{\partial x_1 \partial x_n} \\ \cdot \qquad\qquad \cdot \\ \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_1} \cdots \frac{\partial^2 f(\boldsymbol{x})}{\partial x_n \partial x_n} \end{pmatrix}$$

4. Minimum of a multi-dimensional function

The necessary conditions for a convex function $f(\mathbf{x}^*)$ to achieve its minima at $\mathbf{x}^*$ are: $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

To determine if the point is a minimum (convex) or maximum (concave), certain conditions involving the Hessian must be met. In this course we will generally know whether the function is convex or not, so this condition is not stressed.

A.3 Optimization with Equality Constraints

1. Single equality constraint

We first consider the problem of finding the minimum of a convex function $f(\mathbf{x})$ subject to a single equality constraint:

$$\begin{aligned} &\min_{(\mathbf{x})} f(\mathbf{x}) \\ &\text{subject to}: h(\mathbf{x}) = 0 \\ &\text{where } \mathbf{x} = (x_1\, x_2 \cdots x_j \cdots x_n)^T \end{aligned}$$

To solve this problem, we form the Lagrangean function $L(\boldsymbol{x}, \lambda)$

$$L(\mathbf{x}, \lambda) \equiv f(\mathbf{x}) - \lambda \cdot h(\mathbf{x})$$

Note the similarity of $L(\mathbf{x},\lambda)$ to $f(\mathbf{x})$ for values of $\mathbf{x}$ that satisfy $h(\mathbf{x}) = 0$. Also, we make no assumption about the sign of the Lagrange multiplier λ. To find the necessary conditions for a minimum, we apply the same method as for an unconstrained minimization:

$$\frac{\partial L(\mathbf{x},\lambda)}{\partial x_j} = \frac{\partial f(\mathbf{x})}{\partial x_j} - \lambda \cdot \frac{\partial h(\mathbf{x})}{\partial x_j} = 0, \qquad j = 1,\ldots n$$

$$\frac{\partial L(\mathbf{x},\lambda)}{\partial \lambda} = -h(\mathbf{x}) = 0$$

Thus, we obtain $(n + 1)$ simultaneous equations in $(n + 1)$ unknowns:

$$\begin{aligned} &\nabla_{\mathbf{x}} f(\mathbf{x}) - \lambda \cdot \nabla_{\mathbf{x}} h(\mathbf{x}) = \mathbf{0} \\ &-h(\mathbf{x}) = 0 \end{aligned}$$

where $\nabla_{\mathbf{x}} f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_j} \right); \nabla_{\mathbf{x}} h(\mathbf{x}) = \left(\frac{\partial h(\mathbf{x})}{\partial x_j} \right)$.

Note we could also *add* $h(\mathbf{x})$ to $f(\mathbf{x})$; however, one standard convention is the form shown above.

2. Extension to several equality constraints

$$\begin{aligned} &\min_{(\mathbf{x})} f(\mathbf{x}) \\ &\text{s.t.}: h_i(\mathbf{x}) = 0, \qquad i = 1,\ldots m, \qquad m < n \end{aligned}$$

Forming the Lagrangean function, we obtain,

$$L(\mathbf{x},\lambda) = f(\mathbf{x}) - \sum_{i=1}^{m} \lambda_i \cdot h_i(\mathbf{x})$$

The necessary conditions are:

$$\begin{aligned} &\frac{\partial L(\mathbf{x},\lambda)}{\partial x_j} = \frac{\partial f(\mathbf{x})}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \cdot \frac{\partial h_i(\mathbf{x})}{\partial x_j} = 0, \qquad j = 1,\ldots n \\ &\frac{\partial L(\mathbf{x},\lambda)}{\partial \lambda_i} = -h_i(\mathbf{x}) = 0, \qquad i = 1,\ldots m \end{aligned}$$

Solve the $(n + m)$ equations for the unknowns $\mathbf{x}$ and $\boldsymbol{\lambda}$.

A.4 Optimization with Inequality Constraints

1. One inequality constraint

Next we consider the case of a single inequality constraint: $(h(\mathbf{x}) \geq 0)$

$$\min_{\mathbf{x}} f(\mathbf{x})$$
$$\text{s.t. } h(\mathbf{x}) - s^2 = 0$$

Here s^2 is a slack variable which represents the amount by which $h(\mathbf{x})$ exceeds 0. It allows us to convert an inequality constraint into an equality constraint. Forming the Lagrangian, we obtain:

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda \cdot \left(h(\boldsymbol{x}) - s^2\right)$$

(1) $$\frac{\partial L}{\partial x_j} = \frac{\partial f(\boldsymbol{x})}{\partial x_j} - \lambda \cdot \left(\frac{\partial h(\boldsymbol{x})}{\partial x_j}\right) = 0, \qquad j = 1 \cdots n$$

(2) $$\frac{\partial L}{\partial \lambda} = -h(\mathbf{x}) + s^2 = 0$$

(3) $$\frac{\partial L}{\partial s} = +2\lambda \cdot s = 0$$

The third derivative requires that:

1. $\lambda = 0$, or
2. $s = 0$, or
3. $\lambda = s = 0$.

If $s = 0$, then $h(\boldsymbol{x}) = 0$, and the constraint is an equality. So, (2) and (3) can be replaced by:

$$\lambda(h(\boldsymbol{x})) = 0$$

This expression is called the complementary slackness condition, which can be interpreted as follows:

1. if $\lambda = 0$, then $h(\boldsymbol{x}) \neq 0$;

2. if $h(\boldsymbol{x}) = 0$, then $\lambda \neq 0$;

3. or $\lambda = h(\boldsymbol{x}) = 0$.

However, in the process, the direction of the inequality, $h(\boldsymbol{x}) \geq 0$, has been lost. So we need to re-introduce it as an explicit requirement, resulting in the following optimality conditions, also now known as the Karush-Kuhn-Tucker (KKT) conditions;

$$
\begin{aligned}
&\nabla f(\boldsymbol{x}) - \lambda \cdot \nabla h(\boldsymbol{x}) = 0 \\
(4) \qquad &h(\boldsymbol{x}) \geq 0 \\
&\lambda \cdot h(\boldsymbol{x}) = 0
\end{aligned}
$$

Moreover, we assume that $\lambda \geq 0$. That is, we associate a nonnegative multiplier with the nonnegative constraint.

The necessary conditions for x^* to be a relative minimum of $f(\boldsymbol{x})$ s.t. $h(\boldsymbol{x}) \geq 0$ are that a nonnegative λ exists such that it and x^* satisfy conditions (4).

2. Extension to several inequality constraints

The extension to several inequality constraints is straightforward:

$$
\begin{aligned}
&\min_{x} f(\boldsymbol{x}) \\
&\text{s.t. } h_i(\boldsymbol{x}) \geq 0, \qquad i = 1 \ldots m
\end{aligned}
$$

$$
L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \sum_{i=1}^{m} \lambda_i \cdot h_i(\mathbf{x})
$$

The optimality necessary conditions for a relative minimum are:

$$
\begin{aligned}
&\frac{\partial f(\boldsymbol{x})}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \cdot \left(\frac{\partial h_i(x)}{\partial x_j} \right) = 0, \qquad j = 1 \ldots n \\
&h_i(\boldsymbol{x}) \geq 0, \qquad i = 1 \ldots m \\
&\lambda_i . h_i(\boldsymbol{x}) = 0, \qquad i = 1 \ldots m \\
&\lambda_i \geq 0, \qquad i = 1 \ldots m
\end{aligned}
$$

If $f(\mathbf{x})$ is a strictly convex function, the optimality conditions are also sufficient to define a unique minimum, if the region defined by the constraints is concave.

3. Nonnegative variables

Nonnegativity constraints can be represented like any other inequality constraints $(h_i(\boldsymbol{x}) \geq 0)$. However, it is common and convenient to treat them separately. To see how this is done, let's rewrite the above problem, making the nonnegativity constraints explicit.

$$\min_{\mathbf{x}} f(\mathbf{x})$$
$$\text{s.t.}: h_i(\mathbf{x}) \geq 0, \qquad i = 1 \ldots m$$
$$x_j \geq 0, \qquad j = 1 \ldots n$$
$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \sum_i \lambda_i \cdot (h_i(\mathbf{x})) - \sum_j \theta_j \cdot x_j$$

The KKT necessary conditions are:

(1) $\dfrac{\partial f(\boldsymbol{x})}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \cdot \left(\dfrac{\partial h_i(x)}{\partial x_j} \right) - \theta_j = 0, \qquad j = 1 \ldots n$

(2) $h_i(\mathbf{x}) \geq 0, \qquad i = 1 \ldots m$

(3) $\lambda_i \cdot h_i(\mathbf{x}) = 0, \qquad i = 1 \ldots m$

(4) $x_j \geq 0, \qquad j = 1 \ldots n$

(5) $\theta_j \cdot x_j = 0, \qquad j = 1 \ldots n$

(6) $\lambda_i \geq 0, \qquad i = 1 \ldots m$

(7) $\theta_j \geq 0, \qquad j = 1 \ldots n$

Now rewrite conditions (1) as:

$$\frac{\partial f(\boldsymbol{x})}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \cdot \left(\frac{\partial h_i(x)}{\partial x_j} \right) = \theta_j, \qquad j = 1 \ldots n$$

Substituting θ_j into conditions (5), we obtain:

(5′) $x_j \cdot \left(\dfrac{\partial f(\boldsymbol{x})}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \cdot \left(\dfrac{\partial h_i(x)}{\partial x_j} \right) \right) = 0, \qquad j = 1 \ldots n$

Moreover, since $\theta_j \geq 0$,

$$\frac{\partial f(\boldsymbol{x})}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \cdot \left(\frac{\partial h_i(x)}{\partial x_j} \right) \geq 0$$

Therefore, we can rewrite conditions (1)-(6) as:

(1′) $\dfrac{\partial f(\boldsymbol{x})}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \cdot \left(\dfrac{\partial h_i(x)}{\partial x_j} \right) \geq 0, \qquad j = 1 \ldots n$

(5′) $x_j\left(\frac{\partial f(\mathbf{x})}{\partial x_j}-\sum_{i=1}^{m}\lambda_i\cdot\left(\frac{\partial h_i(x)}{\partial x_j}\right)\right)=0, \quad j=1\ldots n$

(2) $h_i(\mathbf{x})\geq 0, \quad i=1\ldots m$

(3) $\lambda_i\cdot h_i(\mathbf{x})=0, \quad i=1\ldots m$

(4) $x_j\geq 0, \quad j=1\ldots n$

(6) $\lambda_i\geq 0, \quad i=1\ldots m$

Note that θ_j vanishes.

In words, the necessary conditions for $\mathbf{x}^*$ to be a relative minimum of $f(\mathbf{x})$ subject to $h_i(\mathbf{x})\geq 0$ and $\mathbf{x}\geq 0$ are that $\mathbf{x}^*$ satisfies conditions (1′, 2, 3, 4, 5′, 6). If $f(\mathbf{x})$ is convex and the constraints $h_i(\mathbf{x})$ are concave or quasi-concave, then these are also sufficient conditions for a minimum. If $f(\mathbf{x})$ is strictly convex, the minimum is unique.

A.5 More General Formulations

In the applications of constrained optimization methods to network equilibrium problems, we will be faced with the situation where the function to be minimized has the form:

$$f(\mathbf{x})=\int_0^{\mathbf{x}}\mathbf{c}(\mathbf{s})d\mathbf{s}$$

To solve such problems as an optimization problem, the travel cost function **c** must be integrable; otherwise, the function is not well-defined. If **c** is differentiable, the integral is well-defined if and only if the Jacobian matrix $\nabla_{\mathbf{x}}\mathbf{c}(\mathbf{x})$ is symmetric everywhere. That is,

$$\frac{\partial c_i(\mathbf{x})}{\partial x_j}=\frac{\partial c_j(\mathbf{x})}{\partial x_i}, \quad \forall i,j,\forall\mathbf{x}\in X$$, where X is the constraint space of **x**.

One common way to satisfy this requirement is to assume that $c_i(\mathbf{x})=c_i(x_i)$. That is, $c_i(\mathbf{x})$, the cost of element *i*, depends not on the vector **x**, but only on the element x_i. Such a function is termed *separable*.

Example: a link's travel time depends only on its own flow, and not on the flows on opposing or conflicting links. The symmetry requirement means that the effect of the conflicting flow from a minor arterial on a major arterial's travel time must be equal to the effect of the major arterial's flow on the minor arterial's travel time. This condition is unlikely to be met, and therefore the symmetry requirement fails. Another case pertains to the effect of car and truck flows on each other's travel times on the same link.

In order to relax this restriction, we consider generalized forms of the constrained optimization problem, known as the Nonlinear Complementarity Problem and the Variational Inequality Problem. These formulations, however, do not apply to all possible situations, as discussed below.

<u>Nonlinear Complementarity Problem</u>

To illustrate the formulation of the Nonlinear Complementarity Problem, or NCP, we reconsider the optimality conditions for the problem stated in A.4.3,

$$\begin{aligned} &\min_{(\mathbf{x})} f(\mathbf{x}) \\ &\text{s.t.}: h_i(\mathbf{x}) \geq 0, \qquad i = 1 \ldots m \\ &\quad x_j \geq 0, \qquad j = 1 \ldots n \end{aligned}$$

The optimality conditions are:

$$(1') \qquad \frac{\partial f(\mathbf{x})}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \cdot \left(\frac{\partial h_i(x)}{\partial x_j} \right) \geq 0, \qquad j = 1 \ldots n$$

$$(5') \qquad x_j \left(\frac{\partial f(\mathbf{x})}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \cdot \left(\frac{\partial h_i(x)}{\partial x_j} \right) \right) = 0, \qquad j = 1 \ldots n$$

$$(2) \qquad h_i(\mathbf{x}) \geq 0, \qquad i = 1 \ldots m$$

$$(3) \qquad \lambda_i \cdot h_i(\mathbf{x}) = 0, \qquad i = 1 \ldots m$$

$$(4) \qquad x_j \geq 0, \qquad j = 1 \ldots n$$

$$(6) \qquad \lambda_i \geq 0, \qquad i = 1 \ldots m$$

Now, define,

$$\mathbf{F}(\mathbf{y}) = \begin{bmatrix} \left(\dfrac{\partial f(\mathbf{x})}{\partial x_j} - \displaystyle\sum_{i=1}^{m} \lambda_i \cdot \left(\dfrac{\partial h_i(x)}{\partial x_j} \right) \right) & j = 1 \ldots n \\ h_i(\mathbf{x}) & i = 1 \ldots m \end{bmatrix} \geq 0$$

$$\mathbf{y} = \begin{bmatrix} x_j & j = 1 \ldots n \\ \lambda_i & i = 1 \ldots m \end{bmatrix} \geq 0$$

Finding $\mathbf{y}^*$ such that $\mathbf{F}(\mathbf{y}^*)^{\mathrm{T}} \cdot \mathbf{y}^* = 0$ solves the nonlinear complementarity problem (NCP). Note that each element of this vector product must be zero for the result to be zero, since all the elements are non-negative.

Examining the above definitions, we see that $\mathbf{y} \geq 0$ is equivalent to (4) and (6). $F(\mathbf{y}) \geq 0$ is equivalent to (1′) and (2′). And, the requirement that $\mathbf{F}(\mathbf{y}^*)^{\mathrm{T}} \cdot \mathbf{y}^* = 0$ is equivalent to (5′) and (3). Therefore, the NCP is equivalent to the KKT conditions for the minimization of a convex function, subject to inequality and nonnegativity constraints. Note that the feasible region for this problem is simply the nonnegative orthant of dimension (n+m). The advantage of the NCP over the optimization formulation is clear for the case that only the derivatives $\frac{\partial f(\mathbf{x})}{\partial x_j}$ are available, but the function $f(\mathbf{x})$ is not well-defined, as illustrated by the above example.

<u>Variational Inequality Problem</u>

Equivalent to every NCP is a Variational Inequality Problem (VIP) of the form,

$$\mathbf{F}(\mathbf{y}^*)^{\mathrm{T}} \cdot (\mathbf{y} - \mathbf{y}^*) \geq 0 \quad \forall \mathbf{y} \in \mathbf{Y}, \text{ where } \mathbf{Y} \text{ is the feasible region.}$$

To show that this VIP is implied by the NCP, rewrite this formulation as:

$$\mathbf{F}(\mathbf{y}^*)^{\mathrm{T}} \cdot (\mathbf{y}) \geq F(\mathbf{y}^*)^{\mathrm{T}} \cdot (\mathbf{y}^*) = 0 \quad \forall \mathbf{y} \in \mathbf{Y}.$$

The right-hand side is simply the NCP, so it must be equal to 0. Since any $\mathbf{y} \in \mathbf{Y}$ is nonnegative, and $\mathbf{F}(\mathbf{y}^*)$ is also nonnegative, their product must be positive or zero for any $\mathbf{y} \in \mathbf{Y}$. Note that the left-hand side has the form of the objective function of a linear programming problem with a feasible region **Y**, since $\mathbf{F}(\mathbf{y}^*)$ is a constant vector. Minimizing this function with respect to $\mathbf{y}$ yields $\mathbf{y}^*$.

Unlike the NCP's restriction to a nonnegative feasible region, the VIP may have a defined constraint space. For example, let the constraint space $\mathbf{Y}$ be redefined by the inequalities,

(2) $\quad h_i(\mathbf{x}) \geq 0, \qquad i = 1 \ldots m$

(4) $\quad x_j \geq 0, \qquad j = 1 \ldots n$

(6) $\quad \lambda_i \geq 0, \qquad i = 1 \ldots m$

Then, we can redefine the function as,

$$F_j(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \cdot \left(\frac{\partial h_i(x)}{\partial x_j} \right) \geq 0, \qquad j = 1 \ldots n$$

If the constraints are linear in **x**, then the second term will be a constant for both **x** and $\mathbf{x}^*$, and will cancel out. In this case, $\mathbf{F}(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) \geq 0, \forall \mathbf{x} \in \mathbf{Y}$

For further details, see A. Nagurney (1999) *Network Economics*, Kluwer, Norwell, MA, Ch. 1.

Monotonicity and the Fixed Point Problem

The requirements for existence and uniqueness for NCP and VIP parallel those of optimization problems. The convexity requirement is replaced by a monotonicity requirement on $\mathbf{F}(\mathbf{x})$.

$$\mathbf{F}(\mathbf{x}) \text{ is monotone on } X \text{ if } (\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y}))^{\mathrm{T}} \cdot (\mathbf{x} - \mathbf{y}) \geq 0,\ \forall\, \mathbf{x}, \mathbf{y} \in X.$$

If x and y are scalars, then the definition means that the scalar function F is increasing (decreasing if < 0) from x to y. In two or more dimensions, the definition means that the surface defined by the product of the vector-valued function $\mathbf{F}(\mathbf{x})$ and $(\mathbf{x} - \mathbf{y})$ is increasing from $\mathbf{x}$ to $\mathbf{y}$. An example is given in Section 2 of these Notes. As usual, *strict* monotonicity means that > replaces $\geq$ in the definition.

Finally, we give the definition of the Fixed Point Problem: find an $\mathbf{x}^*$ in X such that $F(\mathbf{x}^*) = \mathbf{x}^*$, where F is a mapping from the set X to the n-dimensional space. Historically, the fixed point problem has been used to prove existence of solutions to VIP or NCP models.

A.6 Examples of Constrained Optimization Problems

$$\begin{aligned}
&\min_{\mathbf{x}} f(\mathbf{x}) = x_1^2 + x_2^2 + 4x_1 - 6x_2 \\
&\text{s.t.}: h_1 = 10 - x_1 - x_2 \geq 0 \\
&\qquad h_2 = 18 - 3x_1 - x_2 \geq 0 \\
&\qquad \mathbf{x} \geq 0
\end{aligned}$$

$$L(\mathbf{x}, \lambda) = x_1^2 + x_2^2 + 4x_1 - 6x_2 - \lambda_1(10 - x_1 - x_2) - \lambda_2(18 - 3x_1 - x_2)$$

The optimality conditions are:

$$\begin{aligned}
&\frac{\partial L}{\partial x_1} = 2x_1 + 4 - \lambda_1(-1) - \lambda_2(-3) \geq 0 \\
&\frac{\partial L}{\partial x_2} = 2x_2 - 6 - \lambda_1(-1) - \lambda_2(-1) \geq 0 \\
&x_1(2x_1 + 4 - \lambda_1(-1) - \lambda_2(-3)) = 0 \\
&x_2(2x_2 - 6 - \lambda_1(-1) - \lambda_2(-1)) = 0 \\
&10 - x_1 - x_2 \geq 0 \\
&18 - 3x_1 - x_2 \geq 0 \\
&\lambda_1(10 - x_1 - x_2) = 0 \\
&\lambda_2(18 - 3x_1 - x_2) = 0 \\
&\lambda \geq 0
\end{aligned}$$

Consider the solution **x** = (0 3).

$$\frac{\partial L}{\partial x_1} = 0 + 4 - \lambda_1(-1) - \lambda_2(-3) \geq 0$$
$$\frac{\partial L}{\partial x_2} = 6 - 6 - \lambda_1(-1) - \lambda_2(-1) \geq 0$$
$$0(0 + 4 - \lambda_1(-1) - \lambda_2(-3)) = 0$$
$$3(0 + 6 - 6 - \lambda_1(-1) - \lambda_2(-1)) = 0 \qquad \Rightarrow \lambda_1 = -\lambda_2$$
$$10 - 0 - 3 \geq 0 \qquad \text{OK}$$
$$18 - 0 - 3 \geq 0 \qquad \text{OK}$$
$$\lambda_1(10 - 0 - 3) = 0 \qquad \Rightarrow \lambda_1 = 0$$
$$\lambda_2(18 - 0 - 3) = 0 \qquad \Rightarrow \lambda_2 = 0$$
$$\lambda \geq 0$$

All conditions are met for $\lambda_1 = \lambda_2 = 0$, and $f(0\ 3) = 0 + 9 + 0 - 18 = -9$.
This solution is one of 16 combinations of solutions to **x** and λ. It is the only one that satisfies all of the constraints and therefore is the relative minimum.

As another example, consider **x** = (6 0).

$$\frac{\partial L}{\partial x_1} = 12 + 4 - \lambda_1(-1) - \lambda_2(-3) \geq 0$$
$$\frac{\partial L}{\partial x_2} = 0 - 6 - \lambda_1(-1) - \lambda_2(-1) \geq 0$$
$$6(12 + 4 - \lambda_1(-1) - \lambda_2(-3)) = 0$$
$$0(0 - 6 - \lambda_1(-1) - \lambda_2(-1)) = 0$$
$$10 - 6 - 0 \geq 0 \qquad \text{OK}$$
$$18 - 18 - 0 \geq 0 \qquad \text{OK}$$
$$\lambda_1(10 - 4 - 0) = 0 \qquad \Rightarrow \lambda_1 = 0$$
$$\lambda_2(18 - 18 - 0) = 0 \qquad \Rightarrow \lambda_2 \geq 0$$
$$\lambda \geq 0$$
$$-6 + 0 + \lambda_2 \geq 0 \qquad \Rightarrow \lambda_2 \geq 6$$
$$6(16 + 3\lambda_2) = 0 \qquad \Rightarrow \lambda_2 = -\frac{16}{3}$$

which contradicts the condition that $\lambda_2 \geq 0$.

Exercises

1. Solve the above problem for two or more additional combinations:

    a. a combination on the boundary of the feasible region defined by positive values of Lagrange multipliers;

    b. a combination of values in the interior of the region.

    To understand the problem better, draw the constraints and plot the contours of the objective function around the point that minimizes $f(\mathbf{x})$, if unconstrained.

2. Draw the picture corresponding to the following problem, and find $\mathbf{x}^*$ by assuming the constraint is an equality (active). Then, solve the optimality conditions that correspond to this result, and show that they are satisfied.

$$\min_{\mathbf{x}}\left(6x_1^2 + 5x_2^2\right)$$
$$\text{s.t.}: x_1 + 5x_2 \geq 3$$
$$\mathbf{x} \geq 0$$

3. Add the following constraint to Problem 2 and re-solve;

$$x_1 + 5x_2 \leq 3$$

4. The daily emission of airborne pollutants from two petrochemical facilities is given (in thousands of cubic feet) by

$$E = 2x_1^2 - 80x_1 + 8x_2^2 + 1000$$

where x_1 and x_2 represent hours of operation per day of the two facilities.

    a. What amount of daily operation (in hours) of each of the two facilities would be optimal in the sense of minimizing daily pollution emission?

    b. How would you answer part (a) above if, in addition, there were an upper limit of 12 hours per day of operations at facility 1?

    c. You later discover (as is usually the case) that the E function was incorrectly estimated: it actually should be:

$$E = 2x_1^2 - 80x_1 + 8x_2^2 - 4x_1x_2 + 1100$$

    How would you now answer the question in part a?

    Use the full set of optimality conditions in answering at least one part of this question.

2. Auto Route Choice Principles

We begin our consideration of travel choice models for urban travel forecasting with the problem of how to predict travelers' choice of route through a roadway network, given the origin and destination of travel. The solution of this route choice problem for road and transit networks, with or without congestion, also establishes a method for determining modal travel costs. With this information, it is then possible to consider models of mode and origin-destination choice, as well as models of choice of residential location.

In transportation planning practice, the problem of route choice is traditionally called *traffic* (or *trip*) *assignment*, since origin-destination (O-D) flows were viewed as being mechanically "assigned" to the network. The *user-optimal* model (UO) of route choice described in these notes is now widely accepted and applied, both by practitioners and researchers, but is still not widely understood, even though a convergent solution algorithm was devised by 1968, based on a formulation published in 1956, and implemented in some computer codes by the mid-1970s.

2.1 User-Optimal Route Choice

2.1.1 Assumptions

A British traffic scientist, John Wardrop, proposed the following "*criterion*" for a network equilibrium in a paper published in 1952:

> "The journey times on all routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route."

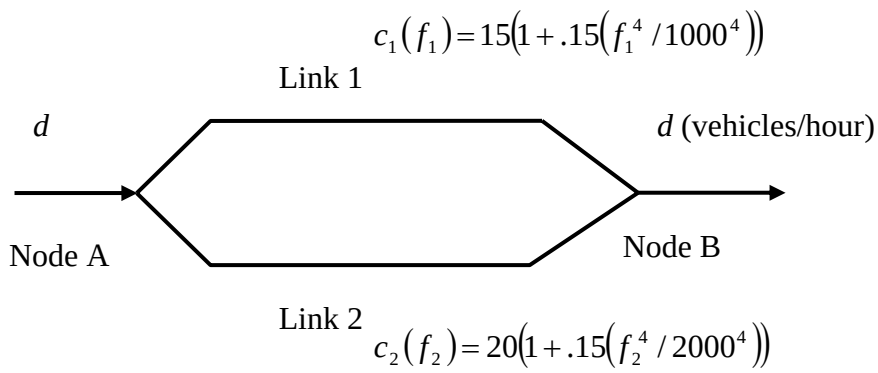
This statement is now known as Wardrop's first principle for *user-optimal* (UO) network equilibrium. An alternative statement is that each traveler chooses the shortest route as measured by travel time. The two statements are equivalent because no traveler can find a shorter route if Wardrop's criterion is met. Beckmann et al (1956) independently stated the same criterion.

By *used routes*, Wardrop meant the routes used from a given origin to a given destination, which in practice are defined in terms of zones. Today, we interpret *journey times* as generalized costs of travel. In general, routes consist of sequences of links; then, route costs may be defined as the sum of the link costs along the route. Since link costs generally depend on link flows, through the cost-flow function, and each link serves (possibly) many routes, then identifying the route costs and flows which satisfy the above Wardrop condition involves solving the route choice problem for the entire system of zone pairs and the entire network simultaneously.

Since link flows have units of vehicles/hour, then the corresponding route flows and O-D flows also have units of vehicles/hour or persons/hour. The resulting route choice model is a *steady-state flow* or *static* model in which there is no concept of a traveler moving from an origin to a destination. Rather, there is an O-D flow, with each link of each used route having a corresponding flow. This formulation results in a highly specific concept of "congestion," in which there are no bottlenecks or traffic jams, only steadily flowing vehicles at speeds determined by the cost-flow performance functions.

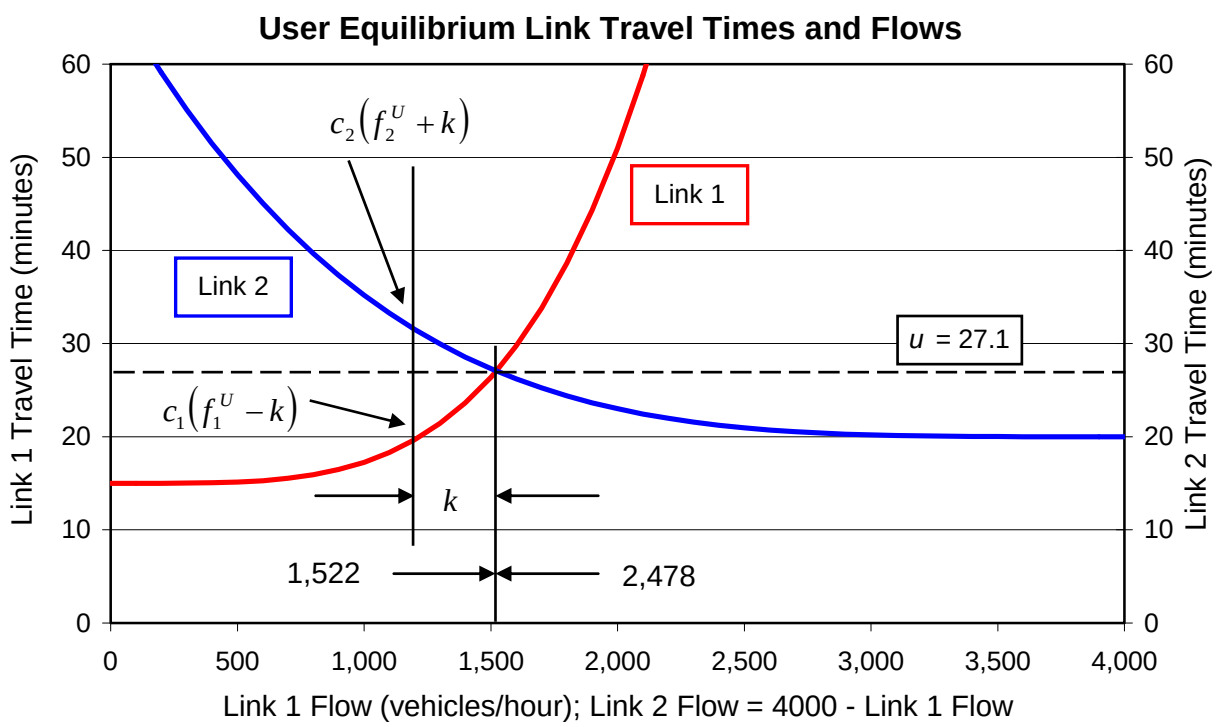
2.1.2 User-Optimal Formulation

In this section, we consider a simple case with a single O-D pair. In Section 2 we shall generalize this case to an entire network. Suppose two nodes A and B are connected by two links, 1 and 2. The fixed O-D flow/hour is *d*.



Two-Link, Two-Node Example

The travel cost on link 1 is given by $c_1 = c_1(f_1)$, an increasing function of total link flow, where f_1 is the flow on link 1; likewise, $c_2 = c_2(f_2)$. Conservation of flow requires that $f_1 + f_2 = d$. If we want the solution such that the link costs are equal at equilibrium, then the user-optimal flows (f_1^U, f_2^U) can be found graphically as shown above. The user-optimal flows correspond to the intersection of $c_1(f_1)$ and $c_2(f_2)$, that is, where the travel costs on the two links are equal.



The equilibrium cost is denoted as *u*; the equilibrium flows are f_1^U and $f_2^U = \left(d - f_1^U\right)$. Consider some other allocation of flows, $\left(f_1^U - k\right)$ and $\left(f_2^U + k\right)$. The travel costs at these flows are $c_1\left(f_1^U - k\right) < c_2\left(f_2^U + k\right)$. Thus, this solution does not satisfy Wardrop's criterion. By an unspecified process, *k* vehicles/hour need to switch routes to achieve the user-optimal solution. The above diagram illustrates another attribute of the solution: the area under the link cost-flow functions is minimized at the UO solution. Any other solution, such as $\left(f_1^U - k\right)$, results in a larger area by the amount of the triangle to the right of the vertical line at $f_1^U - k$. This insight suggests an analytical problem formulation:

User-Optimal Traffic Assignment Problem (UO TAP)

$$\min_{(f_1, f_2)} z(\mathbf{f}) = \int_0^{f_1} c_1(x)dx + \int_0^{f_2} c_2(x)dx \qquad \text{(objective function)}$$

$$\text{st}: f_1 + f_2 = d \qquad \text{(conservation of flow constraint)}$$

$$f_1 \geq 0,\ f_2 \geq 0 \qquad \text{(non-negativity constraints)}$$

A specific form of the travel cost-flow performance function is the BPR function for a link with a zero flow travel time of c_a^0 and a nominal capacity of C_a:

$$c_a = c_a^0\left(1 + 0.15\left(\frac{f_a}{C_a}\right)^4\right)$$

The integral of the BPR function is:

$$\int_0^{f_a} c_a(x)dx = c_a^0\left(f_a + 0.03\left(\frac{f_a^5}{C_a^4}\right)\right)$$

By comparison, the total travel cost on link *a* is:

$$T_a(f_a) \equiv c_a^0\left(f_a + 0.15\left(\frac{f_a^5}{C_a^4}\right)\right)$$

2.1.3 Analysis of User-Optimal Conditions

Let us proceed to demonstrate analytically that the solution of **UO TAP** corresponds to the Wardrop condition.

$$L(\mathbf{f},u)=\int_0^{f_1} c_1(x)dx+\int_0^{f_2} c_2(x)dx-u(f_1+f_2-d)$$

$$\frac{\partial L}{\partial f_1}=c_1(f_1)-u(+1)\geq 0$$

$$\frac{\partial L}{\partial f_2}=c_2(f_2)-u(+1)\geq 0$$

$$\frac{\partial L}{\partial u}=f_1+f_2-d\geq 0 \quad \text{(if written as an inequality constraint)}$$

$$f_1(c_1(f_1)-u)=0$$

$$f_2(c_2(f_2)-u)=0$$

$$u(f_1+f_2-d)=0$$

$$\mathbf{f}\geq \mathbf{0},\ u\geq 0$$

In order to take the first derivatives, we need to know that the derivative of an integral with respect to its upper limit is the integrand evaluated at that limit times the partial derivative of the upper limit. More generally,

$$\frac{\partial}{\partial x}\left(\int_0^z t(y)dy\right)=t(z)\cdot\frac{\partial z}{\partial x}$$

First, we consider the case illustrated by the figure, $\mathbf{f}>\mathbf{0}$. In this event, we have:

$$c_1\left(f_1^U\right)=u, \text{ and}$$

$$c_2\left(f_2^U\right)=u, \text{ so}$$

$$c_1\left(f_1^U\right)=c_2\left(f_2^U\right)$$

Hence, at optimality the travel costs are equal; moreover, since $u>0$, $f_1+f_2=d$. Therefore, Wardrop's first principle is satisfied. Suppose one of the links is sufficiently low in cost that no flow occurs on the other link. To analyze this situation, we proceed as follows. Assume link 1 is the lower cost link that receives all of the flow.

1. If $f_1^U>0, f_2^U=0$, then $c_1\left(f_1^U\right)=u, c_2\left(f_2^U\right)\geq u$ and $f_1^U=d$.

   Note that Wardrop's first principle is satisfied since the cost of link 2 cannot be less than the cost of link 1;

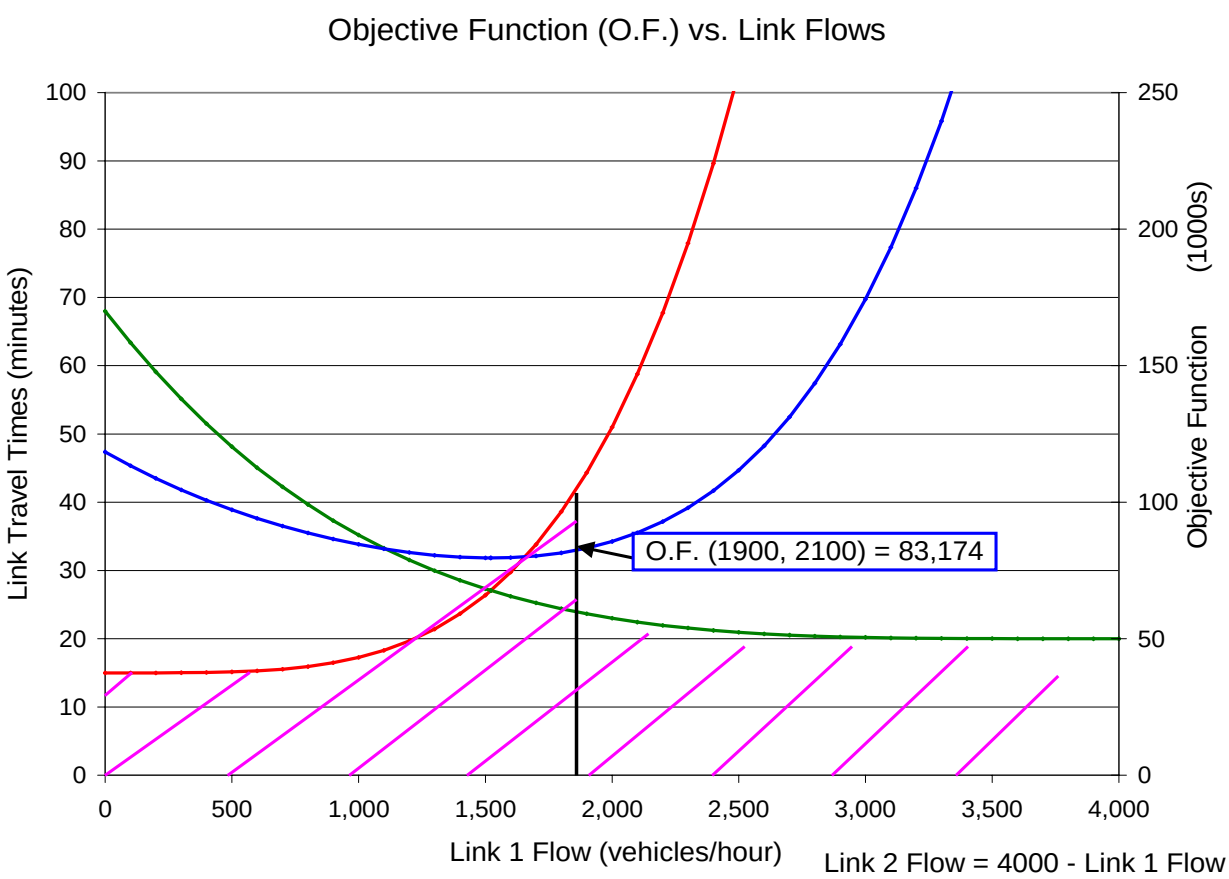
2. If $c_2\left(f_2^U\right)>c_1\left(f_1^U\right)=u$, then $f_2^U=0$.

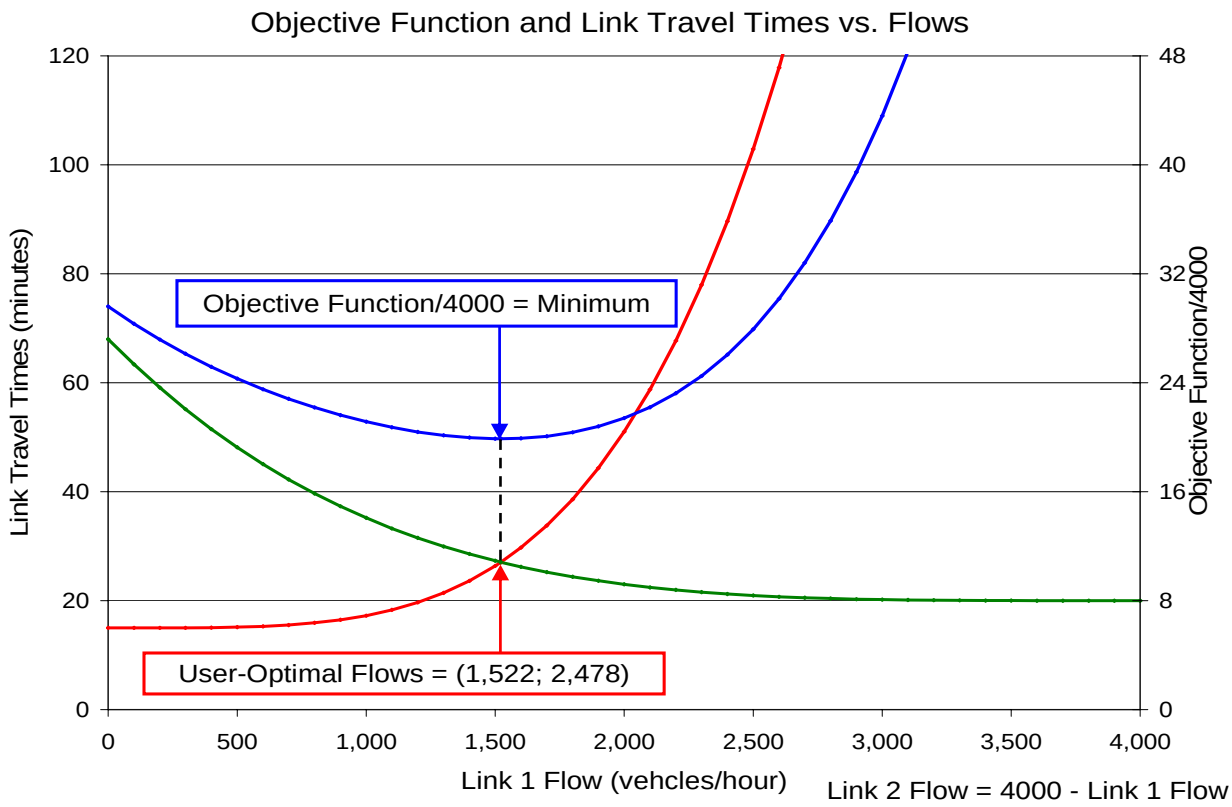
   In this case, the higher cost of link 2 requires that its flow is 0, according to the above optimality conditions.

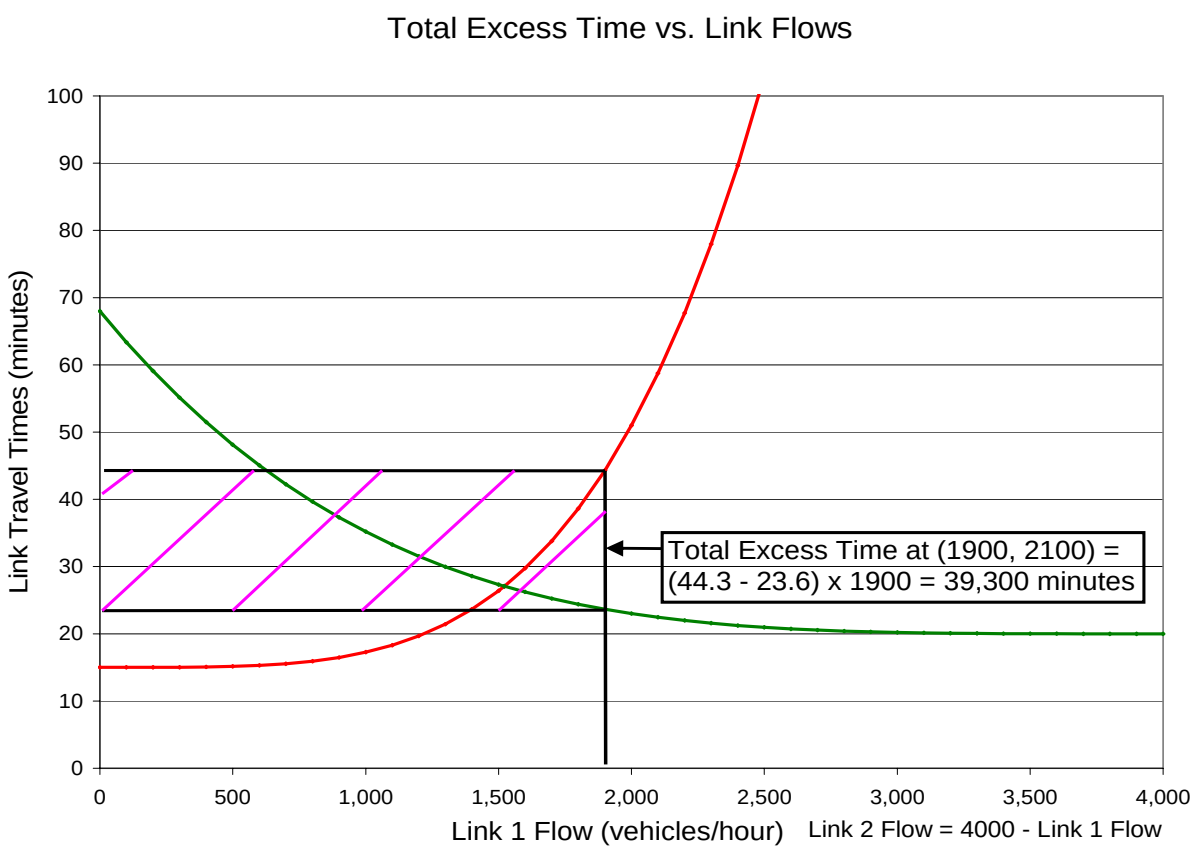
2.1.4 Total Excess Cost – An Intuitive Way to Find the User Equilibrium

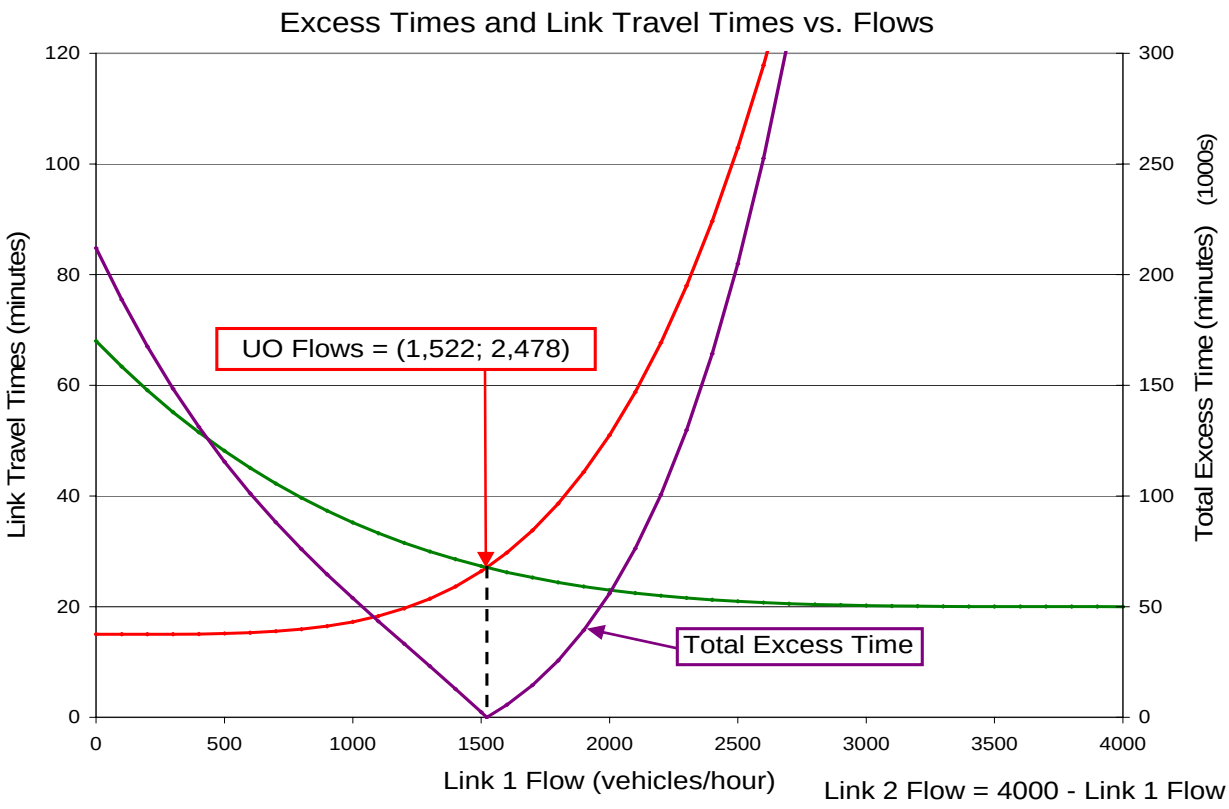
The following figures suggest a more intuitive way to evaluate solutions to network equilibrium problems for large networks, as illustrated with the two-link case. The intersection of the two link cost functions determines the user equilibrium, which is the minimum of the sum-of-the-integrals objective function, as shown below and the top of p. 31. Because the objective function is very flat in the vicinity of its minimum, it does not provide a useful measure for determining the minimum point. For that purpose, another measure is needed; a common choice is the *Gap*, which is equivalent to the *Total Excess Time* or *Total Excess Cost*.

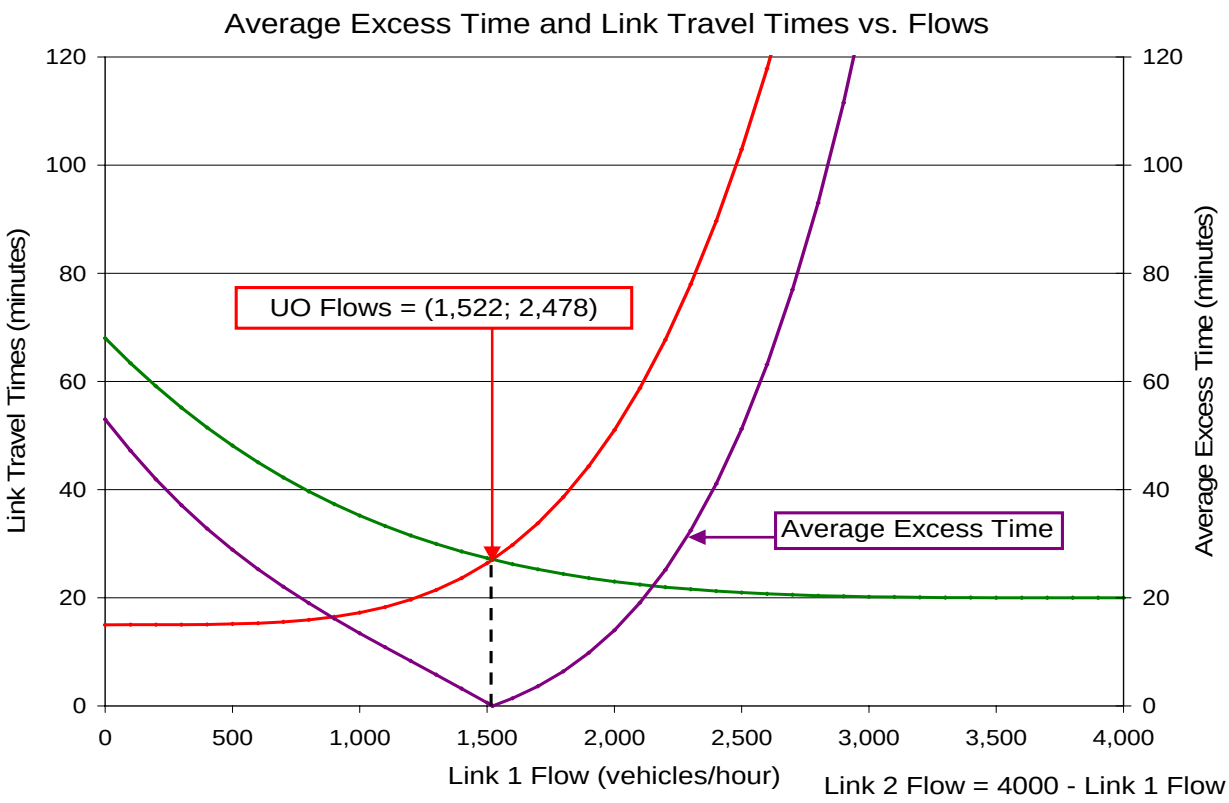
The definition of the Total Excess Time is illustrated by the figures at the bottom of p. 31 and the top of p. 32. The second figure shows the Total Excess Time for each feasible solution of the problem, illustrating that this measure is zero at the user-equilibrium solution. For convenience of plots and their description, the Total Excess Time can also be expressed as the Average Excess Time, as shown at the bottom of p. 32. At the top of p. 33, the Average Excess Time is superimposed on the standardized objective function, showing the advantage of the excess cost measure for determining the minimum point.

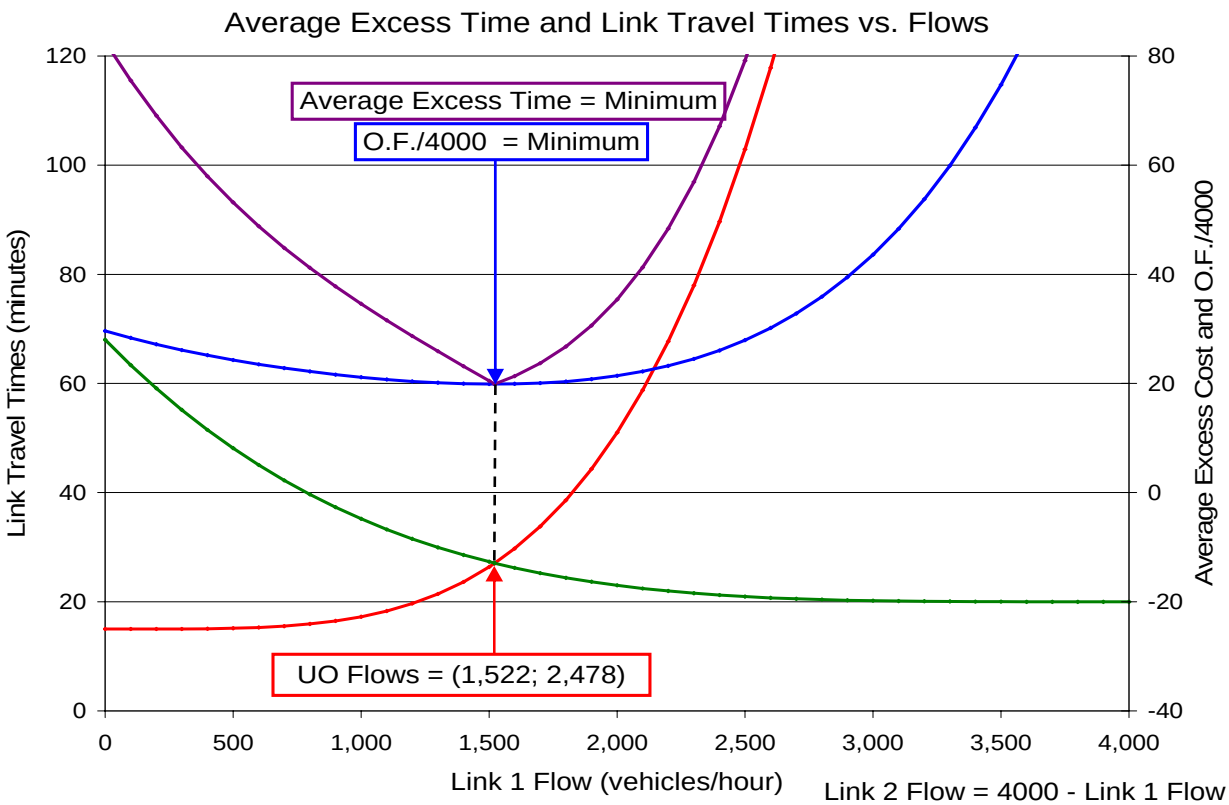



Objective Function and Link Travel Times vs. Flows
Link Travel Times (minutes)
Objective Function/4000
Objective Function/4000 = Minimum
User-Optimal Flows = (1,522; 2,478)
Link 1 Flow (vehcles/hour)
Link 2 Flow = 4000 - Link 1 Flow
0
20
40
60
80
100
120
0
8
16
24
32
40
48
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000


Total Excess Time vs. Link Flows
Link Travel Times (minutes)
Total Excess Time at (1900, 2100) =
(44.3 - 23.6) x 1900 = 39,300 minutes
Link 1 Flow (vehicles/hour)
Link 2 Flow = 4000 - Link 1 Flow
0
10
20
30
40
50
60
70
80
90
100
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000


Excess Times and Link Travel Times vs. Flows
Link Travel Times (minutes)
Total Excess Time (minutes) (1000s)
UO Flows = (1,522; 2,478)
Total Excess Time
Link 1 Flow (vehicles/hour)
Link 2 Flow = 4000 - Link 1 Flow
0
20
40
60
80
100
120
0
50
100
150
200
250
300
0
500
1000
1500
2000
2500
3000
3500
4000


Average Excess Time and Link Travel Times vs. Flows
Link Travel Times (minutes)
Average Excess Time (minutes)
UO Flows = (1,522; 2,478)
Average Excess Time
Link 1 Flow (vehicles/hour)
Link 2 Flow = 4000 - Link 1 Flow
0
20
40
60
80
100
120
0
20
40
60
80
100
120
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000



2.2 NCP and VIP Formulations

To illustrate a more general formulation of the above problem, let's consider travel cost functions that represent interaction between the two links. As a motivation for the example, consider that at the point where the two links diverge at point A that vehicles choosing Link 2 delay vehicles choosing Link 1. The generalized functions are as follows, where $\mathbf{f}^{\mathrm{T}} = (f_1 \;\; f_2)$:

$$c_1(\mathbf{f}) = 15\left(1 + .15\left(f_1^4 / 1000^4\right)\right) + 0.0012 f_2 = c_1(f_1) + 0.0012 f_2$$

$$c_2(\mathbf{f}) = 20\left(1 + .15\left(f_2^4 / 2000^4\right)\right) = c_2(f_2)$$

If we compute the Hessian for this objective function, we learn that it is not symmetric:

$$\frac{\partial c_1(\mathbf{f})}{\partial f_2} = 0.0012 \neq \frac{\partial c_2(\mathbf{f})}{\partial f_1} = 0$$

Therefore, the integral of the cost functions is not well-defined. To form the NCP, define:

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} c_1(\mathbf{f}) - u \\ c_2(\mathbf{f}) - u \\ f_1 + f_2 - d \end{bmatrix} = \begin{bmatrix} 15\left(1 + .15\left(f_1^4 / 1000^4\right)\right) + 0.0012 f_2 - u \\ 20\left(1 + .15\left(f_2^4 / 2000^4\right)\right) - u \\ f_1 + f_2 - d \end{bmatrix} \geq 0, \text{ and } \mathbf{x} = \begin{bmatrix} f_1 \\ f_2 \\ u \end{bmatrix} \geq 0.$$

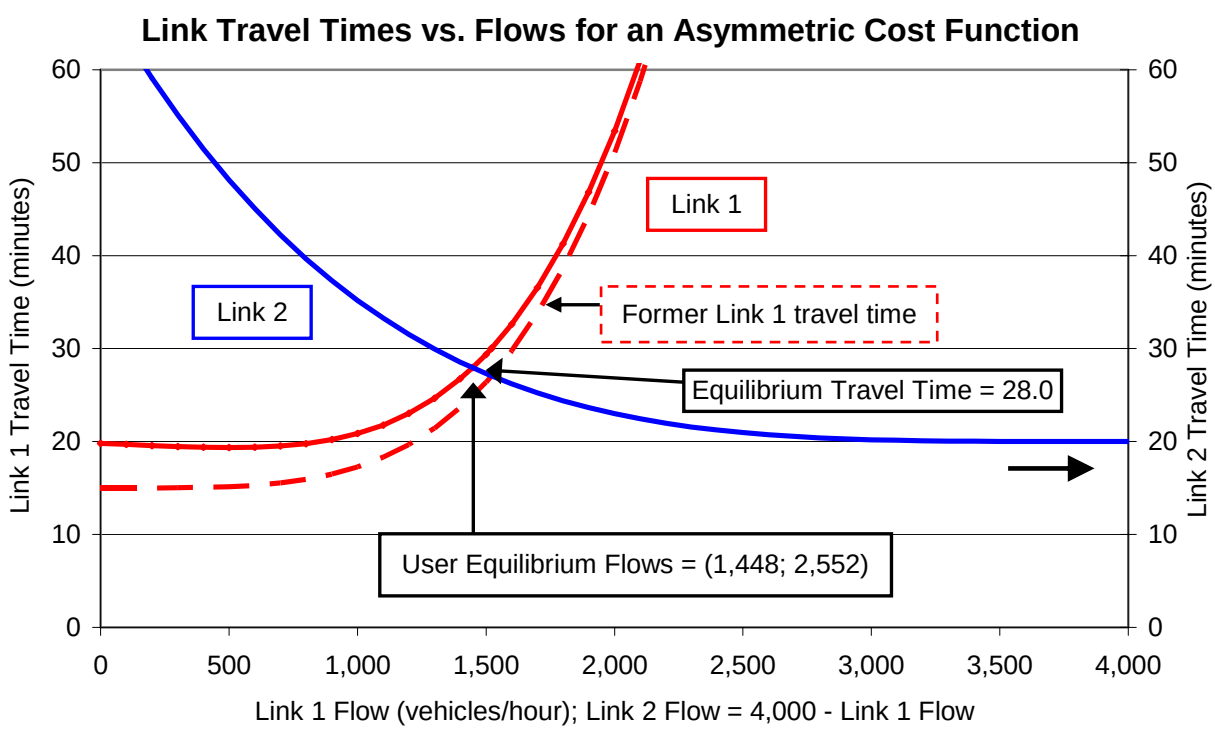
Solution of the NCP requires that the equilibrium values, $\mathbf{x}^*$ and u^* satisfy:

$$\mathbf{F}(\mathbf{x}^*) \cdot \mathbf{x}^* = \begin{bmatrix} 15\left(1 + .15\left(\left(f_1^*\right)^4 / 1000^4\right)\right) + 0.0012 f_2^* - u^* \\ 20\left(1 + .15\left(\left(f_2^*\right)^4 / 2000^4\right)\right) - u^* \\ f_1^* + f_2^* - d \end{bmatrix}^{\mathrm{T}} \bullet \begin{bmatrix} f_1^* \\ f_2^* \\ u^* \end{bmatrix} = 0$$

Performing this vector multiplication yields:

$$\mathbf{F}(\mathbf{x}^*) \cdot \mathbf{x}^* = \left(c_1\left(f_1^*\right) + 0.0012 f_2^* - u^*\right) \cdot f_1^* + \left(c_2\left(f_2^*\right) - u^*\right) \cdot f_2^* + \left(f_1^* + f_2^* - d\right) \cdot u^* = 0$$

By inspection, we can see that if the "asymmetric cost element," $0.0012 f_2$ were not present, the equation would be satisfied by: $c_1\left(f_1^*\right) = c_2\left(f_2^*\right) = u^*$ and $f_1^* + f_2^* = d$. The effect of the additional delay on link 1, therefore, is to shift some flow to link 2, thereby resulting in a new equilibrium solution. A figure showing the new equilibrium flows and travel time is shown below. The new equilibrium travel time is 28 minutes, and a net flow of 74 vehicles per hour is shifted from Link 1 to Link 2.



Also, note that the performance function for Link 1 is no longer strictly increasing.

Next, we consider the VIP formulation of this problem.

$$\begin{bmatrix} 15\left(1+.15\left(\left(f_1^*\right)^4/1000^4\right)\right)+0.0012 f_2^* - u^* \\ 20\left(1+.15\left(\left(f_2^*\right)^4/2000^4\right)\right)-u^* \\ f_1^* + f_2^* - d \end{bmatrix}^{\mathrm{T}} \bullet \left[\begin{pmatrix} f_1 \\ f_2 \\ u \end{pmatrix} - \begin{pmatrix} f_1^* \\ f_2^* \\ u^* \end{pmatrix}\right] \geq 0, \qquad \begin{pmatrix} f_1 \\ f_2 \\ u \end{pmatrix} \geq \mathbf{0}$$

where f_1, f_2, u are nonnegative scalar variables, and d is a constant. Rewriting the VIP as in Section A.5 of the Appendix, we obtain:

$$\begin{bmatrix} 15\left(1+.15\left(\left(f_1^*\right)^4/1000^4\right)\right)+0.0012 f_2^* - u^* \\ 20\left(1+.15\left(\left(f_2^*\right)^4/2000^4\right)\right)-u^* \\ f_1^* + f_2^* - d \end{bmatrix}^{\mathrm{T}} \bullet \begin{bmatrix} f_1 \\ f_2 \\ u \end{bmatrix} \geq \begin{bmatrix} 15\left(1+.15\left(\left(f_1^*\right)^4/1000^4\right)\right)+0.0012 f_2^* - u^* \\ 20\left(1+.15\left(\left(f_2^*\right)^4/2000^4\right)\right)-u^* \\ f_1^* + f_2^* - d \end{bmatrix}^{\mathrm{T}} \bullet \begin{bmatrix} f_1^* \\ f_2^* \\ u^* \end{bmatrix} = 0$$

The right-hand side of the inequality is the same as the equality in the NCP. The VIP states that all nonnegative values of the variables will result in a positive or zero left-hand side. A trivial example is $f_1 = f_2 = u = 0$.

Instead of requiring f_1, f_2, u to be nonnegative, suppose we also require f_1, f_2 to satisfy the conservation of flow constraint, $f_1 + f_2 = d$. Then we may remove the third line from the VIP, redefining the constraint space as follows:

$$\begin{bmatrix} 15\left(1+.15\left(\left(f_1^*\right)^4/1000^4\right)\right)+0.0012 f_2^* - u^* \\ 20\left(1+.15\left(\left(f_2^*\right)^4/2000^4\right)\right)-u^* \end{bmatrix}^{\mathrm{T}} \bullet \left[\begin{pmatrix} f_1 \\ f_2 \end{pmatrix} - \begin{pmatrix} f_1^* \\ f_2^* \end{pmatrix}\right] \geq 0, \ (f_1, f_2) \in \mathrm{F}$$

where $\mathrm{F} = \{ f_1 + f_2 = d; \mathbf{f} \geq \mathbf{0} \}$.

The VIP may then be written more compactly as follows, since the terms related to u^* are redundant:

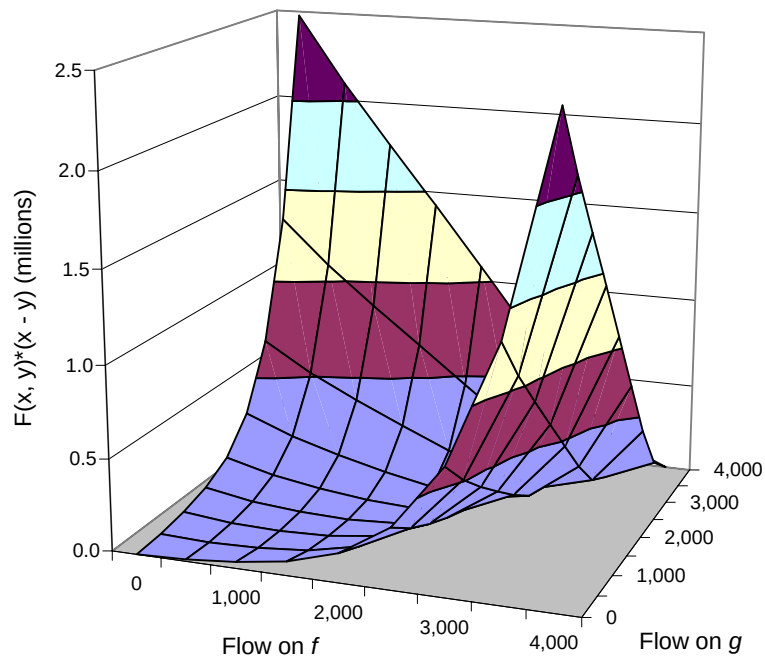
$$\mathbf{c}\left(\mathbf{f}^*\right)^{\mathrm{T}} \cdot \left(\mathbf{f} - \mathbf{f}^*\right) \geq \mathbf{0}, \mathbf{f} \in \mathrm{F}$$

A remaining question is whether the function is monotone on the constraint space F. The requirement for the function to be monotone may be stated as:

$$\left[\mathbf{c}(\mathbf{f}) - \mathbf{c}(\mathbf{g})\right]^{\mathrm{T}} (\mathbf{f} - \mathbf{g}) \geq 0, \forall \mathbf{f}, \mathbf{g} \in \mathrm{F}.$$

To check this condition, the above statement was evaluated for $(0 \le f_1, g_1 \le 4000)$, since $f_2 = 4{,}000 - f_1$ and $g_2 = 4{,}000 - g_1$. These values were systematically varied, and plotted as shown below. As can be readily visualized with the plot, the vector product is nonnegative.

Plot of Monotonicity Condition for Existence and Uniqueness



2.3 System-Optimal Route Choice

Wardrop also stated a second criterion: "The average journey time is a minimum." Today, we refer to this solution as *system-optimal route choice*. It has largely been a matter of theoretical analysis, until renewed interest in road pricing has stimulated additional research on this concept. The reason is that achieving system optimal flows requires that the flows be *directed* along routes in some manner. However, there is a direct interpretation of system-optimal route costs in terms of congestion tolls. We explore this interpretation below.

The system-optimal (SO) problem is formulated *directly*, in contrast to the *indirect* formulation of the user-optimal (UO) problem, as follows:

$$\min_{(f_1, f_2)} \sum_a c_a(f_a) \cdot f_a$$
$$\text{s.t.}: f_1 + f_2 = d$$
$$\mathbf{f} \ge \mathbf{0}$$

To solve this problem for the system-optimal flows $\boldsymbol{f}^s$, we proceed as follows, letting v denote the Lagrange multiplier associated with the conservation of flow constraint:

$$L(\mathbf{f},v)=\sum_a c_a(f_a)\cdot f_a - v(f_1+f_2-d)$$
$$\frac{\partial L}{\partial f_1}=c_1(f_1)+f_1\cdot\frac{\partial c_1(f_1)}{\partial f_1}-v(+1)\geq 0$$
$$\frac{\partial L}{\partial f_2}=c_2(f_2)+f_2\cdot\frac{\partial c_2(f_2)}{\partial f_2}-v(+1)\geq 0$$
$$\frac{\partial L}{\partial v}=f_1+f_2-d\geq 0$$
$$f_1\left(c_1(f_1)+f_1\cdot\frac{\partial c_1(f_1)}{\partial f_1}-v\right)=0$$
$$f_2\left(c_2(f_2)+f_2\cdot\frac{\partial c_2(f_2)}{\partial f_2}-v\right)=0$$
$$v(f_1+f_2-d)=0$$
$$\mathbf{f}\geq\mathbf{0},\ v\geq 0$$

For ease of exposition, define: $m_a(f_a)\equiv c_a(f_a)+f_a\cdot\frac{\partial c_a(f_a)}{\partial f_a}$. Since $m_a(f_a)$ is the derivative of the total cost of link *a*, $(c_a(f_a)\cdot f_a)$, with respect to the flow, f_a, then $m_a(f_a)$ is by definition the *Marginal Cost* of link *a* at flow f_a.

To analyze the optimality conditions, consider the case, $\mathbf{f}>\mathbf{0}$. Then, $m_1(f_1^S)=v=m_2(f_2^S)$, where (f_a^S) are the system-optimal flows.

Since $v>0$ by definition of $m_a(f_a)$, then $(f_1+f_2=d)$, as required. Alternatively, suppose $(f_1>0,\ f_2=0)$. Then,

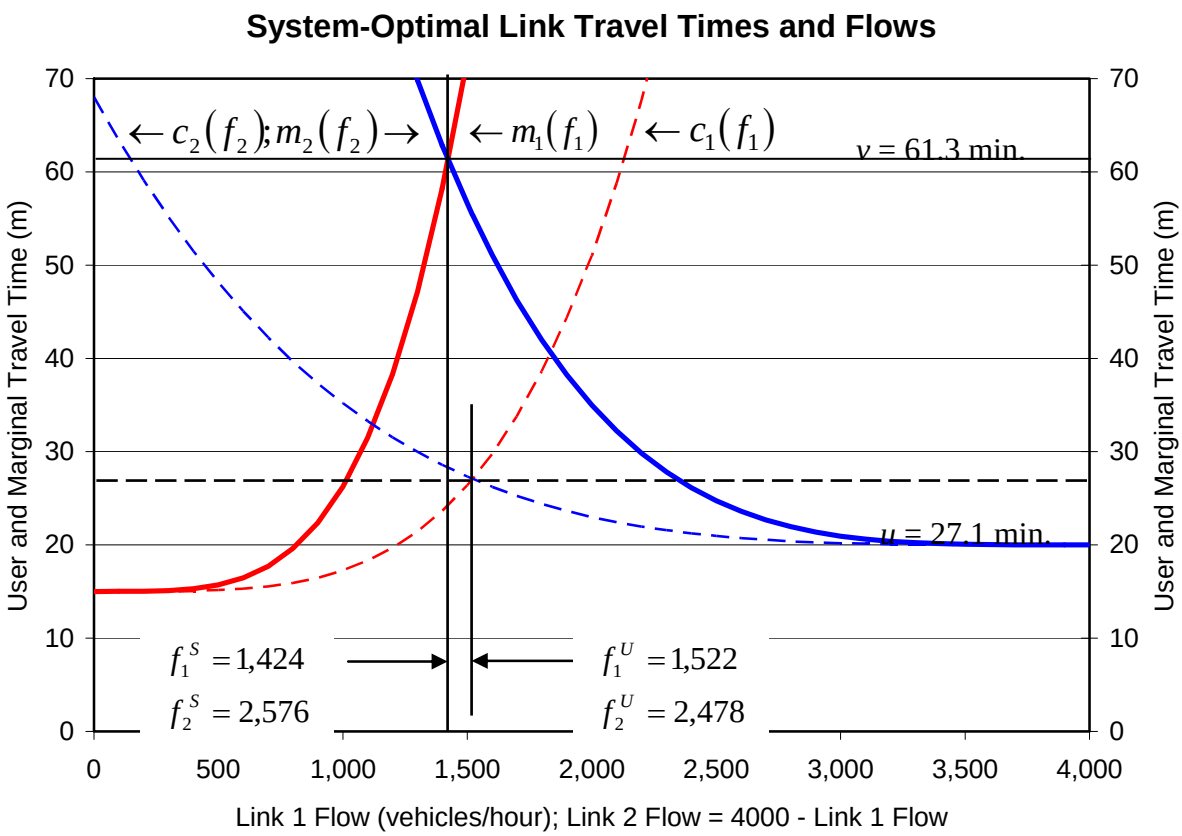
$$m_1(f_1^S)=v$$
$$m_2(f_2^S)\geq v$$
$$v>0\Rightarrow f_1^S=d$$

Finally, suppose $m_2(f_2^S)>v$; then, by the complementary slackness condition on f_2, above, $f_2^S=0$ and $f_1^S=d$. Next consider the value of $m_a(f_a)$ based on the BPR function:

$$T_a(f_a)=c_a^0\left(f_a+0.15\left(\frac{f_a^5}{C_a^4}\right)\right)$$
$$m_a(f_a)\equiv\frac{\partial T_a(f_a)}{\partial f_a}=c_a^0\left(1+0.75\left(\frac{f_a^4}{C_a^4}\right)\right)=c_a^0\left(1+0.15\left(\frac{f_a^4}{C_a^4}\right)\right)+c_a^0\left(0.60\left(\frac{f_a^4}{C_a^4}\right)\right)$$

From the definition of the Marginal Cost, the derivative of the Total Cost with respect to flow, we may learn that the Marginal Cost is the increase in Total Cost resulting from one additional unit of flow. The first term of the right side of the above equation is the cost incurred by that additional unit of flow. The second term on the right is the additional cost imposed on all f_a vehicles by the additional unit of flow. The principle of marginal cost road pricing is that each unit of flow (vehicle) should be charged a toll equal to the cost it imposes on all others. This situation may be visualized in the following figure showing the SO flows determined by equating the marginal link costs.

By examining the figure carefully, we may observe the differences in generalized cost for the SO and UO solutions. Note for Link 1 that the travel time incurred at the SO flow is less than the UO travel time; however, for Link 2, the UO travel time is greater than the SO time. A toll charged to vehicles on Link 1 equal to $\left(c_2\left(f_2^S\right) - c_1\left(f_1^S\right)\right)$ would adjust the total generalized costs of the two flows to equality at the SO flows. This toll is 4.0 minutes per vehicle.



2.4 Solution Algorithm for the Two-Link Problem based on the Frank-Wolfe Method

Next we address the problem of solving for the optimal flows computationally. We first examine this problem for two links. Consider at iteration *n* that we have a feasible solution $\left(f_1^n, f_2^n\right)$, where $\left(f_1^n + f_2^n = d\right)$. For the UO problem, the objective function is: $z\left(\mathbf{f}^n\right) = \sum_a \int_0^{f_a^n} c_a(x)dx$.

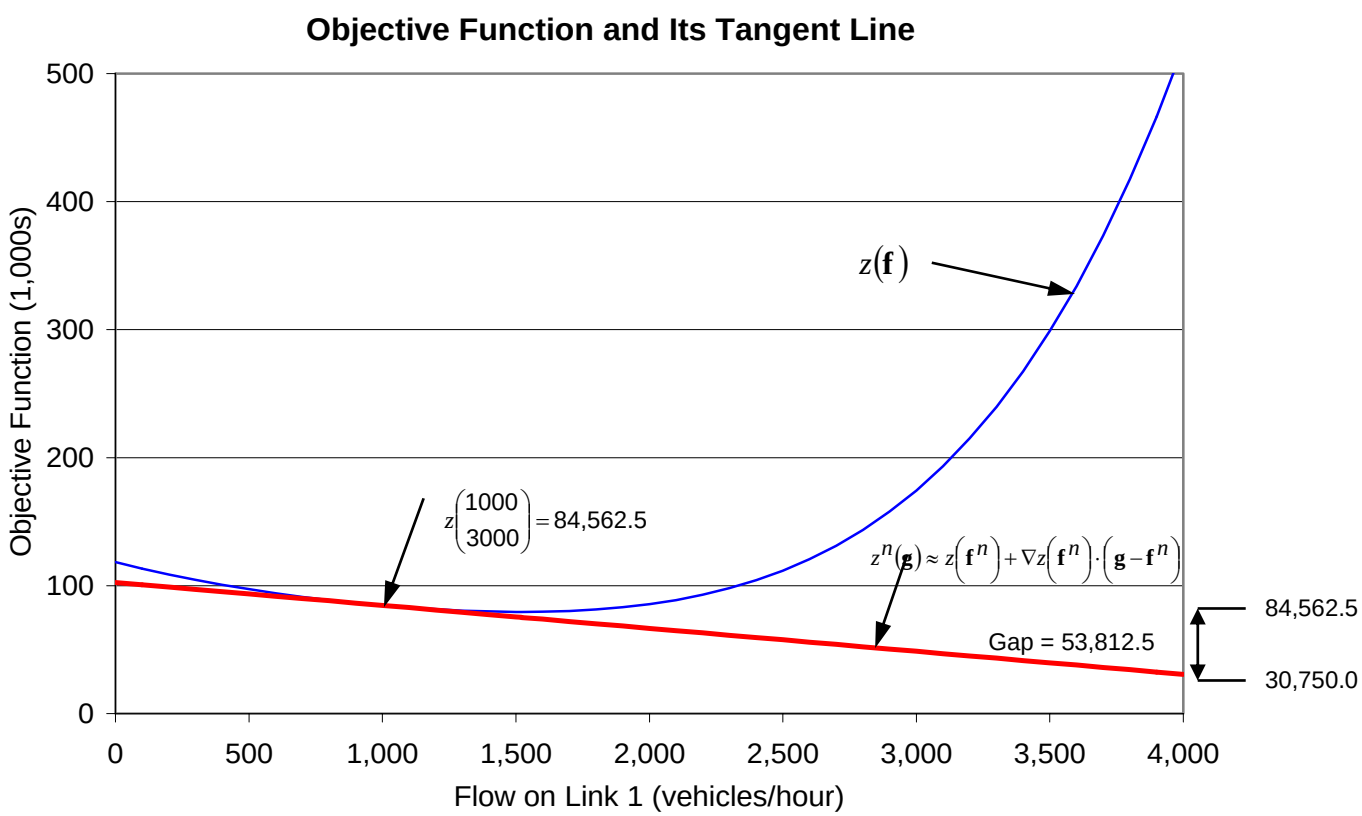
First, we approximate the objective function at iteration *n* by its first-order Taylor series,

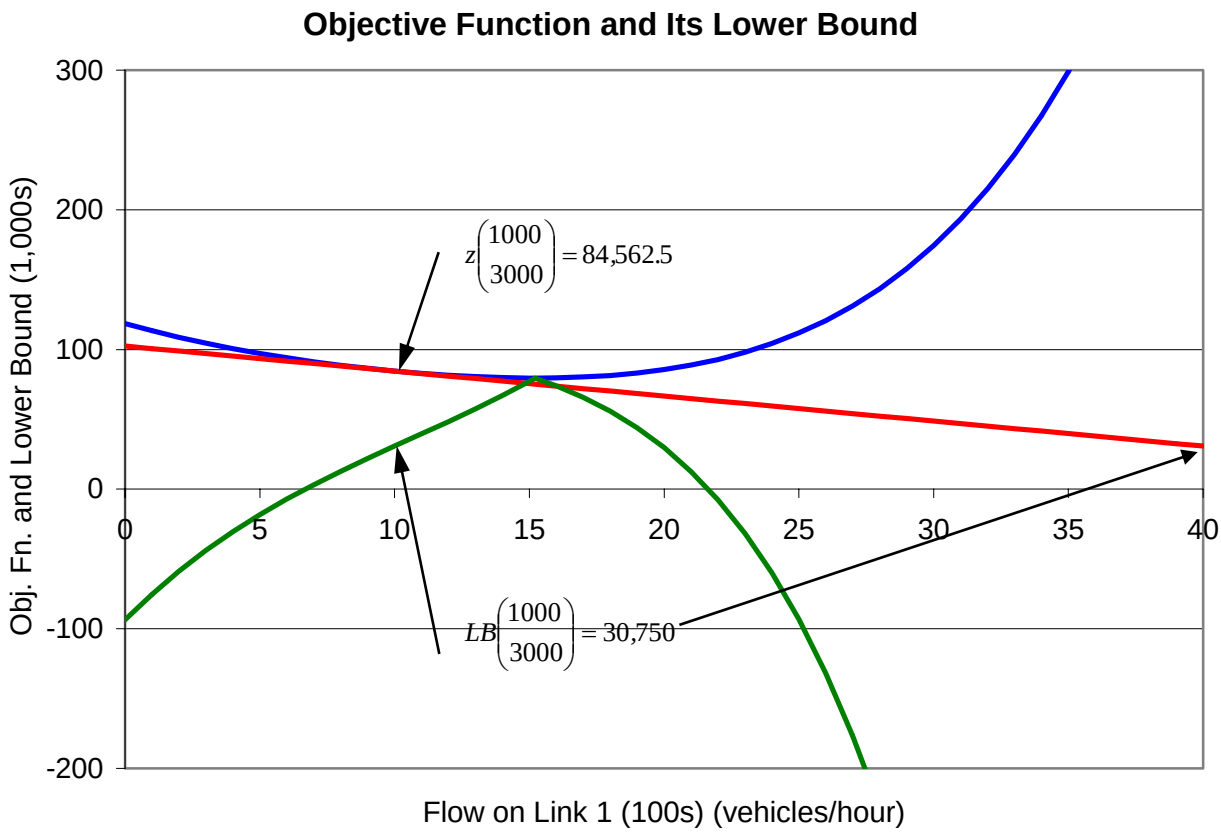
$$z^n(\mathbf{g}) \approx z(\mathbf{f}^n) + \nabla z(\mathbf{f}^n)\cdot(\mathbf{g}-\mathbf{f}^n) = z(\mathbf{f}^n) + \begin{pmatrix} c_1(f_1^n) \\ c_2(f_2^n) \end{pmatrix}^{\mathrm{T}} \cdot \begin{pmatrix} g_1 - f_1^n \\ g_2 - f_2^n \end{pmatrix} = \begin{pmatrix} c_1(f_1^n) - c_2(d - f_1^n) \\ 0 \end{pmatrix}^{\mathrm{T}} \cdot \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} + C .$$

Note we substitute $(d - f_1^n)$ for f_2^n. Then, we move from the current solution $\mathbf{f}^n = (f_1^n \; f_2^n)^{\mathrm{T}}$ in the direction of a decrease in $z(\mathbf{f}^n)$ pointed by the gradient $\nabla z(\mathbf{f}^n)$, called the *Search Direction*, to a new solution $\mathbf{g}^n = (g_1^n \; g_2^n)^{\mathrm{T}}$. We will refer to $\mathbf{f}^n$ as the Main Problem solution at iteration *n*, and $\mathbf{g}^n$ as the Subproblem solution at iteration *n*. This movement is along the *tangent line*:

$$\begin{aligned} z(\mathbf{f}^n) + \nabla z(\mathbf{f}^n)\cdot(\mathbf{g}^n - \mathbf{f}^n) &= \sum_a \int_0^{f_a^n} c_a(x)dx + c_1(f_1^n)\cdot(g_1^n - f_1^n) + c_2(f_2^n)\cdot(g_2^n - f_2^n) \\ &= \sum_a \int_0^{f_a^n} c_a(x)dx + \left[c_1(f_1^n) - c_2(d - f_1^n)\right]\cdot(g_1^n - f_1^n) \end{aligned} ,$$

as shown in the next figure. The convex line shows the Objective Function, with a tangent line at (1000, 3000). The Gap is the difference between the Objective Function and the Subproblem solution. In the next figure, the concave line is the Lower Bound on the Objective Function, which equals the Objective Function plus the Gap, whose value is non-positive.





The Subproblem can be formulated as follows:

$$\min_{(\mathbf{g})} \bar{z}(\mathbf{g}) = \left(c_1\left(f_1^n\right)\cdot\left(g_1^n - f_1^n\right) + c_2\left(f_2^n\right)\cdot\left(g_2^n - f_2^n\right)\right)$$
$$\text{st}: g_1^n + g_2^n = d$$
$$\mathbf{g}^n \geq \mathbf{0}$$

Since $\left(\mathbf{f}^n\right)$ is fixed, we can solve this problem simply by identifying the lower cost link:

$$c_{\hat{a}} = \min_a \left[c_1\left(f_1^n\right), c_2\left(f_2^n\right)\right]$$

Then, we allocate (assign) the total flow d to that link, in order to minimize the total cost. That is, we set $g_{\hat{a}}^n = d$ and $g_a^n = 0, a \neq \hat{a}$. Suppose $\hat{a} = 1$. The resulting value of $\bar{z}\left(\mathbf{g}^n\right)$ is:

$$\bar{z}\left(\mathbf{g}^{\mathbf{n}}\right) = c_1\left(f_1^n\right)\cdot\left(d - f_1^n\right) + c_2\left(f_2^n\right)\cdot\left(0 - f_2^n\right)$$

Since $c_1\left(f_1^n\right) < c_2\left(f_2^n\right)$, then $c_1\left(f_1^n\right)\cdot d - c_1\left(f_1^n\right)\cdot f_1^n - c_2\left(f_2^n\right)\cdot f_2^n \leq 0$. Why?

Because $c_1\left(f_1^n\right)\cdot\left(f_1^n + f_2^n\right) - c_1\left(f_1^n\right)\cdot f_1^n - c_2\left(f_2^n\right)\cdot f_2^n = \left[c_1\left(f_1^n\right) - c_2\left(f_2^n\right)\right]\cdot f_2^n \leq 0.$

The value of $\bar{z}(\mathbf{g}^n)$ achieved by the solution (d, 0) is called the *Gap*, which is the negative of the *Total Excess Cost*, which as we have seen has an intuitive interpretation. In general, define

$$T(\mathbf{f}^n) = \sum_a c_a(f_a^n) \cdot f_a^n = \text{ total travel cost at the current solution at iteration } n$$
$$T(\mathbf{g}^n) = \sum_a c_a(f_a^n) \cdot g_a^n = \text{ total travel cost of the best solution, } \mathbf{g}^n.$$

Then, $\bar{z}(\mathbf{g}^n) = \sum_a c_a(f_a^n) \cdot g_a^n - \sum_a c_a(f_a^n) \cdot f_a^n \le 0.$

For our two-link example, the Gap and Total Excess Cost calculations are performed as follows:

$$\begin{aligned}\text{Gap} &= 17.25 \cdot 4{,}000 + 35.1875 \cdot 0 - (17.25 \cdot 1{,}000 + 35.1875 \cdot 3{,}000) \\ &= 69{,}000 - (17{,}250 + 105{,}562.5) = -53{,}812.5\end{aligned}$$

$$\text{Total Excess Cost} = (35.1875 - 17.25) \cdot 3{,}000 = 53{,}812.5$$

Returning to the figure, note that the function lies above the tangent line. Since the objective function is convex, it must lie everywhere above the tangent line at any point. Accordingly, the true optimal solution must be greater than the current objective function value plus the Gap, since the Gap is defined to be ≤ 0. If the Gap is 0, then the current solution is optimal. The following expression summarizes these points:

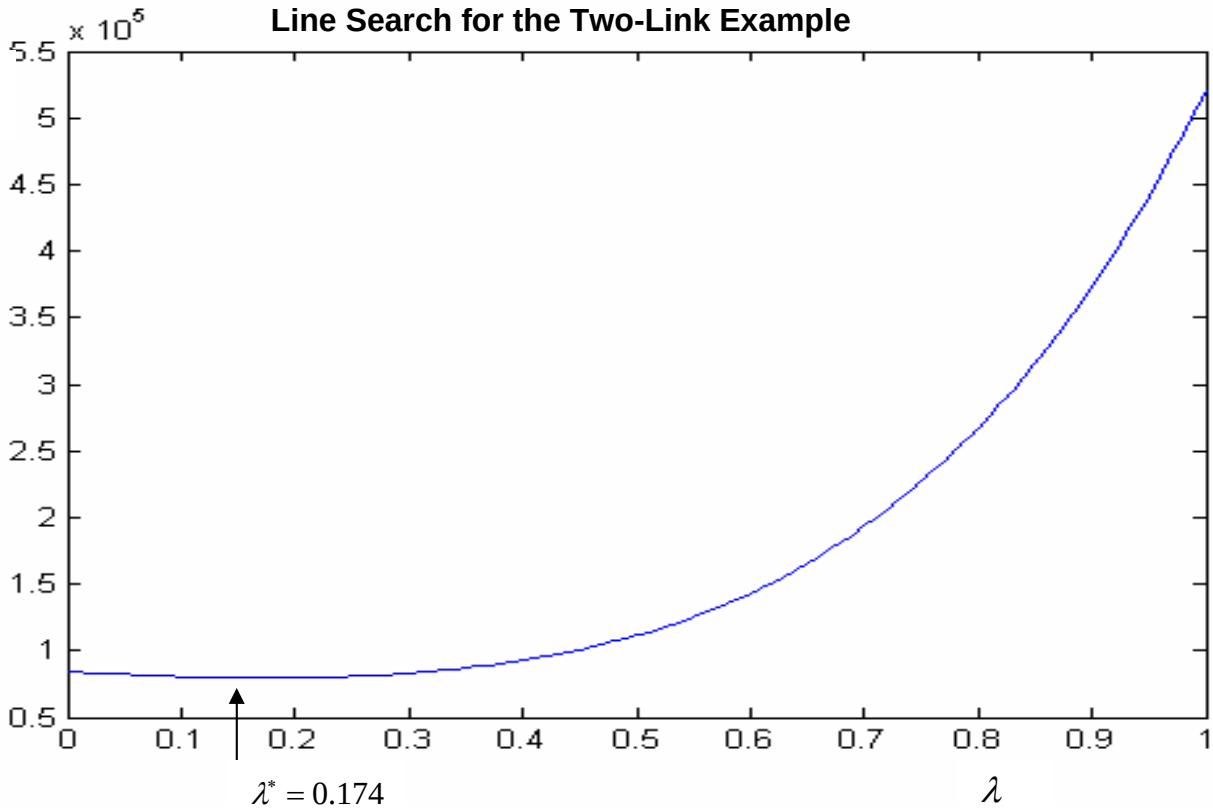
$$z(\mathbf{f}^*) \ge z(\mathbf{f}^n) + \sum_a c_a(f_a) \cdot (g_a^n - f_a^n) \le z(\mathbf{f}^n)$$

For the two-link problem, comparing the cost of the two links tells us whether to shift flow from Link 1 to Link 2 or the opposite. The next question is how much flow to shift, called the *Step Size Problem*, which is stated as follows:

$$\min_{(0 \le \lambda \le 1)} z(\lambda) = \sum_a \int_0^{f_a^{n+1}} c_a(x)dx, \qquad \text{where } f_a^{n+1} \equiv f_a^n + \lambda \cdot (g_a^n - f_a^n) = (1-\lambda) \cdot f_a^n + \lambda \cdot g_a^n$$

We typically solve this problem with a method called *bisection search*, described as follows:

1. initialize $\lambda_1 = 0.5$; set a counter $k = 1$; define $\lambda_0 = 0.0$.
2. find $z'(\lambda_k) = \dfrac{\partial z(\mathbf{f}^{n+1})}{\partial \lambda} = \sum_a c_a(f_a^{n+1}) \cdot (g_a^n - f_a^n) =$ slope of the function $z(\mathbf{f}^{n+1})$.
   a. if $z'(\lambda_k) > 0$, $\lambda_{k+1} = \lambda_k - (\lambda_k - \lambda_{k-1})/2$
   b. if $z'(\lambda_k) < 0$, $\lambda_{k+1} = \lambda_k + (\lambda_k - \lambda_{k-1})/2$
3. stop when $(\lambda_{k+1} - \lambda_k) < \varepsilon$, and $z'(\lambda_k) < 0$ at least once; set $\lambda^* = \lambda_{k+1}$.



The entire solution algorithm may be stated as follows:

1. initialize $\mathbf{f}^1$; set a counter to $n = 1$;

2. solve the Subproblem for $\mathbf{g}^n$;

3. check for convergence: Is the Relative Gap at iteration n, $\mathrm{RG}^n \equiv \left| \dfrac{z(\mathbf{f}^n) - \mathrm{BLB}^n}{\mathrm{BLB}^n} \right| < \varepsilon$?

   where $\mathrm{BLB}^n \equiv \max\limits_{n' \le n} \left(z(\mathbf{f}^{n'}) + \nabla z(\mathbf{f}^{n'}) \cdot (\mathbf{g}^{n'} - \mathbf{f}^{n}) \right) \equiv \max\limits_{n' \le n} \left(z(\mathbf{f}^{n'}) + \mathrm{Gap}^{n'} \right)$ = Best Lower Bound at n

   $\mathrm{Gap}^n = \nabla z(\mathbf{f}^n) \cdot (\mathbf{g}^{\mathbf{n}} - \mathbf{f}^n)$

4. perform line search to obtain λ^* and update $\mathbf{f}^n$ to $\mathbf{f}^{n+1}$, where $f_a^{n+1} = (1 - \lambda^*) \cdot f_a^n + \lambda^* \cdot g_a^n$

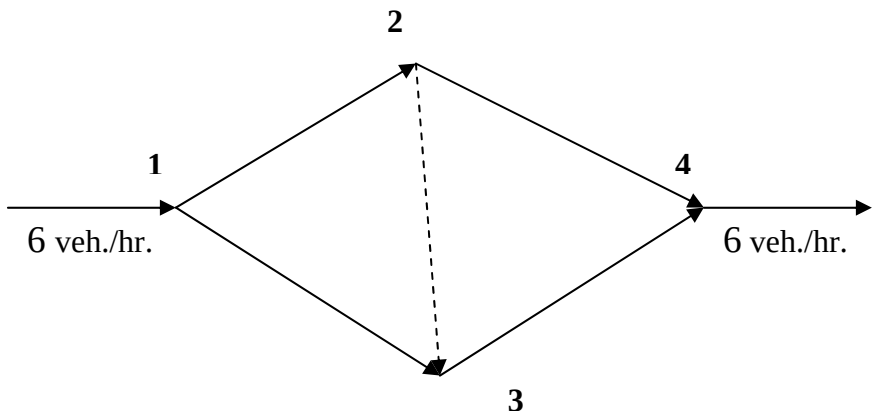
5. check for convergence again with the new value of the objective function $z(\mathbf{f}^{n+1})$; if not converged, update the flows according to the above equation, increment n and go to step 2.

This solution method is based on a procedure first proposed by M. Frank and P. Wolfe (1956). For this reason, it is generally referred to as the Frank-Wolfe method or algorithm.

2.5 Braess's Paradox

Several apparent paradoxes of traffic assignment problem have been identified, the most famous being known as Braess's Paradox. The network shown below (LeBlanc, 1975) exhibits Braess's Paradox when link (2, 3) is added to the network, which may be stated as follows:

> If the route flows are user-optimal, there may exist additions of new links that result in a new User-Optimal solution with a *higher* total user cost.



link	cost functions
$(1,2)$	$40. + 0.5 f_{12}^4$
$(3,4)$	$40. + 0.5 f_{34}^4$
$(1,3)$	$185. + 0.9 f_{13}^4$
$(2,4)$	$185. + 0.9 f_{24}^4$
$(2,3)$	$15.4 + 1.0 f_{23}^4$

Without link (2,3), the UO flows on each link are 3 vehicles/hour. Are these the system-optimal flows as well? Add link (2, 3) and determine the UO and SO flows, route costs and total costs. The solutions for flows ranging from 1 to 12 vehicles per hour are given at the end of this section.

References

Bar-Gera, H. (2002) Origin-Based Algorithm for the Traffic Assignment Problem, *Transportation Science* **36**, 398-417.

Braess, D., A. Nagurney and T. Wakolbinger (2005) On a Paradox of Traffic Planning, *Transportation Science* **39**, 446-450 (translated from the original German published in 1968).

Frank, M. and P. Wolfe (1956) An Algorithm for Quadratic Programming, *Naval Research Logistics Quarterly* **3**, 95-110.

LeBlanc, L. J. (1975) An Algorithm for the Discrete Network Design Problem, *Transportation Science* **9**, 183-199

Nagurney, A. and D. Boyce (2005) Preface to "On a Paradox of Traffic Planning," *Transportation Science* **39**, 443-445.

Wardrop, J. G. (1952) Some Theoretical Aspects of Road Traffic Research, *Proceedings of the Institution of Civil Engineers, Part II* **1**, 325-378.

Exercises

1. Consider the following travel cost functions for the two-link problem shown in Section 2.1.2.

$$c_1(f_1) = 15\left(1 + 0.15\left(\frac{f_1}{1000}\right)^4\right) \qquad c_2(f_2) = 20\left(1 + 0.15\left(\frac{f_2}{2000}\right)^4\right)$$

Solve graphically for the solutions of the user-optimal and system-optimal problems for the two-link network with d = 5,000 vehicles/hour. Plot the user travel time functions and the marginal travel time functions in the region of their intersection to yield a good graphical solution.

Compute the total travel time for both solutions. How much time is saved by the system-optimal solution? What is the difference in the users' travel times on the two links at the system-optimal solution? How do these times compare with the users' travel time for the user-optimal solution? Be careful: compare the users' travel times, not their marginal times.

2. Apply the Frank-Wolfe algorithm, described in the notes, to solve for the UO flows for the above two-link problem plus the addition of a third link with the following travel cost function:

$$c_3(f_3) = 25\left(1 + 0.15\left(\frac{f_3}{500}\right)^4\right)$$

For your initial solution, take the UO solution that you found for Exercise 1 above. Compute the travel cost for the following links: $c_1\left(f_1^U\right), c_2\left(f_2^U\right), c_3(0)$, where $\left(f_1^U, f_2^U\right)$ is the solution to Exercise 1. Choose the minimum cost link of these three. Compute the Objective Function, Gap and Lower Bound on the Objective Function using the work sheet given below. Note that the Gap is negative by definition, except at the optimum, where it is zero. Therefore, the Lower Bound should be less than the Objective Function. Perform the line search step to find λ^*:

$$\min_{(0 \le \lambda \le 1)} z(\lambda) = \sum_a \int_0^{f_a^{n+1}} c_a(x)dx, \qquad \text{where } f_a^{n+1} \equiv f_a^n + \lambda \cdot \left(g_a^n - f_a^n\right) = (1-\lambda) \cdot f_a^n + \lambda \cdot g_a^n$$

You may find λ^* in one of two ways. One way is to compute $z(\lambda)$ for several values of λ, and plot $z(\lambda)$ versus λ, and interpolate the result. If you choose this method, you must find λ sufficiently precisely so that you obtain a decrease in the Objective Function value. For each iteration, you will need to increase the precision of λ from tenths to hundredths to thousandths. The other procedure is to apply the bisection method. In this case, you must continue the method until you obtain a negative value for the slope of the function $z(\lambda)$.

Tabulate your computations for three iterations of the algorithm following the initial solution. That is, solve for three values of λ*. Plot the objective function and lower bound (y-axis) vs. the iteration number (x-axis).

Worksheet for Frank-Wolfe Method, Exercise 2

link	f_a^1	$c_a\left(f_a^1\right)$	$\int_0^{f_a^1} c_a(x)dx$	g_a^1	$\left(g_a^1 - f_a^1\right)$	$c_a\left(f_a^1\right)\cdot\left(g_a^1 - f_a^1\right)$	f_a^2
1							
2							
3							
	Objective Function		**=**	**5,000**	**Gap**	**=**	**5,000**
Lower Bound = Objective Function + Gap						**=**	
Relative Gap = Gap/Best Lower Bound						**=**	

link	f_a^2	$c_a\left(f_a^2\right)$	$\int_0^{f_a^2} c_a(x)dx$	g_a^2	$\left(g_a^2 - f_a^2\right)$	$c_a\left(f_a^2\right)\cdot\left(g_a^2 - f_a^2\right)$	f_a^3
1							
2							
3							
	Objective Function		**=**	**5,000**	**Gap**	**=**	**5,000**
Lower Bound = Objective Function + Gap						**=**	
Relative Gap = Gap/ Best Lower Bound						**=**	

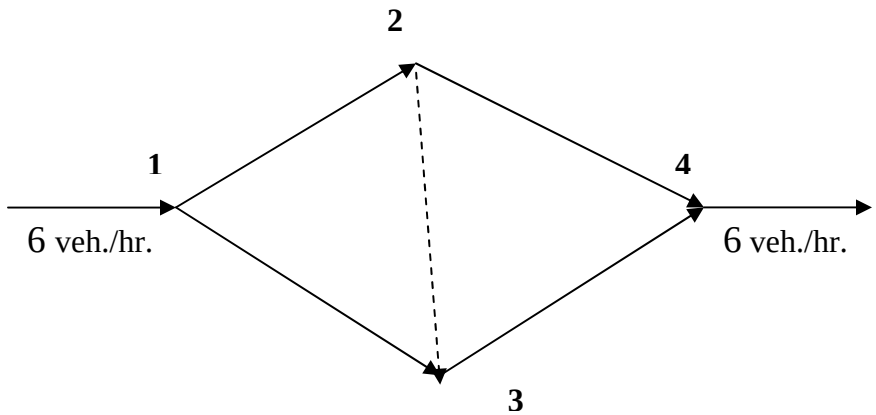
link	f_a^3	$c_a\left(f_a^3\right)$	$\int_0^{f_a^3} c_a(x)dx$	g_a^3	$\left(g_a^3 - f_a^3\right)$	$c_a\left(f_a^3\right)\cdot\left(g_a^3 - f_a^3\right)$	f_a^4
1							
2							
3							
	Objective Function		**=**	**5,000**	**Gap**	**=**	**5,000**
Lower Bound = Objective Function + Gap						**=**	
Relative Gap = Gap/ Best Lower Bound						**=**	

3. This exercise explores the properties of Braess's Paradox. It also introduces networks with routes of more than one link.

a. Solve the Braess's paradox network shown below for the user-optimal flows. You can do this by applying the Frank-Wolfe method, or you may be able to do it more quickly by trial and error. In the latter case, begin with the network without link (2,3) and solve for the flows by noting the symmetry of the network. If the route costs are equal, then you have found the user-optimal solution, since the solution for this network is unique.

<u>link cost functions</u>

$(1,2)$	$40. + 0.5 f_{12}^4$
$(3,4)$	$40. + 0.5 f_{34}^4$
$(1,3)$	$185. + 0.9 f_{13}^4$
$(2,4)$	$185. + 0.9 f_{24}^4$
$(2,3)$	$15.4 + 1.0 f_{23}^4$



b. Add link (2,3) and try moving 1 vehicle/hour to the new lower cost route that includes link (2,3); continue until you find the user-optimal solution, that is the solution in which all three routes have equal travel costs.. Check your answer by comparing it with the result in the Table, at the end of Section 2. Compare the UO route costs for the two solutions, and note that the network with link (2,3) has a higher UO cost.

c. Derive the marginal cost functions for the links of this network. Using these marginal cost functions, solve the system-optimal problem for a total flow of 6 vehicles/hour with and without link (2,3). What is the total travel time for each solution? How do these travel times compare with the total travel time for the user-optimal solutions with and without link (2,3)?

To solve the SO problem with link (2,3) you will need to consider non-integer link flows. Therefore, you cannot use trial and error, since the number of possibilities is infinite. You can solve this problem using the Frank-Wolfe method, but it is somewhat more involved than the simple three-link problem you solved so far. The reason is that you need to work with link flows and route flows, not just link flows. More formally, you need to work in two different spaces, link flow space and route flow space.

The tables below and on the next pages are provided to facilitate your work. Begin with the route table and enter the solution without link (2, 3). Compute the link costs in the link table, and then determine the route costs.

d. Now re-solve for the UO and SO flows with a total flow of 12 vehicles/hour entering and leaving the network. For this problem, start with the solution without link (2,3), and then apply

the Frank-Wolfe method for one iteration. If you calculate the value of λ carefully, I believe you will obtain the answer in the table.

What can you conclude about Braess's paradox from your answer, and from examining the table and figures at the end of Section 2?

Iteration 1 – Search Direction

link space	f_a^1	$c_a(f_a^1)$	$\int_0^{f_a^1} c_a(x)dx$	g_a^1	$(g_a^1 - f_a^1)$	$c_a(f_a^1)\cdot(g_a^1 - f_a^1)$	f_a^2
1,2							
2,4							
1,3							
3,4							
2,3							
	OF =				**Gap =**		
Lower Bound = Objective Function + Gap =							
Relative Gap = Gap/ Best Lower Bound =							

route space	main problem route flow h_r^1	links in route	route cost C_r^1	subproblem route flow m_r^1	$(m_r^1 - h_r^1)$	$C_r^1(m_r^1 - h_r^1)$
1						
2						
3						
	d			***d***	**Gap** =	

Iteration 1 –Step Size

link	f_a^2 $(\lambda = 0.0)$	$\int_0^{f_a^2} c_a(x)dx$	f_a^2 $(\lambda = 0.1)$	$\int_0^{f_a^2} c_a(x)dx$	f_a^2 $(\lambda = 0.2)$	$\int_0^{f_a^2} c_a(x)dx$	f_a^2 $(\lambda = 0.3)$	$\int_0^{f_a^2} c_a(x)dx$
1,2								
2,4								
1,3								
3,4								
2,3								
	OF =		**OF =**		**OF =**		**OF =**	

The Objective Function can only be directly computed in terms of links flows. The Gap can be computed in terms of either link flows or route flows. The relationship of the computation in terms of route flows to the computation in link flows is derived as follows:

$$\sum_a c_a\left(f_a^n\right)\cdot\left(g_a^n - f_a^n\right) = \sum_a c_a\left(f_a^n\right)\cdot\left(\sum_r m_r^n \delta_r^a - \sum_r h_r^n \delta_r^a\right) = \sum_a c_a\left(f_a^n\right)\cdot\sum_r \delta_r^a\left(m_r^n - h_r^n\right)$$
$$= \sum_r\sum_a c_a\left(f_a^n\right)\delta_r^a\cdot\left(m_r^n - h_r^n\right) = \sum_r C_r\cdot\left(m_r^n - h_r^n\right)$$

where $\delta_r^a = 1$ if link *a* belongs to route r, and 0 otherwise, and m_r^n = the Subproblem route flow. So, if you have the route costs and route flows for the current Main Problem solution, you can compute the gap using the first table above.

Iteration 2 – Search Direction

link space	f_a^2	$c_a\left(f_a^2\right)$	$\int_0^{f_a^2} c_a(x)dx$	g_a^2	$\left(g_a^2 - f_a^2\right)$	$c_a\left(f_a^2\right)\cdot\left(g_a^2 - f_a^2\right)$	f_a^3
1,2							
2,4							
1,3							
3,4							
2,3							
	OF =				**Gap =**		
Lower Bound = Objective Function + Gap =							
Relative Gap = Gap/ Best Lower Bound =							

route space	main problem route flow h_r^2	links in route	route cost C_r^2	subproblem route flow m_r^2	$\left(m_r^2 - h_r^2\right)$	$C_r^2\left(m_r^2 - h_r^2\right)$
1						
2						
3						
	d			***d***	**Gap =**	

Iteration 2 –Step Size

link	f_a^3 $(\lambda = 0.0)$	$\int_0^{f_a^3} c_a(x)dx$	f_a^3 $(\lambda = \quad)$	$\int_0^{f_a^3} c_a(x)dx$	f_a^3 $(\lambda = \quad)$	$\int_0^{f_a^3} c_a(x)dx$	f_a^3 $(\lambda = \quad)$	$\int_0^{f_a^3} c_a(x)dx$
1,2								
2,4								
1,3								
3,4								
2,3								
	OF =		**OF =**		**OF =**		**OF =**	

Iteration 3 – Search Direction

route space	main problem route flow h_r^3	links in route	route cost C_r^3	subproblem route flow m_r^3	$\left(m_r^3 - h_r^3\right)$	$C_r^3\left(m_r^3 - h_r^3\right)$
1						
2						
3						
	d			***d***	**Gap** =	

link space	f_a^3	$c_a\left(f_a^3\right)$	$\int_0^{f_a^3} c_a(x)dx$	g_a^3	$\left(g_a^3 - f_a^3\right)$	$c_a\left(f_a^3\right)\cdot\left(g_a^3 - f_a^3\right)$	f_a^4
1,2							
2,4							
1,3							
3,4							
2,3							
	OF =		**=**		**Gap =**		
Lower Bound = Objective Function + Gap =							
Relative Gap = Gap/ Best Lower Bound =							

Iteration 3 –Step Size

link	f_a^4 $(\lambda = 0.0)$	$\int_0^{f_a^4} c_a(x)dx$	f_a^4 $(\lambda = \quad)$	$\int_0^{f_a^4} c_a(x)dx$	f_a^4 $(\lambda = \quad)$	$\int_0^{f_a^4} c_a(x)dx$	f_a^4 $(\lambda = \quad)$	$\int_0^{f_a^4} c_a(x)dx$
1,2								
2,4								
1,3								
3,4								
2,3								
	OF =		**OF =**		**OF =**		**OF =**	

Example of Braess's Paradox with Original Capacities from LeBlanc (1975) *TS*, 186-188.
Flows computed with the Origin-Based Traffic Assignment Algorithm of Bar-Gera (2002).

total flow: 1 to 4	**with link 2 - 3**					**without link 2 - 3**				
1	From	To	Volume	Link Cost	Total Cost	From	To	Volume	Link Cost	Total Cost
	1	2	1	40.5	40.5	1	2	0.5	40.0313	20.02
	1	3	0	185.0	0	1	3	0.5	185.0563	92.53
	2	4	0	185.0	0	2	4	0.5	185.0563	92.53
	2	3	1	16.4	16.4	2	3	0	∞	0
	3	4	1	40.5	40.5	3	4	0.5	40.0313	20.02
	total cost				97.4					225.09
	mean cost				97.4					225.09
2	From	To	Volume	Link Cost	Total Cost	From	To	Volume	Link Cost	Total Cost
	1	2	2	48.0	96.0	1	2	1	40.5	40.5
	1	3	0	185.0	0	1	3	1	185.9	185.9
	2	4	0	185.0	0	2	4	1	185.9	185.9
	2	3	2	31.4	62.8	2	3	0	∞	0
	3	4	2	48.0	96.0	3	4	1	40.5	40.5
	total cost				254.8					452.8
	mean cost				127.4					226.4
3	From	To	Volume	Link Cost	Total Cost	From	To	Volume	Link Cost	Total Cost
	1	2	3	80.5	241.5	1	2	1.5	42.5313	63.80
	1	3	0	185.0	0	1	3	1.5	189.5563	284.33
	2	4	0	185.0	0	2	4	1.5	189.5563	284.33
	2	3	3	96.4	289.2	2	3	0	∞	0
	3	4	3	80.5	241.5	3	4	1.5	42.5313	63.80
	total cost				772.2					696.26
	mean cost				257.4					232.09
4	From	To	Volume	Link Cost	Total Cost	From	To	Volume	Link Cost	Total Cost
	1	2	3.4056	107.2579	365.28	1	2	2	48.0	96.0
	1	3	0.5944	185.1123	110.03	1	3	2	199.4	398.8
	2	4	0.5944	185.1123	110.03	2	4	2	199.4	398.8
	2	3	2.8112	77.8545	218.86	2	3	0	∞	0
	3	4	3.4056	107.2579	365.28	3	4	2	48.0	96.0
	total cost				1169.48					989.6
	mean cost				292.37					247.4
5	From	To	Volume	Link Cost	Total Cost	From	To	Volume	Link Cost	Total Cost
	1	2	3.7236	136.1225	506.87	1	2	2.5	59.5313	148.83
	1	3	1.2764	187.3888	239.18	1	3	2.5	220.1563	550.39
	2	4	1.2764	187.3888	239.18	2	4	2.5	220.1563	550.39
	2	3	2.4472	51.2663	125.46	2	3	0	∞	0
	3	4	3.7236	136.1225	506.87	3	4	2.5	59.5313	148.83
	total cost				1617.56					1398.43
	mean cost				323.51					279.69

6	From	To	Volume	Link Cost	Total Cost	From	To	Volume	Link Cost	Total Cost
	1	2	4	168.0	672.0	1	2	3	80.5	241.5
	1	3	2	199.4	398.8	1	3	3	257.9	773.7
	2	4	2	199.4	398.8	2	4	3	257.9	773.7
	2	3	2	31.4	62.8	2	3	0	∞	0
	3	4	4	168..0	672.0	3	4	3	80.5	241.5
	total cost				2204.4					2030.4
	mean cost				367.4					338.4
7	From	To	Volume	Link Cost	Total Cost	From	To	Volume	Link Cost	Total Cost
	1	2	4.2998	210.9013	906.82	1	2	3.5	115.0313	402.61
	1	3	2.7002	232.8469	628.74	1	3	3.5	320.0562	1120.20
	2	4	2.7002	232.8469	628.74	2	4	3.5	320.0562	1120.20
	2	3	1.5995	21.9456	35.102	2	3	0	∞	0
	3	4	4.2998	210.9013	906.82	3	4	3.5	115.0313	402.61
	total cost				3106.24					3045.61
	mean cost				443.75					435.09
8	From	To	Volume	Link Cost	Total Cost	From	To	Volume	Link Cost	Total Cost
	1	2	4.6678	277.3730	1294.73	1	2	4	168.0	672.0
	1	3	3.3322	295.9557	986.17	1	3	4	415.4	1661.6
	2	4	3.3322	295.9557	986.17	2	4	4	415.4	1661.6
	2	3	1.3357	18.5830	24.82	2	3	0	∞	0
	3	4	4.6678	277.3730	1294.73	3	4	4	168.0	672.0
	total cost				4586.63					4667.2
	mean cost				573.33					583.4
9	From	To	Volume	Link Cost	Total Cost	From	To	Volume	Link Cost	Total Cost
	no results for flow 9					1	2	4.5	245.0313	1102.64
	the program shuts down automatically					1	3	4.5	554.0562	2493.25
						2	4	4.5	554.0562	2493.25
						2	3	0	∞	0
						3	4	4.5	245.0313	1102.64
	total cost									7191.79
	mean cost									799.09
10	From	To	Volume	Link Cost	Total Cost	From	To	Volume	Link Cost	Total Cost
	1	2	5.5599	517.8072	2878.98	1	2	5	352.5	1762.5
	1	3	4.4401	534.7801	2374.45	1	3	5	747.5	3737.5
	2	4	4.4401	534.7801	2374.45	2	4	5	747.5	3737.5
	2	3	1.1199	16.9729	19.01	2	3	0	∞	0
	3	4	5.5599	517.8072	2878.98	3	4	5	352.5	1762.5
	total cost				10525.87					11000.0
	mean cost				1052.59					1100.0
11	From	To	Volume	Link Cost	Total Cost	From	To	Volume	Link Cost	Total Cost
	1	2	6.0484	709.1750	4289.39	1	2	5.5	497.5313	2736.42
	1	3	4.9516	726.0225	3594.95	1	3	5.5	1008.5560	5547.06
	2	4	4.9516	726.0225	3594.95	2	4	5.5	1008.5560	5547.06
	2	3	1.0969	16.8474	18.48	2	3	0	∞	0
	3	4	6.0484	709.1750	4289.39	3	4	5.5	497.5313	2736.42
	total cost				15787.17					16566.96
	mean cost				1435.20					1506.09

12	From	To	Volume	Link Cost	Total Cost	From	To	Volume	Link Cost	Total Cost
	1	2	6.5516	961.2068	6297.43	1	2	6	688.0	4128.0
	1	3	5.4484	978.0879	5329.02	1	3	6	1351.4	8108.4
	2	4	5.4484	978.0879	5329.02	2	4	6	1351.4	8108.4
	2	3	1.1032	16.8811	18.62	2	3	0	∞	0
	3	4	6.5516	961.2068	6297.43	3	4	6	688.0	4128.0
	total cost			3895.47	23271.54				4078.8	24472.8
	mean cost				1939.30					2039.4

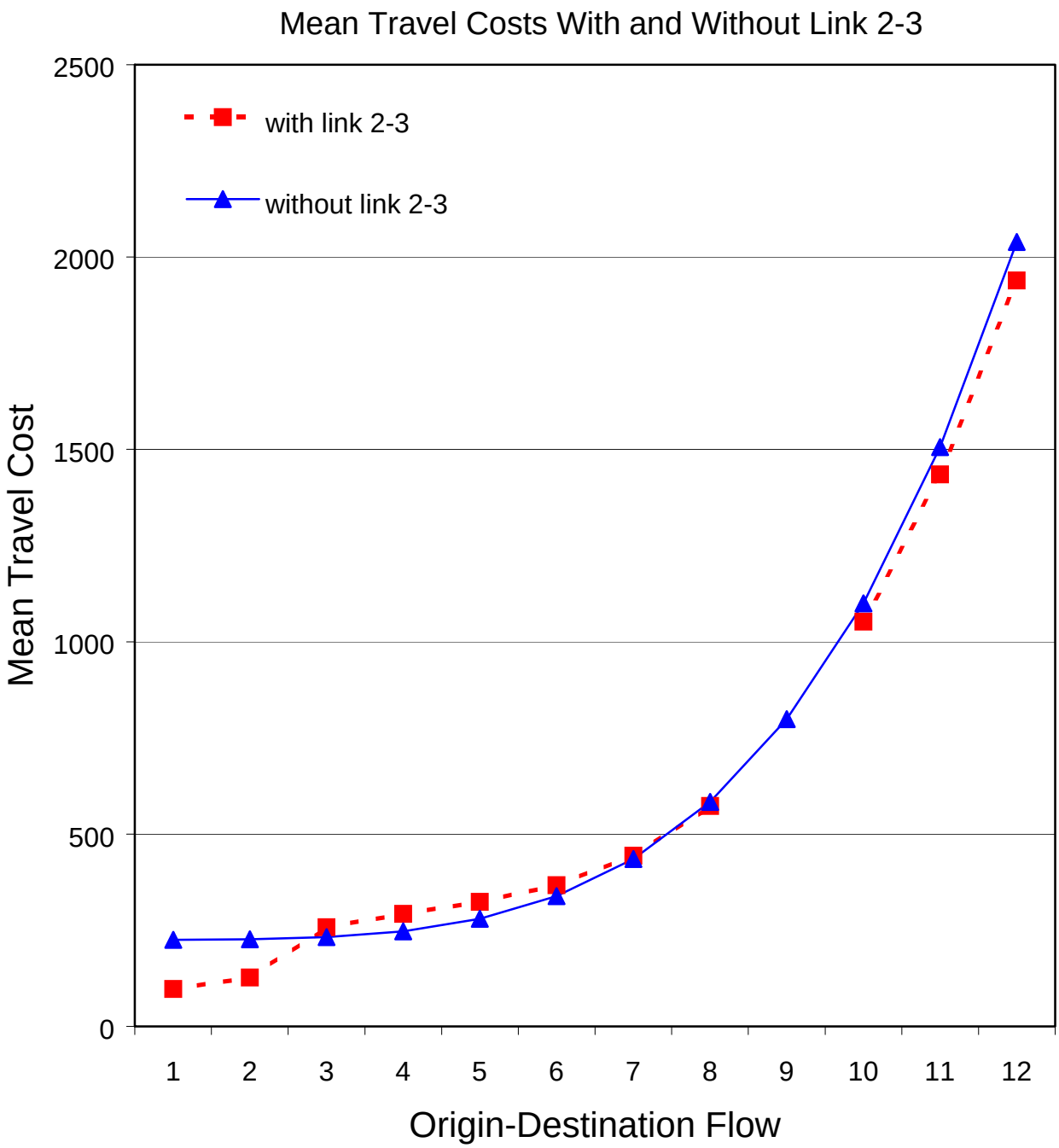
T12(f12) = 40. + 0.5(f12)**4

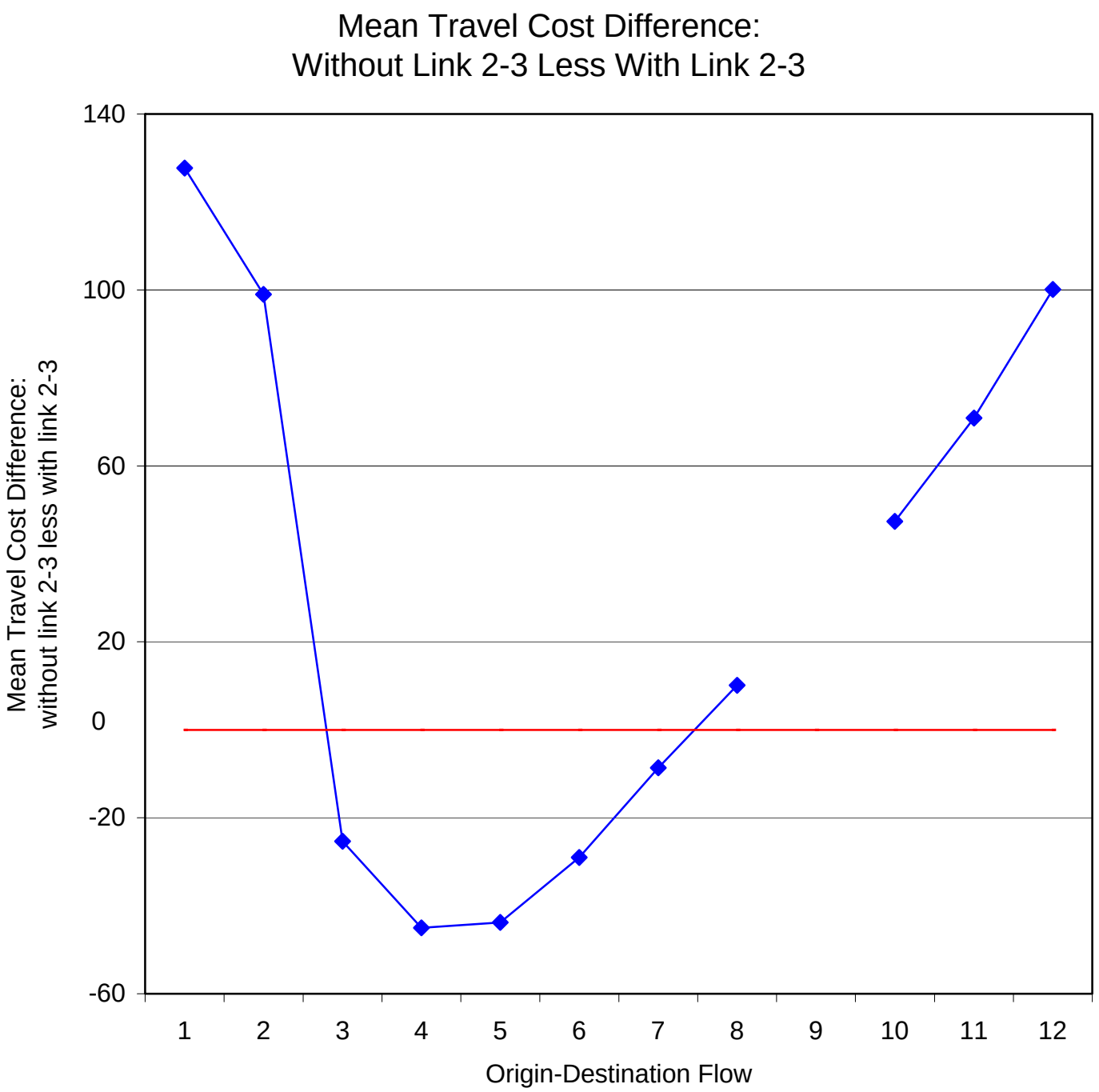
T13(f13) = 185. + 0.9(f13)**4

T24(f24) = 185. + 0.9(f24)**4

T34(f34) = 40. + 0.5(f34)**4

T23(f23) = 15.4 + f(23)**4




Mean Travel Cost Difference:
Without Link 2-3 Less With Link 2-3
140
100
60
20
0
-20
-60
Mean Travel Cost Difference:
without link 2-3 less with link 2-3
1
2
3
4
5
6
7
8
9
10
11
12
Origin-Destination Flow

3. Road and Transit Network Route Choice

3.1 Network Concepts and Shortest Route Algorithms

One of the first uses of computers for solving large-scale planning and engineering problems was for the *shortest route problem* for urban transportation networks of realistic size and detail. This capability enabled transportation planners to estimate origin-destination flows and link flows for alternative network proposals for the first time. Thus, the shortest route problem was used as a key portion of a heuristic for the traffic assignment problem. This application was first made in 1958 using an IBM 704 computer with 32,000 words of memory.

3.1.1 Network Concepts and Definitions

A transportation network as used in this problem is a mathematical or abstract representation of a system of roadways or transit lines. The treatment here has road systems in mind; these concepts can be extended to include transit systems.

node — a location on a network that is the intersection of two or more links

link — (or arc) a pair of nodes which are connected directly, and a number or function representing the generalized cost of travel between the two nodes; usually links are *directed*, or one-way, which means they represent the traffic moving from the tail node (a-node) to the head node (b-node).

network — a set of nodes and a set of links defined on the nodes

centroid — a node representing a zone, or small geographic area, through which flows enter and leave the network

route — sequence of directed links connecting an origin centroid with a destination centroid (also called a *path*)

tree — a network in which there is one and only one route from the node *R* to every other node, and no circuits; that is, it is not possible to return to *R* after leaving it; node *R* is called the origin or root of the tree.

shortest route tree — a tree in which every route is the shortest route from *R* to each node (also called a *minimum cost tree* or a *skim tree*).

3.1.2 Properties of Trees

1. The number of links in a tree is n-1, where n is the number of nodes.

2. The shortest route from *R* to a node *k* includes as subroutes the shortest routes from *R* to the intermediate nodes along the route. A *predecessor* or *back node* is the first node back towards the origin node *R* from a given node *i*.

3. The shortest route tree of origin *R* can be completely described by two arrays of length *n*:

    a. a predecessor node array, *P*(I), giving the predecessor node of I, where, *P*(R) = 0.
    b. a travel cost array, *C*(I), giving the minimum cost from *R* to I.

3.1.3 Shortest Route Problem

The shortest route (or path) problem for a centroid *R* is to find the shortest route tree rooted at origin *R*. The shortest route problem for *N* centroids is to find all *N* shortest route trees. This problem can also be solved for all nodes of the network.

There are two basic approaches to solving the shortest route problem:

1. tree-building algorithms which find the shortest route tree for a given origin node using a labeling technique;

2. matrix methods that find the shortest routes from every node to every other node in one well-defined sequence of matrix operations.

Although the matrix method concept of finding all shortest routes at once seems appealing, it is not practical for large networks because a matrix of size (*n* times *n*) can be very large; also, it finds the shortest routes between all pairs of nodes, not just between all centroids. For the purpose of traffic assignment only the latter is needed. Typically, the ratio of nodes to centroids is about 5:1.

Tree building algorithms consist of two basic types:

1. label setting methods: destination nodes are permanently labeled with their shortest route costs in order of increasing cost from the origin;

2. label correcting method: destinations are initially labeled with their route costs as they are encountered, and then revised if a shorter route is identified.

The disadvantage of type 1 is that the minimum value of an array of candidate link costs must be found repeatedly. The disadvantage of type 2 is that route costs must be updated when a shorter route is found. These types of algorithms are based on papers by Dijkstra and others, and Moore respectively. Many improvements have been found in these algorithms since they were proposed in the late 1950's. The best algorithm for a given application depends on the specific network characteristics and the computer on which it is solved. Nowadays, research on shortest path algorithms and computer codes lies more in the field of computer science than in the field of transportation science and engineering. Students transportation science should have a basic understanding of how such algorithms work; however, in my personal view, they should not consider the creation of new codes, or the refinement of existing ones, to be an appropriate research topic. Literature reviews on shortest route algorithms may be found in three papers listed at the end of Section 3.

3.1.4 The Dijkstra Algorithm

Assume that all link costs c_{ij} are nonnegative. Let C_j be the cost from the origin node to any node j.

1. Label the origin node as node 0 with $C_j = 0$.
2. Let 1, 2, 3, … k be labels at increasing cost from node 0; that is, $C_1 < C_2 < C_3 < ... < C_k$.

3. To label next node, given that *k-1* nodes are labeled,

   a. find node j that is connected to the tree of labeled nodes by *at least one* link from some labeled node i;
   b. form $(C_i + c_{ij})$ for each candidate defined in 3a, where c_{ij} is the cost of the link from node i to node j;
   c. find $c_j^* = \min_{ij}(C_i + c_{ij})$ and label the end node k; designate node i as the predecessor node of k; in case of ties, label one node k and the other k+1.

4. Repeat step (3) until all nodes have been labeled, or until all centroids are reached.

3.2 Auto Route Choice on a Network with Fixed OD Flows

Next, we generalize the route choice formulation for the two-link problem in Section 2 to a general network. To solve the problem, we will use the capability described above to find shortest routes through the network.

3.2.1 UO-TAP on a Network

$$\min_{(\mathbf{h})} z(\mathbf{h}) = \sum_a \int_0^{f_a} c_a(x)dx$$

$$\text{st}: \sum_{r \in R_{pq}} h_r = d_{pq}, p \in P; q \in Q$$

$$h_r \geq 0, \quad r \in R_{pq,}\ p \in P; q \in Q$$

$$\text{where } f_a \equiv \sum_{pq} \sum_{r \in R_{pq}} h_r \delta_r^a, a \in A$$

where $d_{pq} =$ fixed flow from zone p to zone q (vehicles/hour)

$h_r =$ flow on route r, a route from zone p to zone q that belongs to the set of routes R_{pq} connecting zone p to zone q

$\delta_r^a = 1$, if link a belongs to route r from zone p to zone q, and 0 otherwise

$c_a(f_a) =$ generalized travel cost function for link a, depending only on f_a, the flow on link a

$P, Q =$ sets of origin and destination zones, respectively

Note that the unknown variables are the route flows (h_r), and that the link flows (f_a) are defined in terms of the route flows. The link-route correspondence array (δ_r^a) is assumed to be known.

However, it is determined only as needed by use of the shortest route algorithm described above for problems of any size.

The optimality conditions for **UO-TAP** are determined as follows. Define the Lagrangean function:

$$L(\mathbf{h},\mathbf{u}) = \sum_a \int_0^{f_a} c_a(x)dx - \sum_{pq} u_{pq}\left(\sum_{r\in R_{pq}} h_r - d_{pq}\right)$$

Consider the derivative of the objective function with respect to the route flow h_r, the general unknown variable: $\frac{\partial z(\mathbf{h})}{\partial h_r} = \frac{\partial z(\mathbf{h})}{\partial f_a} \cdot \frac{\partial f_a}{\partial h_r} = \sum_a c_a(f_a) \cdot \delta_r^a$

The optimality conditions may then be stated as follows:

$$\sum_a c_a(f_a) \cdot \delta_r^a - u_{pq}(+1) \geq 0, r \in R_{pq}, \qquad p \in P, q \in Q$$

$$h_r \cdot \left(\sum_a c_a(f_a) \cdot \delta_r^a - u_{pq}(+1)\right) = 0, r \in R_{pq}, \qquad p \in P, q \in Q$$

$$\left(\sum_{r\in R_{pq}} h_r - d_{pq}\right) \geq 0, \qquad p \in P, q \in Q$$

$$u_{pq} \cdot \left(\sum_{r\in R_{pq}} h_r - d_{pq}\right) = 0, \qquad p \in P, q \in Q$$

$$\mathbf{h} \geq \mathbf{0}, \mathbf{u} \geq \mathbf{0}$$

These conditions may be interpreted in the following way for OD pair *pq*:

1. assume $h_r > 0$; then, $\left(\sum_a c_a(f_a) \cdot \delta_r^a - u_{pq}(+1)\right) = 0$, or $C_r \equiv \sum_a c_a(f_a) \cdot \delta_r^a = u_{pq}$;
2. assume $h_{r'} = 0$; then, $\left(\sum_a c_a(f_a) \cdot \delta_{r'}^a - u_{pq}(+1)\right) \geq 0$, or $C_{r'} \geq u_{pq}$;
3. assume $C_{r''} > u_{pq}$; then $h_{r''} = 0$.

Hence, every used route connecting zone *p* to zone *q* has the equal generalized travel cost u_{pq}, and no unused route has a lower travel cost. Therefore, Wardrop's first principle is satisfied for this formulation. A similar formulation may be formulated for the System-Optimal problem by replacing $c_a(f_a)$ by the marginal cost function $m_a(f_a)$.

If the generalized cost functions are strictly increasing with link flow, then the solution is unique in the link flows; however, it is not unique in the route flows. The reason is that the link flows are a linear function of the route flows; therefore, the objective function is not *strictly* convex in the route flows.

3.2.2 Nonlinear Complementarity and Variational Inequality Formulations of UO-TAP

Following the procedure described in Appendix A.5, we can formulate the NCP corresponding to the above problem as follows:

$$\mathbf{F}(\mathbf{y}) = \begin{bmatrix} \sum_a c_a(f_a)\cdot\delta_r^a - u_{pq} & r \in R_{pq}, p \in P, q \in Q \\ \sum_{r\in R_{pq}} h_r - d_{pq} & p \in P, q \in Q \end{bmatrix} \geq \mathbf{0} \qquad \mathbf{y} = \begin{bmatrix} h_r & r \in R_{pq}, p \in P, q \in Q \\ u_{pq} & p \in P, q \in Q \end{bmatrix} \geq \mathbf{0}$$

$$\begin{aligned}
\mathbf{F}(\mathbf{y})^{\mathrm{T}} \cdot \mathbf{y} &= \begin{bmatrix} \sum_a c_a(f_a)\cdot\delta_r^a - u_{pq} & r \in R_{pq}, p \in P, q \in Q \\ \sum_{r\in R_{pq}} h_r - d_{pq} & p \in P, q \in Q \end{bmatrix}^{\mathrm{T}} \cdot \begin{bmatrix} h_r & r \in R_{pq}, p \in P, q \in Q \\ u_{pq} & p \in P, q \in Q \end{bmatrix} = 0 \\
&= \left[\sum_{p\in P, q\in Q} \sum_{r\in R_{pq}} h_r \cdot \left(\sum_a c_a(f_a)\cdot\delta_r^a - u_{pq} \right) + \sum_{p\in P, q\in Q} u_{pq} \cdot \left(\sum_{r\in R_{pq}} h_r - d_{pq} \right) \right] \\
&= \left[\sum_{p\in P, q\in Q} \sum_{r\in R_{pq}} h_r \cdot C_r(\mathbf{f}) - \sum_{p\in P, q\in Q} u_{pq} \cdot \sum_{r\in R_{pq}} h_r + \sum_{p\in P, q\in Q} u_{pq} \cdot \sum_{r\in R_{pq}} h_r - \sum_{p\in P, q\in Q} d_{pq} \cdot u_{pq} \right] \\
&= \left[\sum_{p\in P, q\in Q} \sum_{r\in R_{pq}} h_r \cdot C_r(\mathbf{f}) - \sum_{p\in P, q\in Q} d_{pq} \cdot u_{pq} \right] = 0, \text{ where } C_r(\mathbf{f}) \equiv \sum_a c_a(f_a)\cdot\delta_r^a .
\end{aligned}$$

This requirement states at equilibrium that the total travel costs determined in terms of route costs and route flows must equal the total travel costs stated in terms of the fixed OD flows and the user-optimal OD costs. Note that the asterisks denoting that all flows are equilibrium are omitted in order to simplify the expression. The equivalence of total route-based costs and total OD-based costs is a basic property of the equilibrium solution, which holds for each OD pair.

The equivalent VIP may be stated as follows:

$$\begin{bmatrix} \sum_a c_a(f_a^*)\cdot\delta_r^a - u_{pq}^* & r \in R_{pq}, p \in P, q \in Q \\ \sum_{r\in R_{pq}} h_r^* - d_{pq} & p \in P, q \in Q \end{bmatrix}^{\mathrm{T}} \cdot \begin{bmatrix} (h_r - h_r^*) & r \in R_{pq}, p \in P, q \in Q \\ (u_{pq} - u_{pq}^*) & p \in P, q \in Q \end{bmatrix} \geq 0, \quad \begin{matrix} \mathbf{h} \geq \mathbf{0} \\ \mathbf{u} \geq \mathbf{0} \end{matrix}.$$

By moving the conservation flow constraints into the definition of the constraint space, we may simplify the formulation as follows:

$$\left[\sum_a c_a(f_a^*)\cdot\delta_r^a \quad r \in R_{pq}, p \in P, q \in Q \right]^{\mathrm{T}} \cdot \left[(h_r - h_r^*) \quad r \in R_{pq}, p \in P, q \in Q \right] \geq 0, \text{ or}$$

$$\sum_{pq} \sum_{r\in R_{pq}} C_r(\mathbf{f}^*)\cdot(h_r - h_r^*) \geq 0, \quad h_r \in \mathrm{H} = \{ \sum_{r\in R_{pq}} h_r^* - d_{pq},\ p \in P, q \in Q; \mathrm{R}_+^n \}, \text{ which is equivalent to:}$$

$$\sum_a c_a(f_a^*)\cdot(f_a - f_a^*) \geq 0, \quad f_a \in \mathrm{F} = \{ \sum_{r\in R_{pq}} h_r - d_{pq},\ p \in P, q \in Q;\ f_a \equiv \sum_{pq} \sum_{r\in R_{pq}} h_r \delta_r^a, a \in A; \mathbf{h} \geq \mathbf{0} \}$$

3.3 Solution Algorithm for UO-TAP based on the Frank-Wolfe Method

A solution algorithm for UO-TAP for the case of a general network with *separable* link costs $(c_a \equiv c_a(f_a))$ may be readily defined as a generalization of the algorithm for the two-link problem described in Section 2. This algorithm is generally known as the Frank-Wolfe algorithm or method; it is known to converge relatively slowly after the first few iterations. The algorithm may be stated as follows:

1. initialize $\mathbf{f}^n$; set a counter to n = 1; a typical method is to perform an all-or-nothing assignment to the zero-flow link generalized costs; see the following subsection for a definition of all-or-nothing assignment;

2. solve the Subproblem for $\mathbf{g}^n$, the Subproblem link flows, as described below;

3. check for convergence of the Relative Gap at iteration n: Is $\mathrm{RG}^n \equiv \left| \dfrac{z(\mathbf{f}^n) - BLB^n}{BLB^n} \right| < \varepsilon$?

   where $BLB^n \equiv \max\limits_{n' \le n} \left(z(\mathbf{f}^{n'}) + \nabla z(\mathbf{f}^{n'}) \cdot (\mathbf{g}^{n'} - \mathbf{f}^{n'}) \right) = \max\limits_{n' \le n} \left(\left(z(\mathbf{f}^{n'}) + \mathrm{Gap}^{n'} \right) \right)$

   $\mathrm{Gap}^n \equiv \nabla z(\mathbf{f}^n) \cdot (\mathbf{g}^n - \mathbf{f}^n)$

4. perform a line search, as described below, to obtain λ^n and set $\mathbf{f}^{n+1} = (1 - \lambda^n) \cdot \mathbf{f}^n + \lambda^n \cdot \mathbf{g}^n$

5. check for convergence again with the new value of the objective function $z(\mathbf{f}^{n+1})$; if not converged, increment n to n+1, and go to step 2.

<u>Solution of the Subproblem (Step 2)</u>

The Subproblem is found by approximating the objective function at solution $\bar{\mathbf{f}}^n$, as follows:

$$z(\mathbf{g}^n) \approx \sum_a \int_0^{\bar{f}_a^n} c_a(x)dx + \sum_a c_a(\bar{f}_a^n) \cdot (g_a^n - \bar{f}_a^n) = \sum_a c_a(\bar{f}_a^n) \cdot g_a^n + \text{constant}$$

To solve the Subproblem subject to the conservation of flow constraints, we formulate the following problem, where $\hat{\mathbf{h}}$ is now the subproblem route flow:

$$\min_{(\mathbf{g})} \sum_a c_a(\bar{f}_a^n) \cdot g_a^n$$

$$\text{st}: \sum_{r \in R_{pq}} \hat{h}_r = d_{pq}, p \in P; q \in Q$$

$$\hat{\mathbf{h}} \ge \mathbf{0}$$

$$\text{where } g_a \equiv \sum_{pq} \sum_{r \in R_{pq}} \hat{h}_r \delta_r^a, a \in A$$

Since $c_a\left(\bar{f}_a^n\right)$ is temporarily fixed at $\bar{\mathbf{f}}^n$, this is a linear programming problem in $\mathbf{g}^n$. The above problem in the space of link flows may be transformed to the space of route flows, as follows:

$$\sum_a c_a\left(\bar{f}_a^n\right)\cdot g_a^n = \sum_a c_a\left(\bar{f}_a^n\right)\cdot \sum_{pq}\sum_{r\in R_{pq}} \hat{h}_r \delta_r^a = \sum_{pq}\sum_{r\in R_{pq}} \hat{h}_r \cdot \sum_a c_a\left(\bar{f}_a^n\right)\cdot \delta_r^a = \sum_{pq}\sum_{r\in R_{pq}} C_r\left(\bar{\mathbf{f}}^n\right)\cdot \hat{h}_r$$

where $C_r\left(\bar{\mathbf{f}}^n\right) \equiv \sum_a c_a\left(\bar{f}_a^n\right)\cdot \delta_r^a$ is the travel cost on route *r* with flows $\bar{\mathbf{f}}^n$. This transformation is carried out by rearranging the terms in the above equation and refactoring, may be shown by writing out the equation in detail.

Since $C_r\left(\bar{\mathbf{f}}^n\right)$ is a constant, the problem, $\min_{\left(\hat{\mathbf{h}}\geq \mathbf{0}\right)} \sum_{pq}\sum_{r\in R_{pq}} C_r\left(\bar{\mathbf{f}}^n\right)\cdot \hat{h}_r$

may be solved by identifying the minimum cost route from *p* to *q*, and placing all of the OD flow d_{pq} on that route. Such a solution is traditionally called an *all-or-nothing solution*. A more meaningful name would be *minimum cost route* solution. The solution is readily found by applying the shortest route algorithm to the network with link costs $\left(c_a\left(\bar{f}_a^n\right)\right)$ and placing the entire OD flow on each minimum cost route. Summing these route flows to link flows yields the Subproblem solution $\mathbf{g}^n$.

<u>Solution of the Line Search (Step 4)</u>

The line search is defined as: $\min_{\left(0\leq \lambda^n \leq 1\right)} z(\lambda) = \sum_a \int_0^{f_a^{n+1}} c_a(x)dx$,

where $f_a^{n+1} = f_a^n + \lambda^n \cdot \left(g_a^n - f_a^n\right) = \left(1-\lambda^n\right)\cdot f_a^n + \lambda^n \cdot g_a^n$

This problem consists of the minimization of a nonlinear function with respect to a scalar. Such problems may be readily solved by applying a line search technique, such as bisection search. To apply this technique, we require the partial derivative of the objective function with respect to the unknown variable λ^n, which may be stated as follows:

$$\frac{\partial z(\lambda)}{\partial \lambda} = \sum_a c_a\left(f_a^{n+1}\right)\cdot\left(g_a^n - f_a^n\right)$$

Using this result, the bisection search proceeds as follows:

1. initialize $\lambda_1 = 0.5$; set k = 1, and $\lambda_0 = 1.0$.

2. find $z'(\lambda_k) = \sum_a c_a \left(f_a^{n+1}\right) \cdot \left(g_a^n - f_a^n\right)$ = slope of the function $z(\lambda)$.

   a. if $z'(\lambda_k) > 0, \quad \lambda_{k+1} = \lambda_k - (\lambda_k - \lambda_{k-1})/2$

   b. if $z'(\lambda_k) < 0, \quad \lambda_{k+1} = \lambda_k + (\lambda_k - \lambda_{k-1})/2$

3. stop when $(\lambda_k - \lambda_{k-1}) < \varepsilon$, and $z'(\lambda_k) < 0$ at least once.

The last requirement assures that the chosen value of λ results in a decrease in the Main Problem objective function. If this requirement were not fulfilled, the chosen value of λ would lie to the right of its minimum value, and could correspond to a higher value of the Main Problem objective function than the present value. Draw the curve of the Main Problem objective function vs. λ to see this point. In computational experiments for a large network, it was shown that ε can be as large as 0.25 without increasing the overall solution time significantly. Selecting a value as large as 0.25 may substantially reduce the computational effort of the line search.

3.4 Origin-Based Traffic Assignment Algorithm

3.4.1 Overview

The algorithm described in this section was formulated, implemented and extensively tested on large networks by Hillel Bar-Gera (1999, 2002), starting with his Ph.D. thesis, and continuing since that time. This section is intended to give a short, relatively nonmathematical overview of the algorithm. The Frank-Wolfe algorithm may be characterized as being link-based, since the solution variables are link flows. Another type of algorithms is route-based, in that each route flow is a solution variable. Route-based algorithms were proposed by Bothner and Lutter (1982) and Larson and Patriksson (1992); the former is the basis for one practitioner software system.

The concept of the Origin-Based Assignment (OBA) algorithm is to define the solution variables in an intermediate way between links and routes. In this way the algorithm seeks to solve larger networks efficiently than can be solved by either of the two approaches. For examples of comparisons of solutions, see Bar-Gera (2002).

The main solution variables in this algorithm are *origin-based approach proportions*, α_{pa}, for every origin *p* and every link *a*, such that for every origin *p* and node *i* the sum of origin-based approach proportions over all links ending at node *i* is equal to one. Using origin-based approach proportions, *route proportions* are determined by the product of approach proportions of all the links along the route, that is $\gamma_{pqr} = \prod_{a \subseteq r} \alpha_{pa}$. Route flows are determined by the product of OD flow and route proportion, that is $h_{pqr} = d_{pq} \cdot \gamma_{pqr}$. It can be shown (Bar-Gera, 2002, eq. 14) that if link *a* goes from node *i* to node *j*, and if the total flow from origin *p* to node *j* is g_{pj}, then the total flow from origin *p* that arrives at node *j* through link *a* is $f_{pa} = \alpha_{pa} \cdot g_{pj}$; in that respect α_{pa} is indeed the proportion of flow on approach *a* to node *j* for origin *p*, as implied from the name of these variables.

The representation of the solution by origin-based approach proportions allows storing a complete description of the route flows very efficiently. The efficiency of the representation is further enhanced using the fact that at most nodes one link receives approach proportion value of one, while the value of all other links ending at the same node is zero. The availability of route flows can be useful for solution analysis. It is also useful in searching for the equilibrium solution, which is a major difference from many alternative solution procedures, including the Frank-Wolfe algorithm, which store only total link flows during the iterative process.

A key point in this algorithm is that for every origin an *a-cyclic* restricting subnetwork is chosen, A_p, such that for origin p approach proportions of links that are not included in A_p are restricted to 0. Using the equation for route proportions it can be seen that under these restrictions, for every origin, only routes that are limited to the links in its restricted subnetwork can be used. In particular, since A_p is a-cyclic, meaning that it does not contain a directed cycle of links, any cyclic route must contain at least one link that does not belong to A_p, and hence the flow along any cyclic route must be zero. It is important to note that the restriction to a-cyclic subnetworks does exclude many solutions that do not use cyclic routes, which are usually considered legitimate. It can be shown (for example, Bar-Gera, 2002; Lemma 3) that there is always an equilibrium solution that is a-cyclic by origin. Therefore, this restriction does not prevent the algorithm from converging to the true equilibrium solution.

The restriction to solutions that are a-cyclic by origin has several important advantages. First, the simple route flow interpretation presented above is, in fact, only valid for solutions that are a-cyclic by origin. Second, a-cyclic subnetworks allow a definition of a topological order of the nodes, which is an origin-specific ordering of the nodes, such that every link in the restricting subnetwork goes from a node of lower topological order to a node of higher topological order. Most computations in the proposed algorithm are done in a single pass over the nodes, either in ascending or descending topological order. The time required by such computations is a linear function of the number of links in the network. This computational efficiency is the main reason for restricting to a-cyclic solutions.

In solving the traffic assignment problem, the algorithm starts with trees of minimum cost routes as restricting subnetworks, leading to an all-or-nothing assignment. Then, the algorithm considers all origins in a sequential order. For each origin the restricting subnetwork is updated, and the origin-based approach proportions are adjusted within the given restricting subnetwork. To update a restricting subnetwork, unused links are removed; v_i the maximum cost from the origin to node i within the restricting subnetwork is computed for all nodes, and all links [i,j] such that $v_i < v_j$ are added to the restricting subnetwork. Once a new restricting subnetwork is found, several computationally intensive steps are needed including reorganization of the data structure. However, restricting subnetworks tend to stabilize fairly quickly. Therefore, it is useful to update origin-based approach proportions while keeping the restricting subnetworks fixed. This is done by introducing *inner iterations* as described below in the flow chart.

To update origin-based approach proportions within a given restricting subnetwork, a search direction based on shifting flow from high cost alternatives to low cost alternatives is used. In addition to current costs, estimates of cost derivatives are used to improve the search direction in a quasi-Newton fashion. When two independent routes are considered, the amount of flow

shifted by the search direction equals the difference between route costs divided by the sum of route cost derivatives, which is exactly what would be obtained by considering objective function second-order approximation.

The second-order search direction described above is viewed as desirable flow shifts. These are scaled by a step size between zero and one, and then truncated to guarantee feasibility. We refer to this technique as the boundary search procedure, as it tends to choose solutions along the boundary, although it does consider interior points as well. This technique is somewhat different then conventional line search techniques, where shifts are first truncated to guarantee feasibility and only then scaled by a step size. The importance of the boundary search for origin-based assignment is that it is effective in eliminating residual flows, i.e. small flows on sub-optimal routes. The elimination of residual flows is critical for algorithm convergence. See Bar-Gera (2002) for details.

In order to guarantee descent of the objective function, and convergence of the algorithm, the search considers step size values of 1, 1/2, 1/4, 1/8 etc. The stopping condition is based on the concept of *social pressure* introduced by Kupsizewska and Van Vliet (1999). The basic idea is that every traveler shifted from route r_1 to r_2 puts pressure (positive or negative), which is equal to his/her gain (or loss) as a result from the shift, that is according to the difference in route costs. The total social pressure is the sum of the pressure from all the travelers. Our search direction is good in the sense that it always enjoys positive social pressure for small step sizes. As the step size increases, the social pressure decreases, and eventually it may become negative. Our goal is to find the largest step size, i.e. the first in the sequence 1, 1/2, 1/4, 1/8… with positive social pressure. This social pressure principle is in fact equivalent to the stopping condition of the line-search in the Frank-Wolfe algorithm, only that this principle is applicable in certain cases where the line-search optimization rule does not.

Numerous large-scale applications have been demonstrated that the Origin-Based algorithm is capable of achieving highly accurate solutions, limited only by the accuracy of the computer's arithmetic, and is much faster than the Frank-Wolfe algorithm. The algorithm consists of two main steps: update restricting subnetwork, and update origin-based link flows. The algorithm is described by the following pseudocode:

<u>Initialization</u>:

```
for every origin p
    Let A_p be a tree of minimum cost routes under free flow conditions from p
    Let α_pa equal 1 for all links in A_p and 0 otherwise. (all-or-nothing assignment)
end for
```

<u>Main loop</u>:

```
for n=1 to number of main iterations
    for every origin p
        update restricting subnetwork A_p
        update origin-based approach proportions α_pa
    end for
    for m=1 to number of inner iterations
```

```
        for every origin p
            update origin-based approach proportions α_pa
        end for
    end for
end for
```

<u>Update restricting subnetwork for origin *p*:</u>

```
remove unused links from A_p
for every node i compute the maximum cost v_i from p to i
for every link a=[i,j]
    if v_i < v_j add link a to A_p
find new topological order for new A_p
update data structures
```

<u>Update origin-based approach proportions for origin *p*:</u>

```
compute average costs and Hessian approximations
    for step size 1, 1/2, 1/4, 1/8…
        compute flow shifts and scale by step size
        project and aggregate flow shifts
        if social pressure is positive then stop
    end for
apply flow shifts
update total link flows and link costs
```

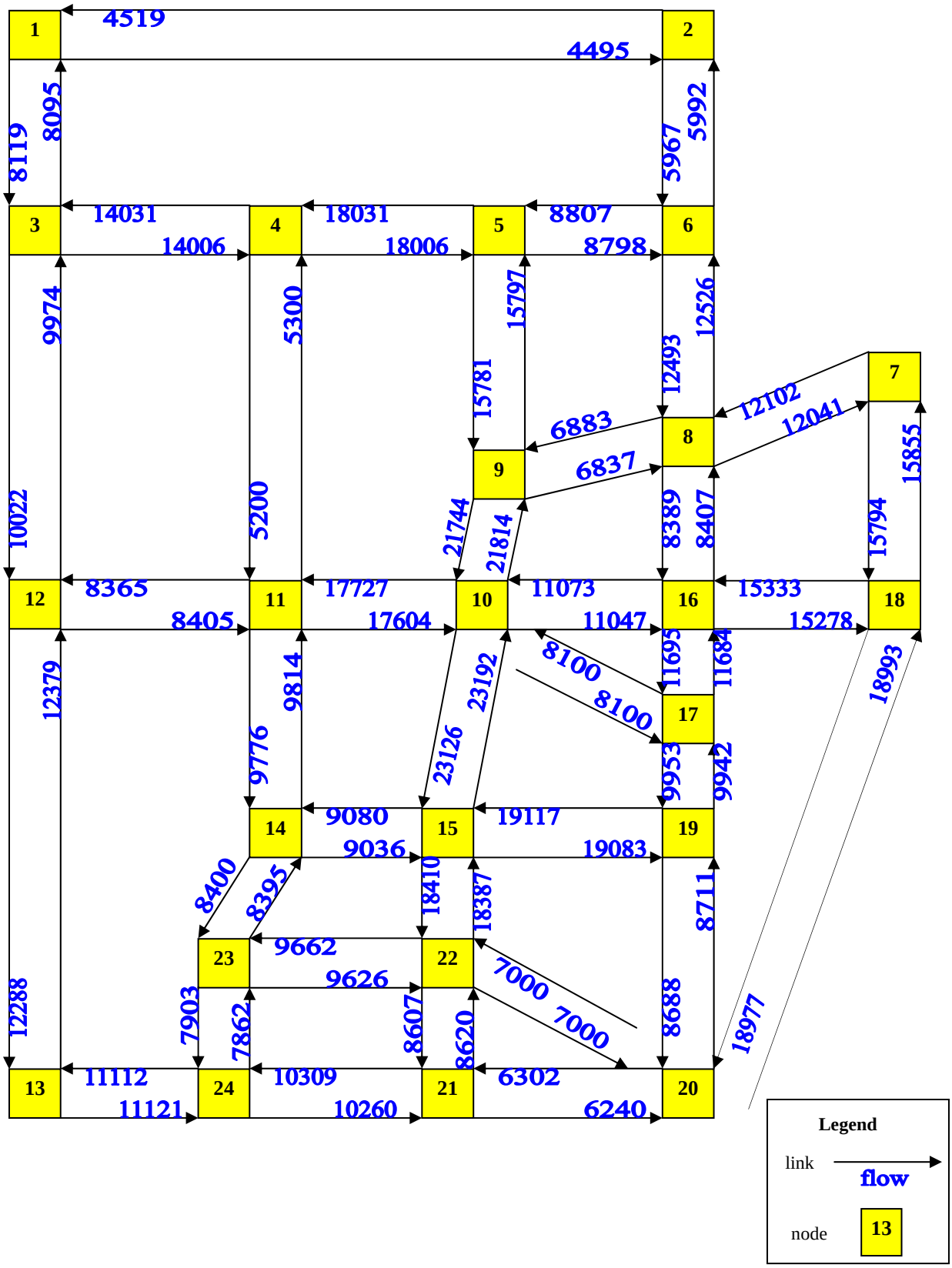
3.5.2 Example of the Application of the Origin-based Assignment Algorithm

LeBlanc (1975) published a test network roughly based on the small U.S. city of Sioux Falls, South Dakota. Because other test networks were not available or acquired, this small test network has been used quite often. This shortcoming of our field has now been remedied by the creation of a test network web site in 2001 by Bar-Gera at http://www.bgu.ac.il/~bargera/tntp/. In these notes, the Sioux Falls test network of LeBlanc is used to illustrate the solution with the Origin-Based Assignment algorithm. The Sioux Falls network has 24 zones and 76 one-way links. It is an atypical network in that each node of the network is also a zone centroid. A standard trip table is used to solve for the user-optimal flows for this network.

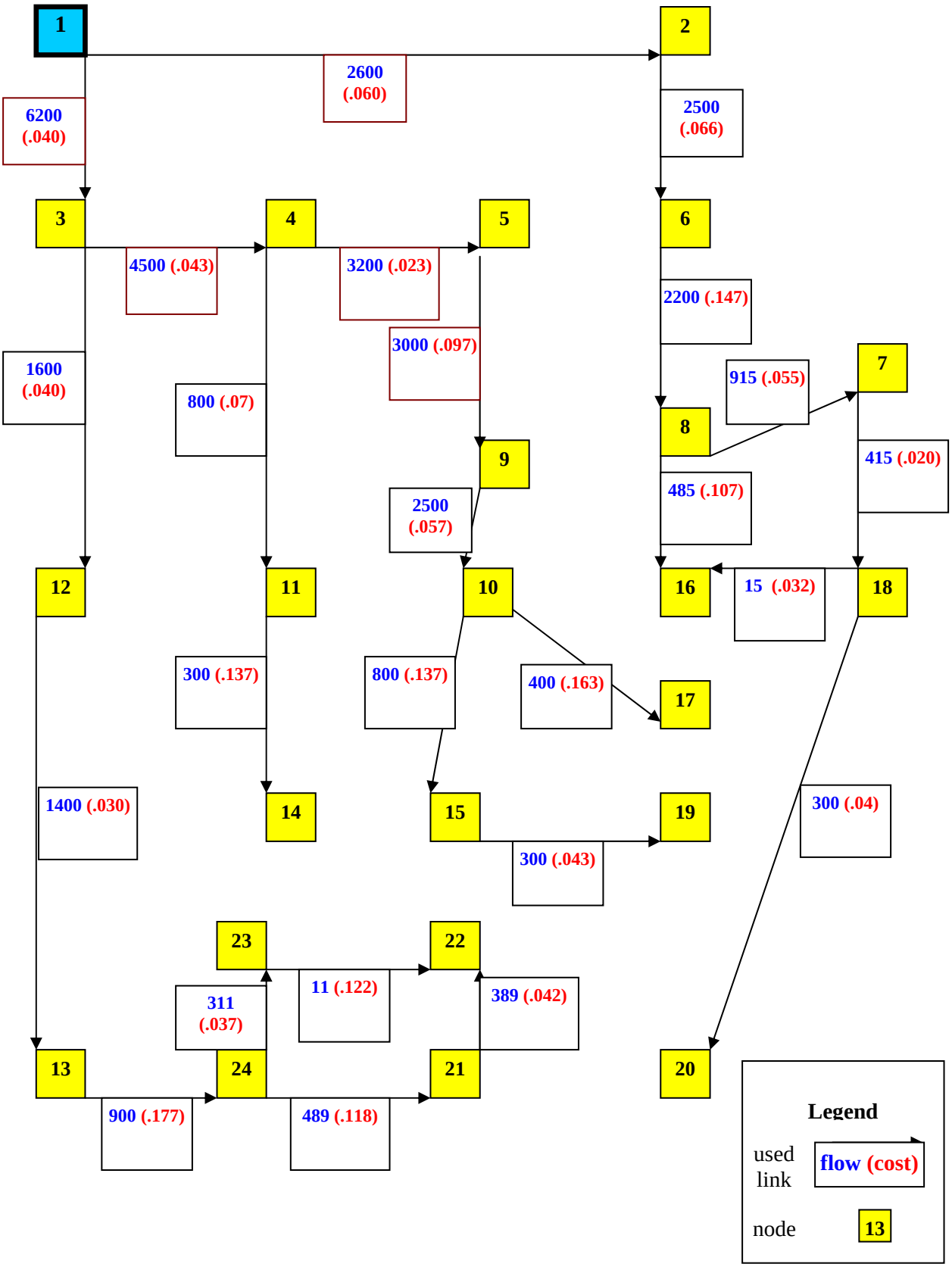
The following five charts show the results of OBA applied to this network. Chart 1 shows the total link flows for the user-optimal solution; Charts 2 and 3 show the origin-based subnetworks for nodes 1 and 12. Chart 4 shows the origin-based flow proportions for node 12, illustrating how to compute the route flow for OD pair 12-16. Chart 5 shows all routes from node 12 to node 16.

The following figure shows the Gap for the Frank-Wolfe and Origin-based algorithms, but does not provide much information concerning where to stop the convergence process. For this important decision, the judgment of the professional practitioner is essential. One idea is the algorithm should be terminated when the link flows become stable among the solutions for plan alternatives. To illustrate this point, we next consider a case study for the Philadelphia network.

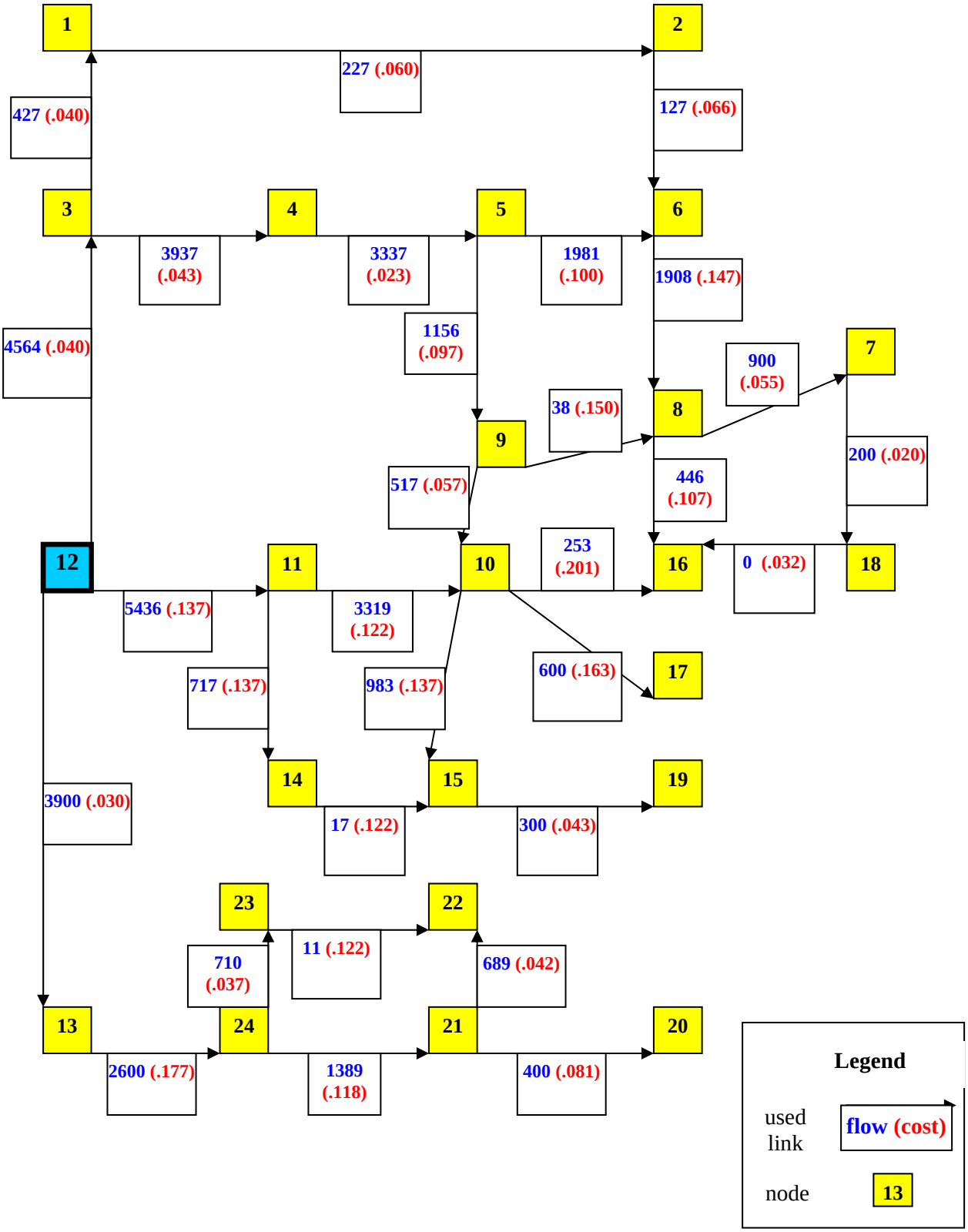
1. Sioux Falls Network - Total Link Flows



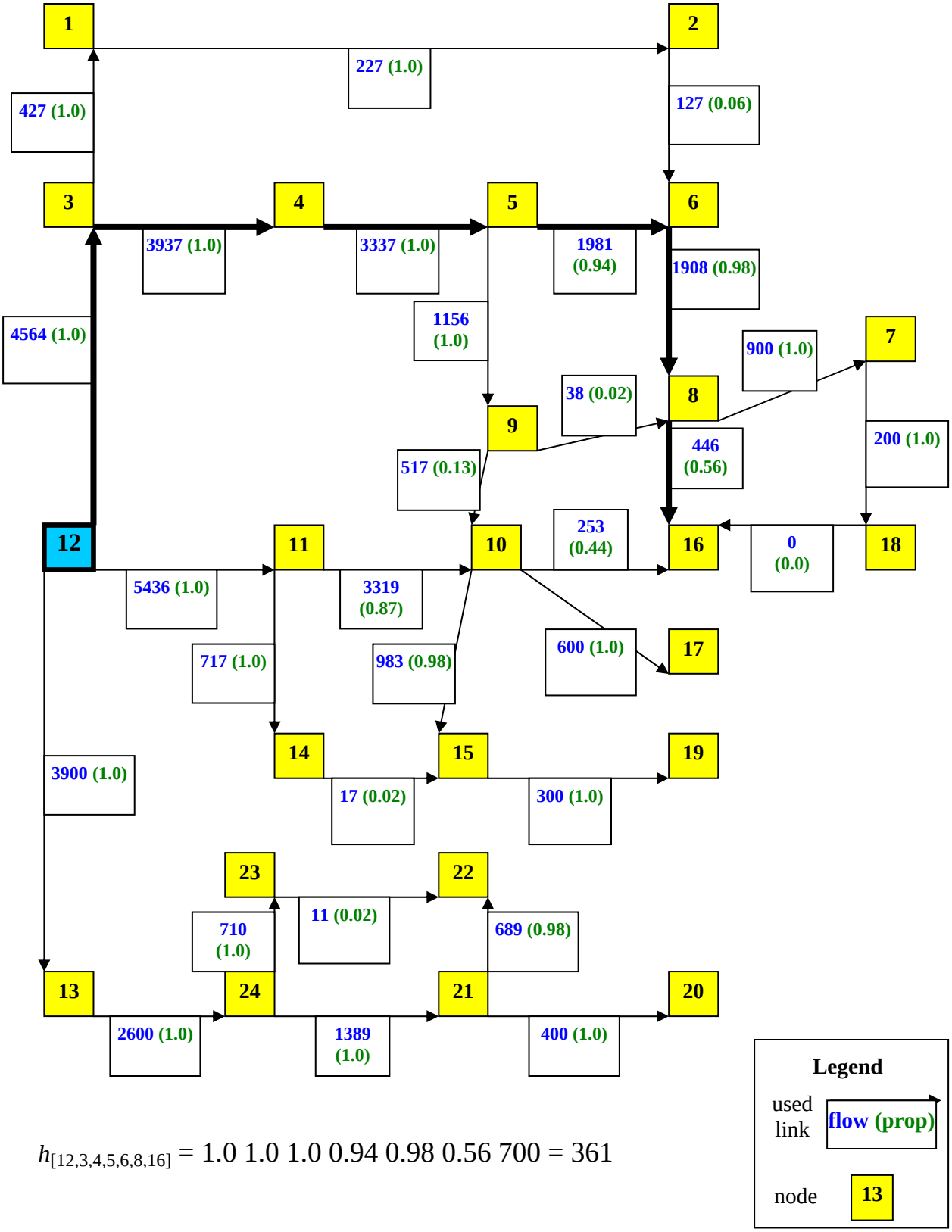
2. Origin-based solution - Node 1



3. Origin-based Solution - Node 12

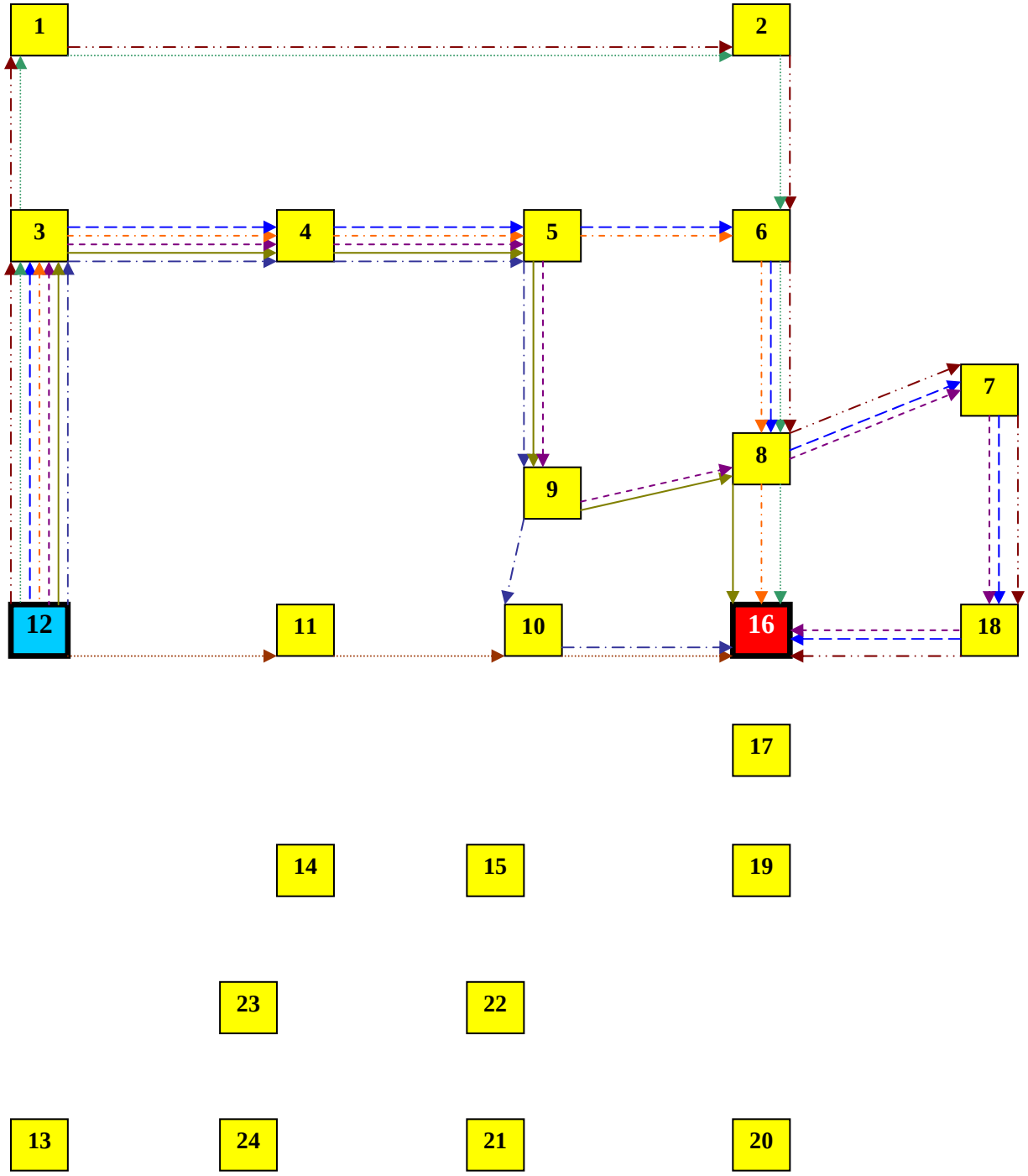


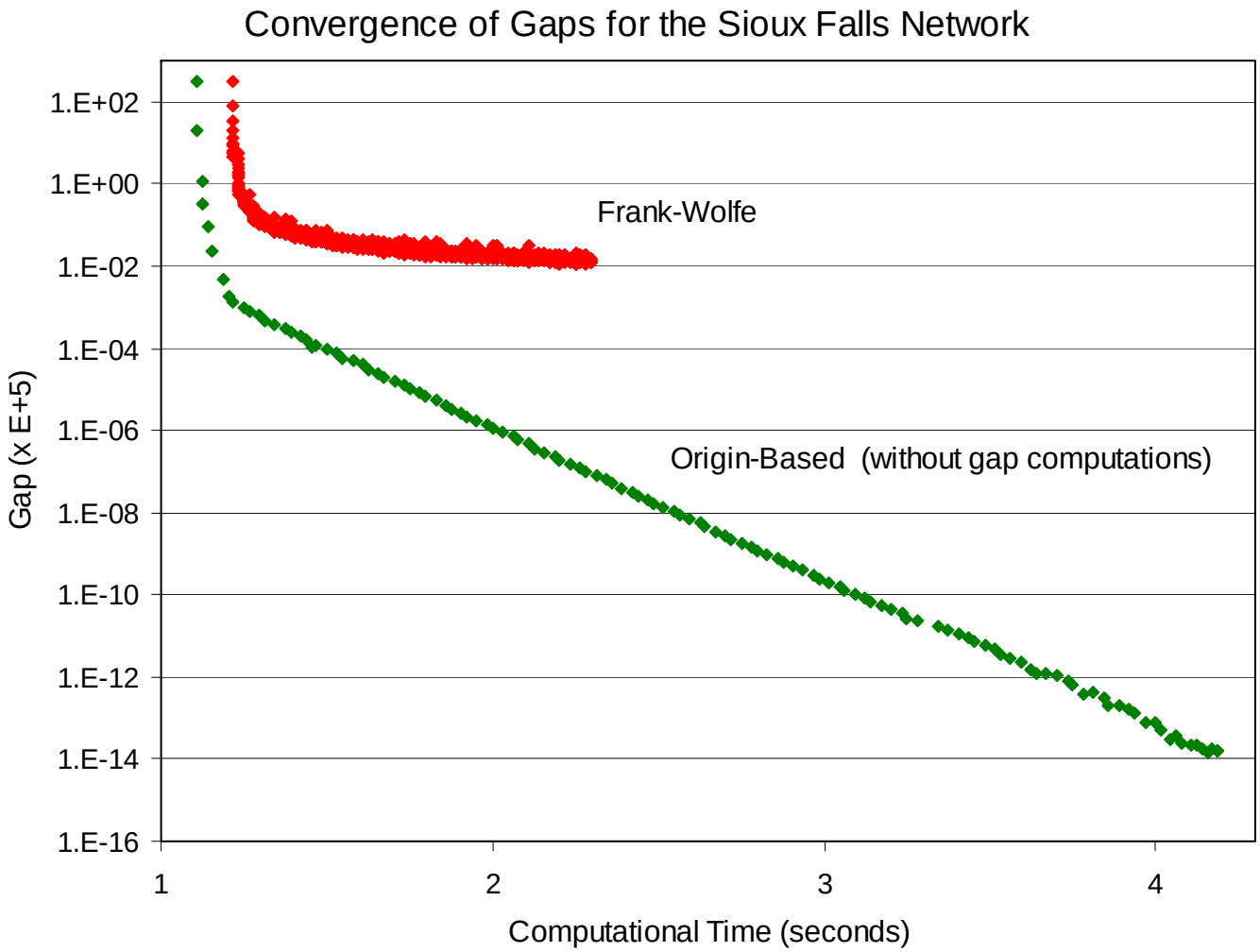
4. Route Flow Interpretation – Node 12



$h_{[12,3,4,5,6,8,16]}$ = 1.0 1.0 1.0 0.94 0.98 0.56 700 = 361

5. Routes - Nodes 12 to 16





3.5 Case Study of Convergence of UO TAP

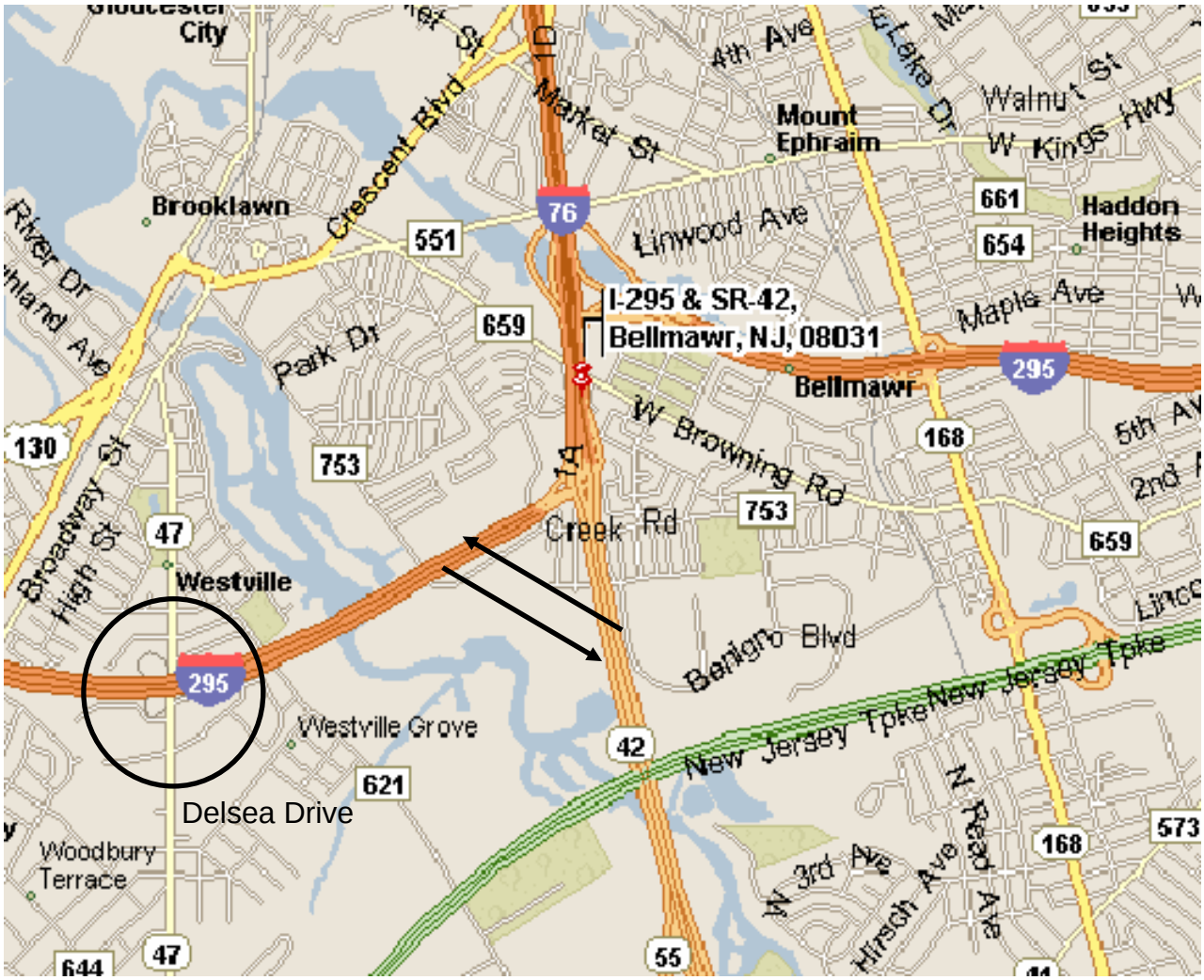
Several years ago, the planning agency for the Delaware Valley region in southeastern Pennsylvania – southern New Jersey was asked to evaluate the benefits of constructing two ramps at an interchange of two major freeways. Their analysis showed that link flows fluctuated all over the region as a result of this relatively minor change in the network. Inspection of the results suggested that insufficient convergence of the traffic assignments was the reason.

The Origin-Based Assignment algorithm was applied to these data to investigate the effect of convergence on the results. Our studies showed that achieving stability of flows in the interchange area requires better convergence than was possible with the software used in the original study or with most software used by practitioners. The findings are summarized in Boyce et al (2004). The following maps and eight figures are extracted from that paper.

Fig. 1. Study Area Location in the Delaware Valley Region

In the Figures 3-4, we see that the link flows in the vicinity of the ramps, as well as the ramps themselves, fluctuate substantially as the convergence of the assignment improves from left to right. In Figures 5-6, we can observe the relative error in these link flows, as compared with a highly converged solution. Finally, Figures 7-8 show the fluctuations on a major arterial affected by the introduction of the ramps.

Fig. 2. Interchange of I-295 with SR-42 Showing Proposed Ramps



Next, we investigated the effect of convergence of the traffic assignments on 8,000 links nearest to the ramps. These results are shown in the next pair of maps. The links are colored in each map according to the difference in link flow between the build and no-build networks, using an order of magnitude scale. For the maps on the next page for Relative Gaps of 1% and 0.0001%, note how the colors change on the west side of the Delaware River, as well as in the vicinity of the ramps, which are shown in the center of the map. A relative gap of 1% may be regarded as best current practice; a relative gap of 0.0001% is well beyond the level of convergence needed for professional practice, but is shown here to illustrate that a precise level of convergence is possible.

The next pair of maps shows the errors in the link flow differences, as compared with a very highly converged solution. Note the change in the scale of the color codes in this second pair of maps. For a relative gap of 0.0001%, nearly all link flow differences lie in the interval from +10 to -10.

Fig. 3. Link flow differences (build less no-build) versus relative gap
New Jersey freeway links with positive flow differences

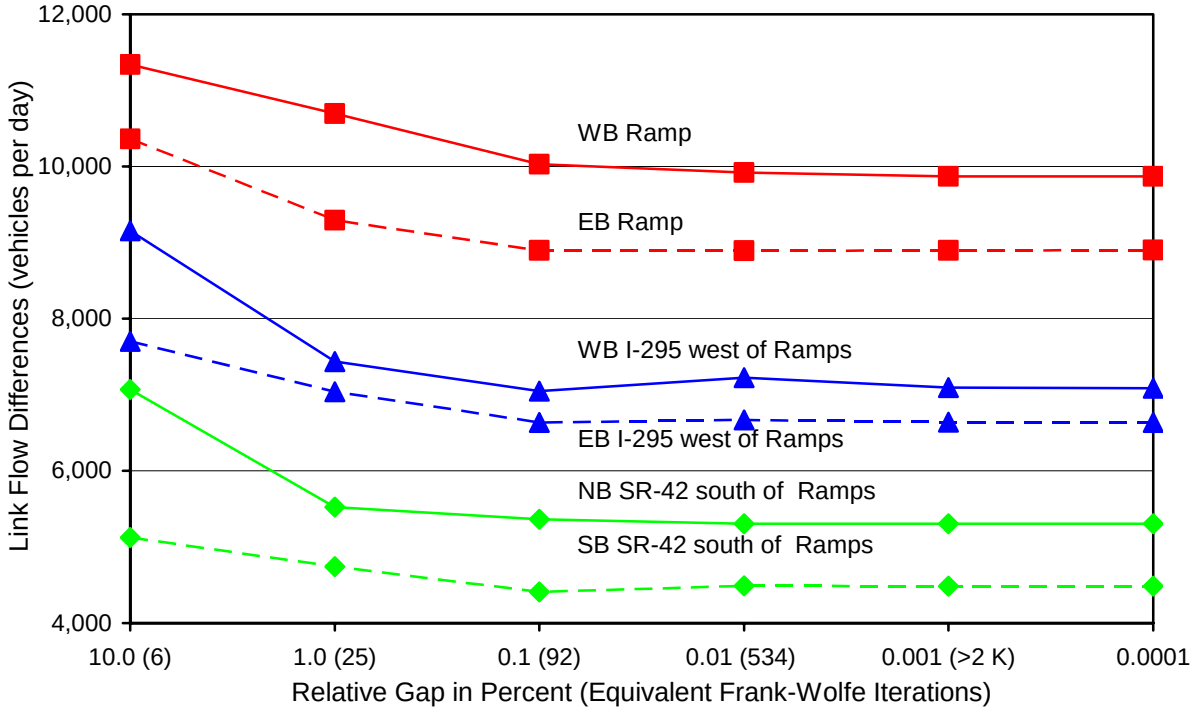


Fig. 4. Link flow differences (build less no-build) versus relative gap
New Jersey freeway links with negative flow differences

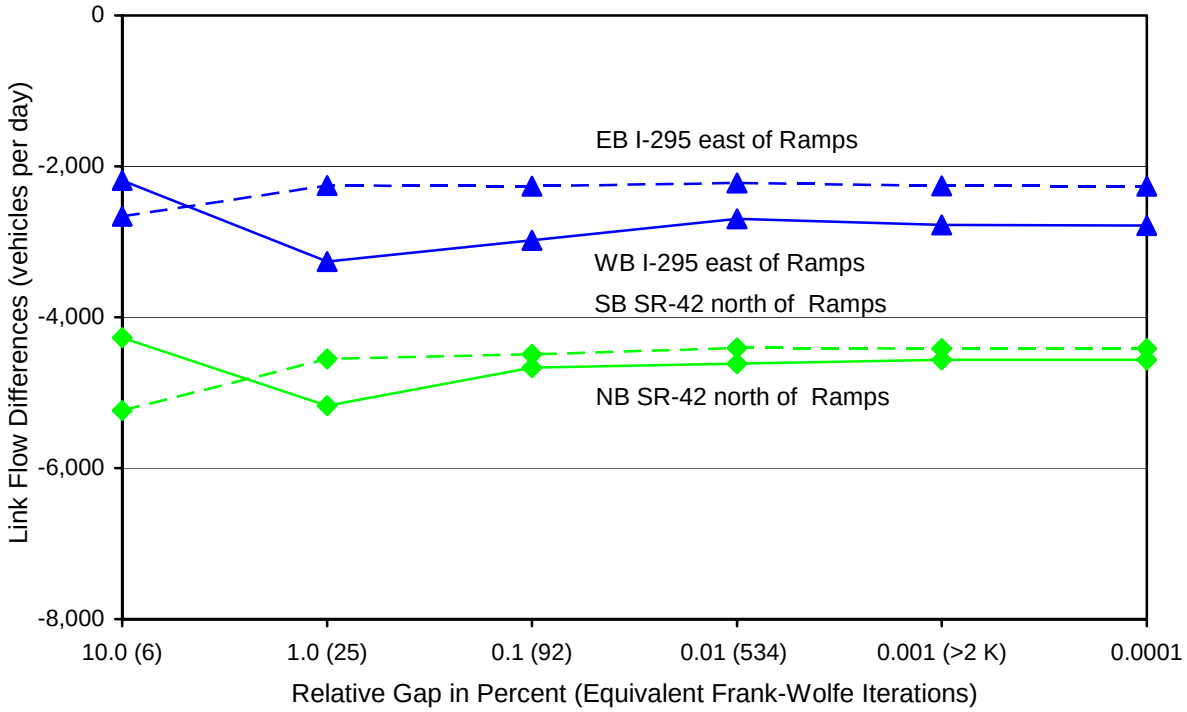


Fig. 5. Error relative to the converged solution versus relative gap
New Jersey freeway links with positive flow differences

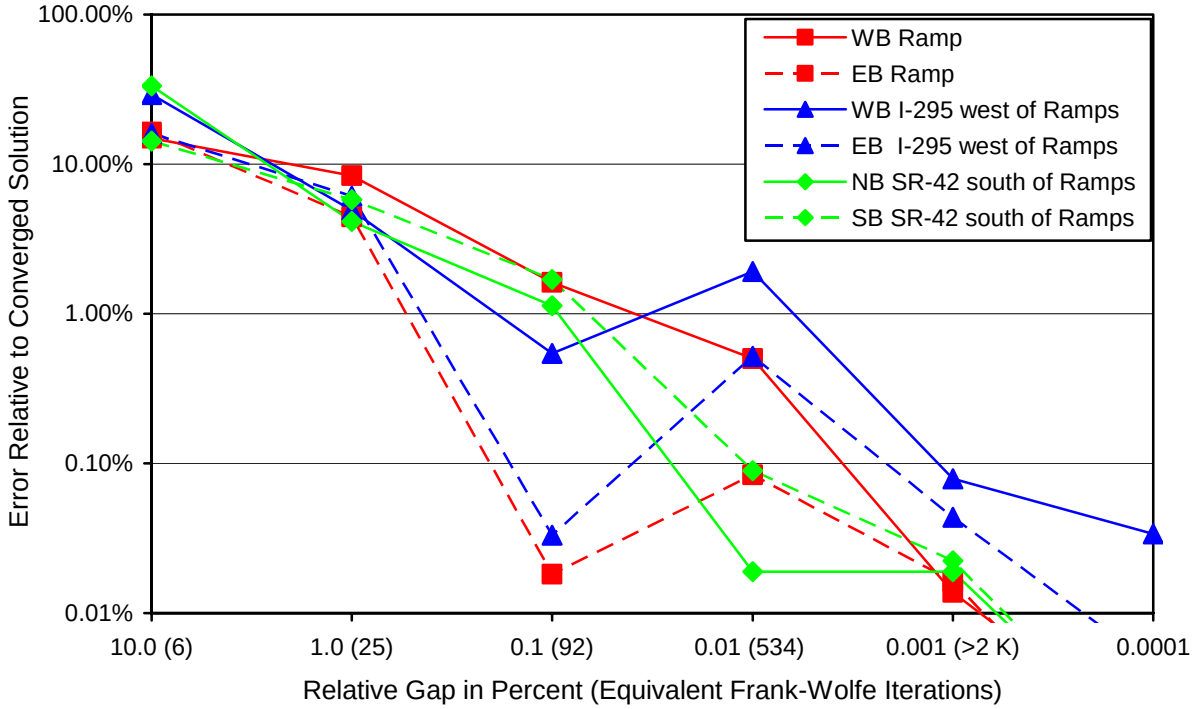


Fig. 6. Error relative to the converged solution versus relative gap
New Jersey freeway links with negative flow differences

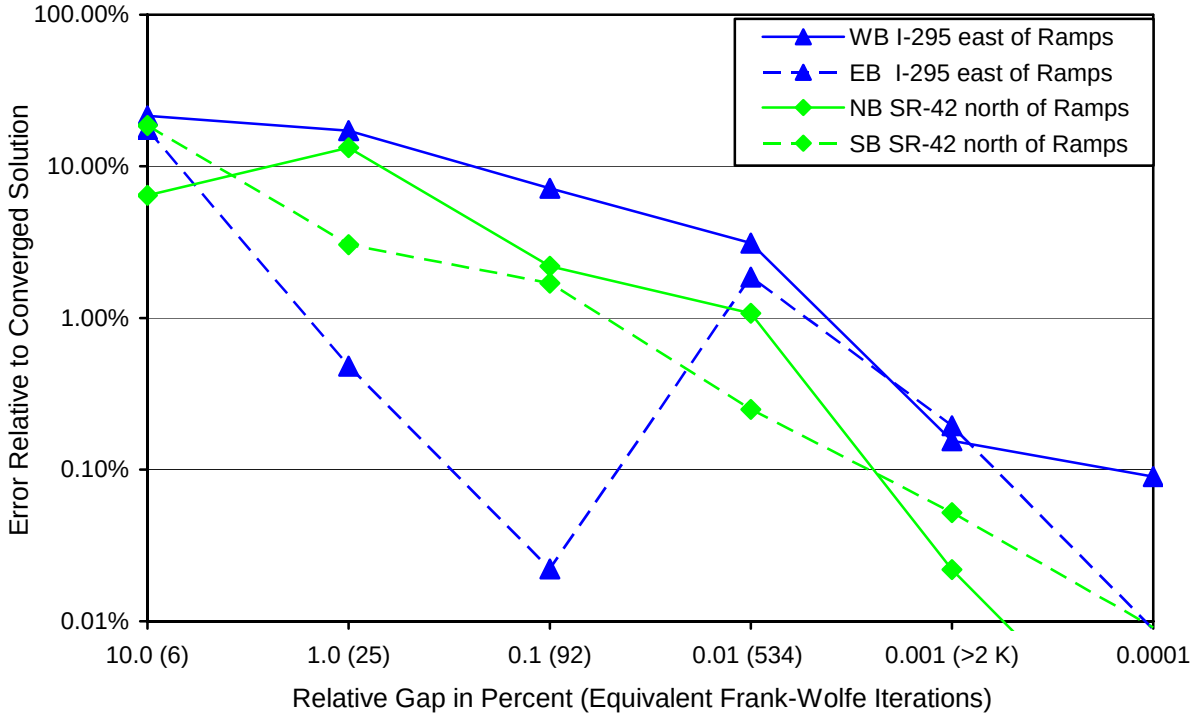


Fig. 7. Link flow differences (build less no-build) versus relative gap
Delsea Drive crossing I-295 west of SR-42

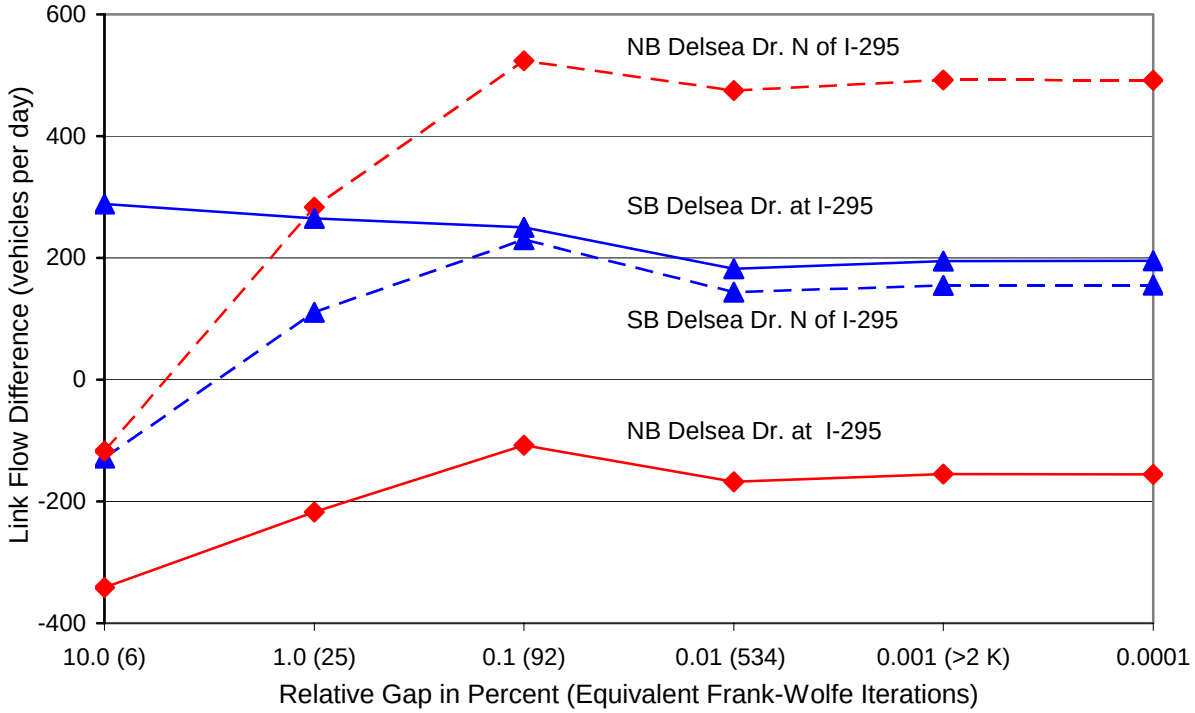
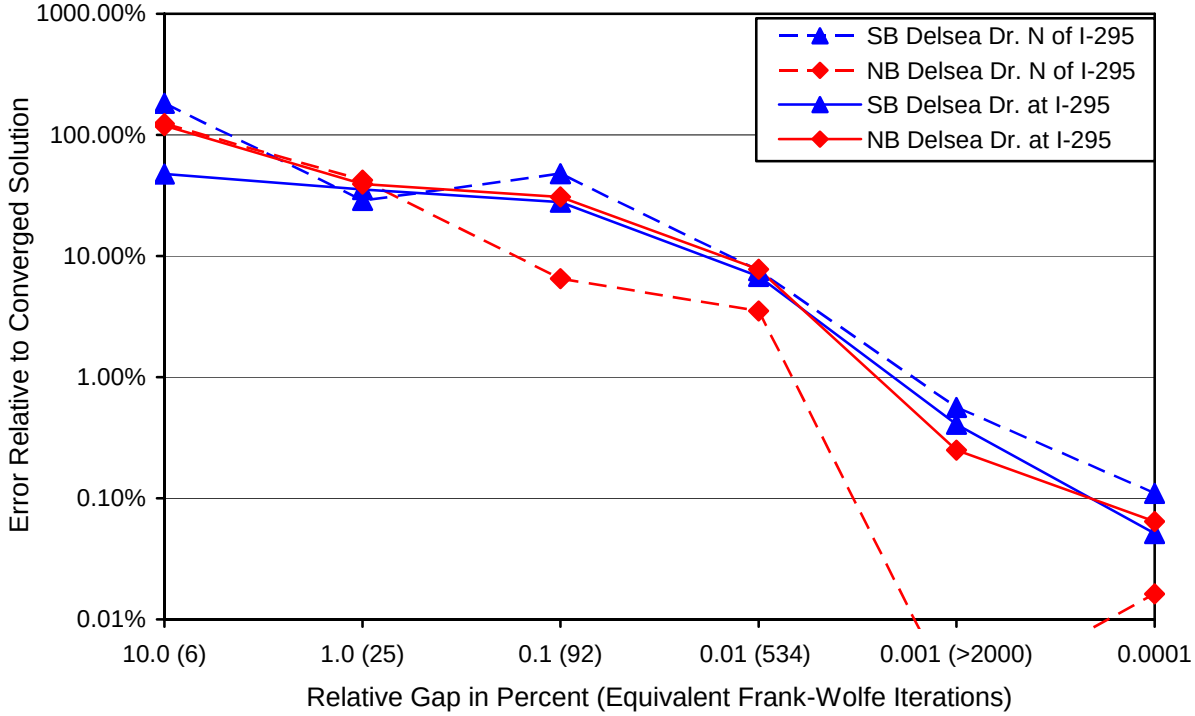
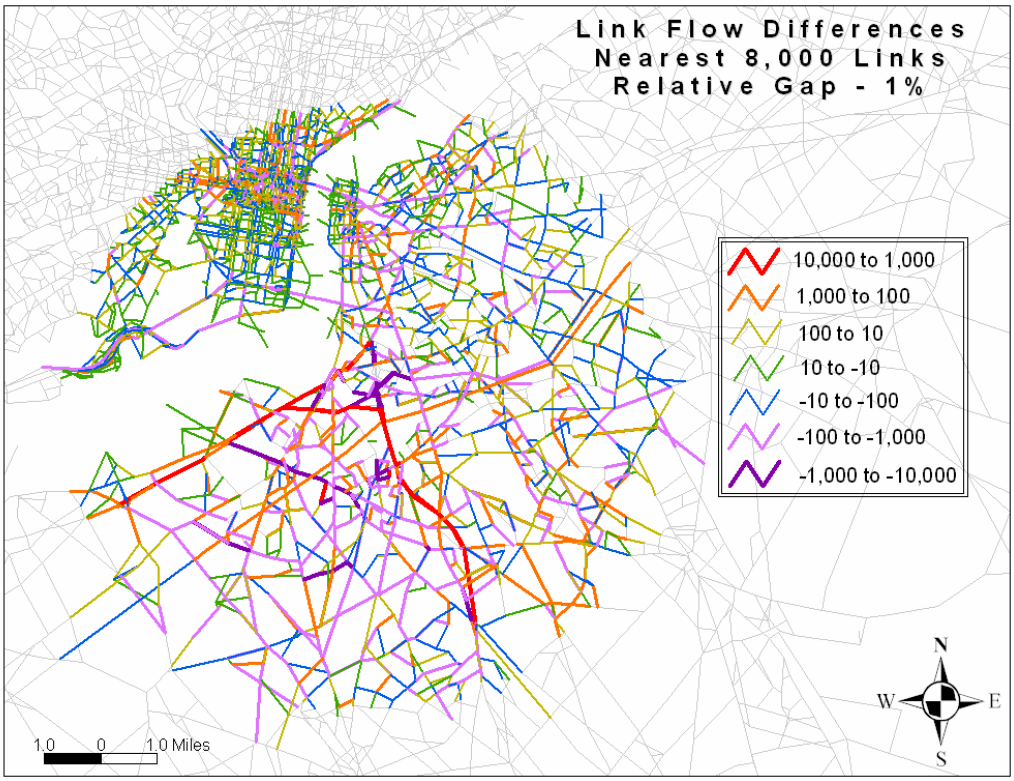
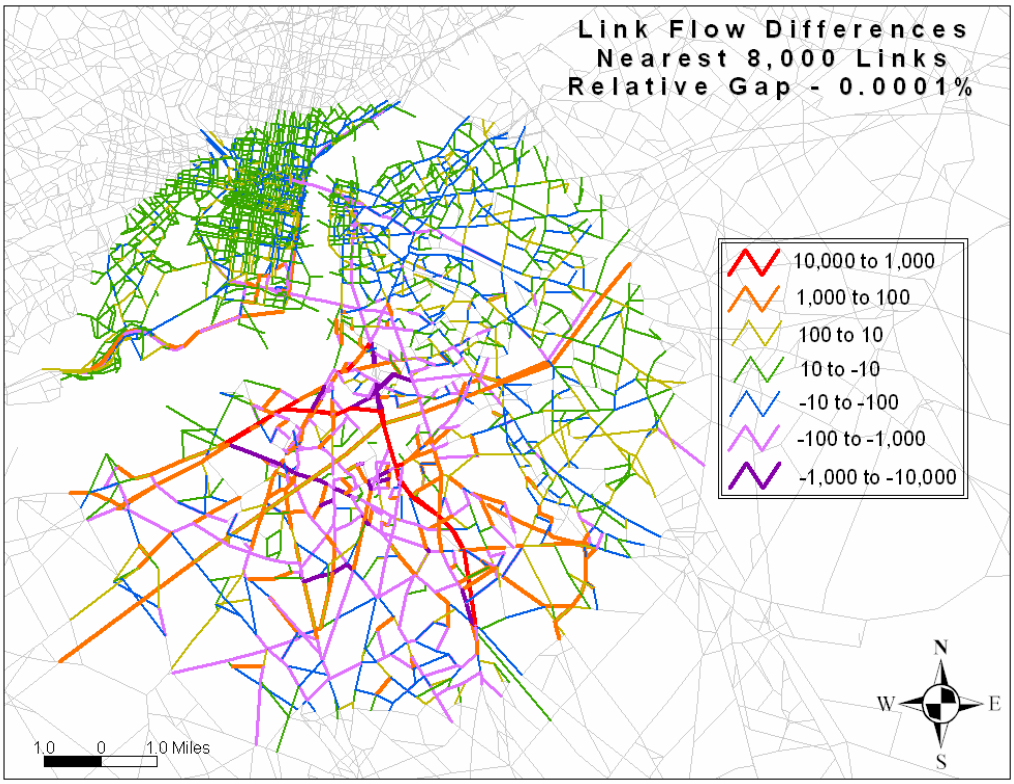
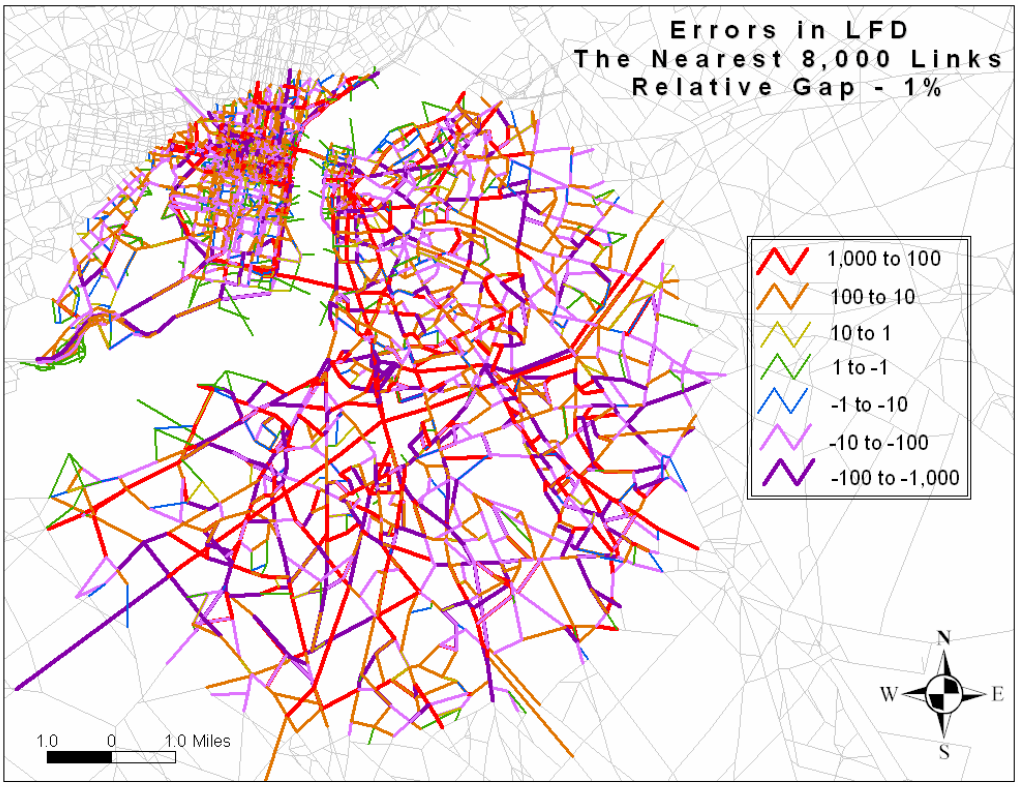


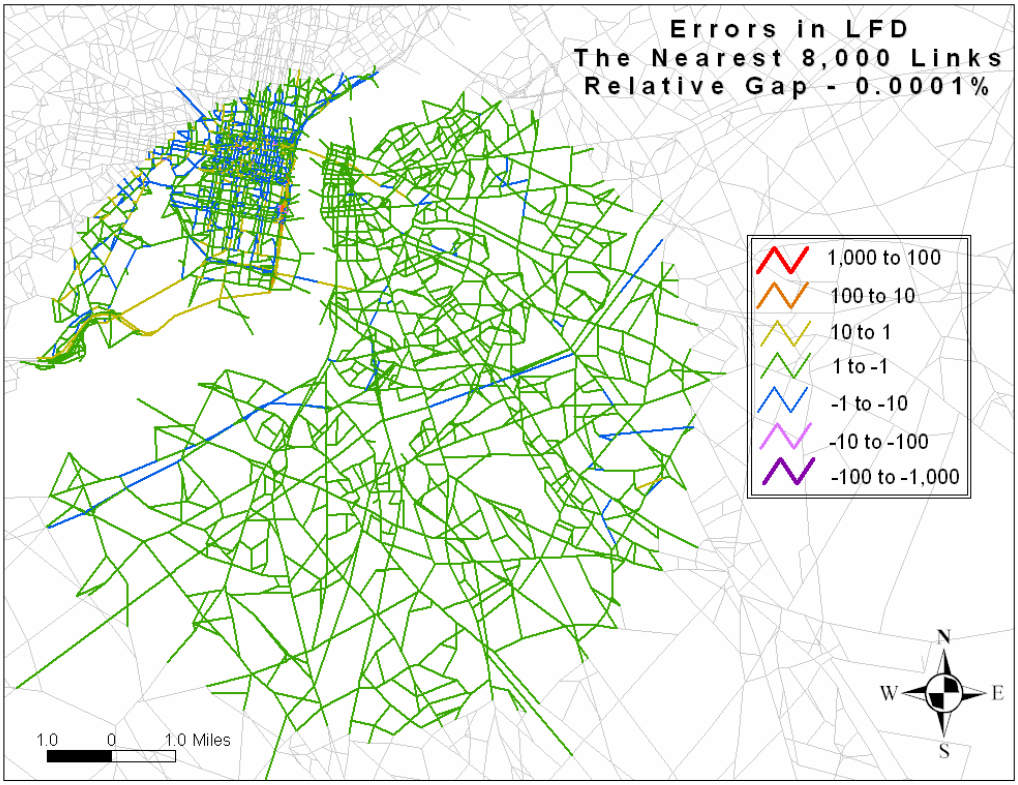
Fig. 8. Error relative to the converged solution versus relative gap
Delsea Drive crossing I-295 west of SR-42

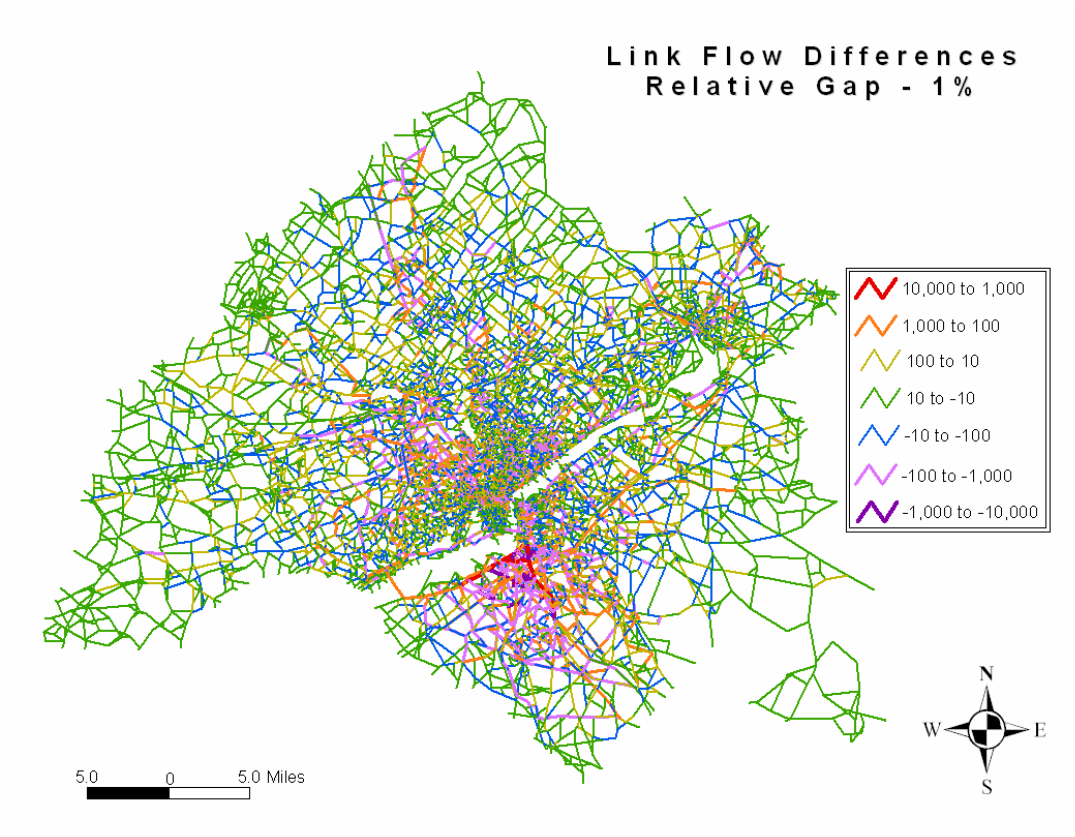


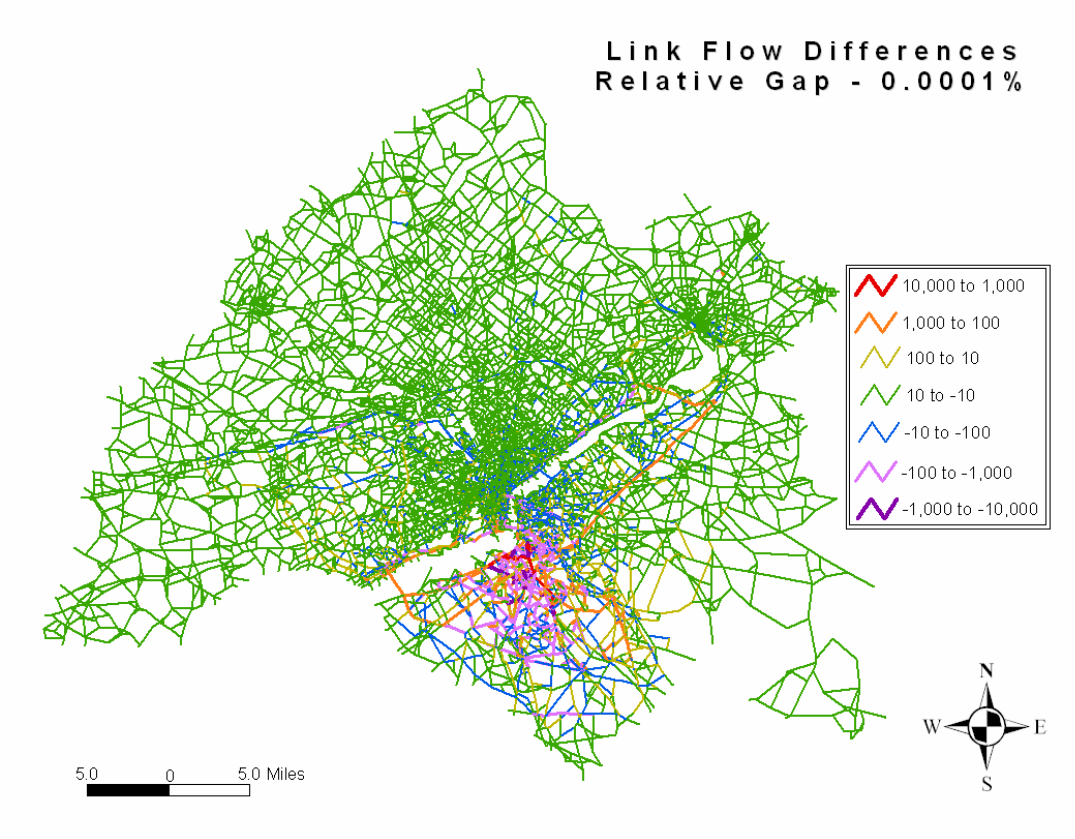

Link Flow Differences
Nearest 8,000 Links
Relative Gap - 1%
10,000 to 1,000
1,000 to 100
100 to 10
10 to -10
-10 to -100
-100 to -1,000
-1,000 to -10,000
N
W
E
S
1.0 0 1.0 Miles

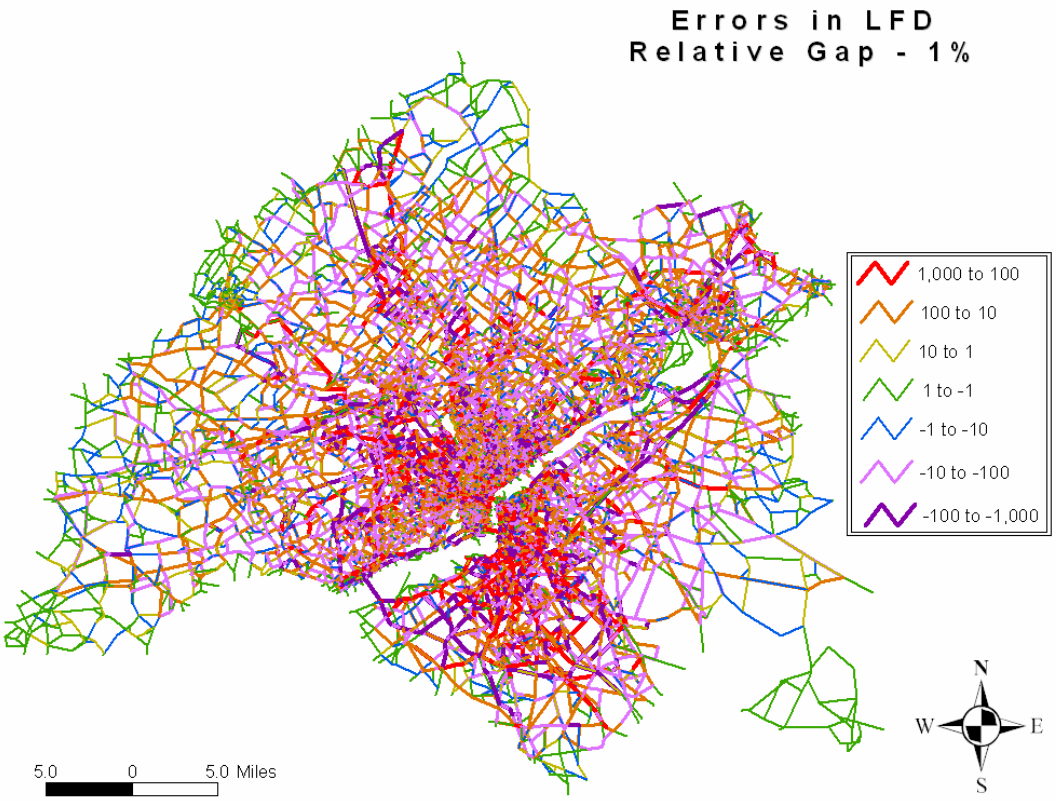

Link Flow Differences
Nearest 8,000 Links
Relative Gap - 0.0001%
10,000 to 1,000
1,000 to 100
100 to 10
10 to -10
-10 to -100
-100 to -1,000
-1,000 to -10,000
N
W
E
S
1.0 0 1.0 Miles

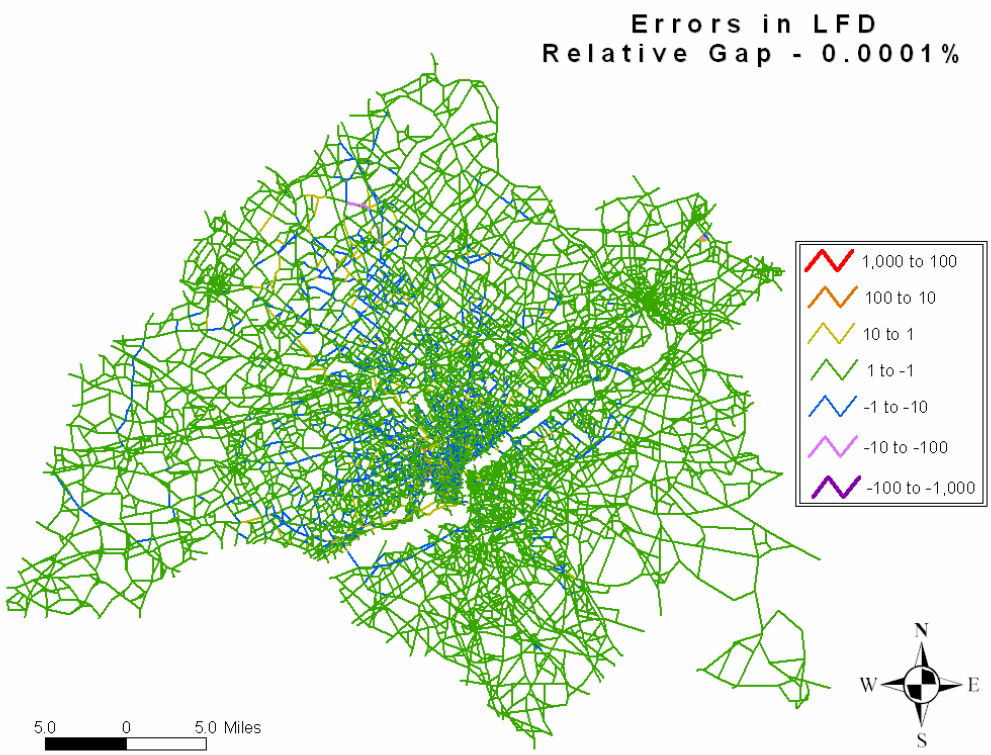

Errors in LFD
The Nearest 8,000 Links
Relative Gap - 1%
1,000 to 100
100 to 10
10 to 1
1 to -1
-1 to -10
-10 to -100
-100 to -1,000
N
W
E
S
1.0 0 1.0 Miles

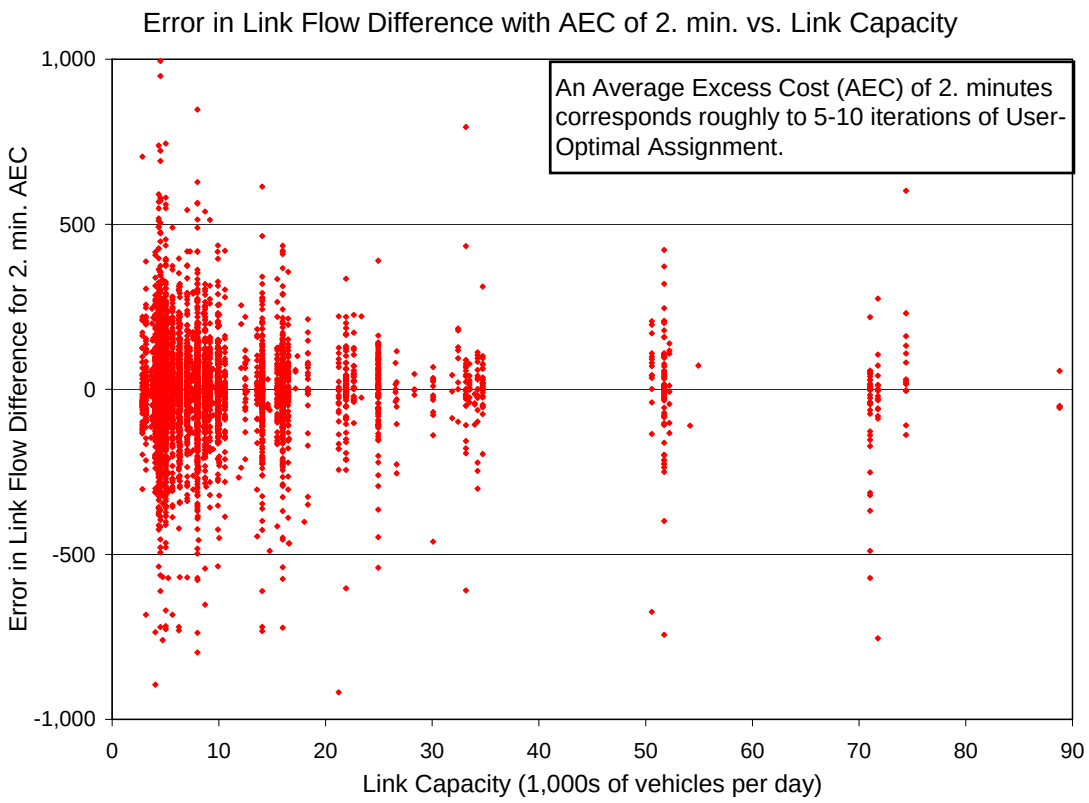

Errors in LFD
The Nearest 8,000 Links
Relative Gap - 0.0001%
1,000 to 100
100 to 10
10 to 1
1 to -1
-1 to -10
-10 to -100
-100 to -1,000
N
W
E
S
1.0 0 1.0 Miles


Link Flow Differences
Relative Gap - 1%
10,000 to 1,000
1,000 to 100
100 to 10
10 to -10
-10 to -100
-100 to -1,000
-1,000 to -10,000
N
W
E
S
5.0
0
5.0 Miles


Link Flow Differences
Relative Gap - 0.0001%
10,000 to 1,000
1,000 to 100
100 to 10
10 to -10
-10 to -100
-100 to -1,000
-1,000 to -10,000
N
W
E
S
5.0
0
5.0 Miles


Errors in LFD
Relative Gap - 1%
1,000 to 100
100 to 10
10 to 1
1 to -1
-1 to -10
-10 to -100
-100 to -1,000
N
W
E
S
5.0
0
5.0 Miles


Errors in LFD
Relative Gap - 0.0001%
1,000 to 100
100 to 10
10 to 1
1 to -1
-1 to -10
-10 to -100
-100 to -1,000
N
W
E
S
5.0
0
5.0 Miles


Error in Link Flow Difference with AEC of 2. min. vs. Link Capacity
An Average Excess Cost (AEC) of 2. minutes corresponds roughly to 5-10 iterations of User-Optimal Assignment.
Error in Link Flow Difference for 2. min. AEC
1,000
500
0
-500
-1,000
0
10
20
30
40
50
60
70
80
90
Link Capacity (1,000s of vehicles per day)

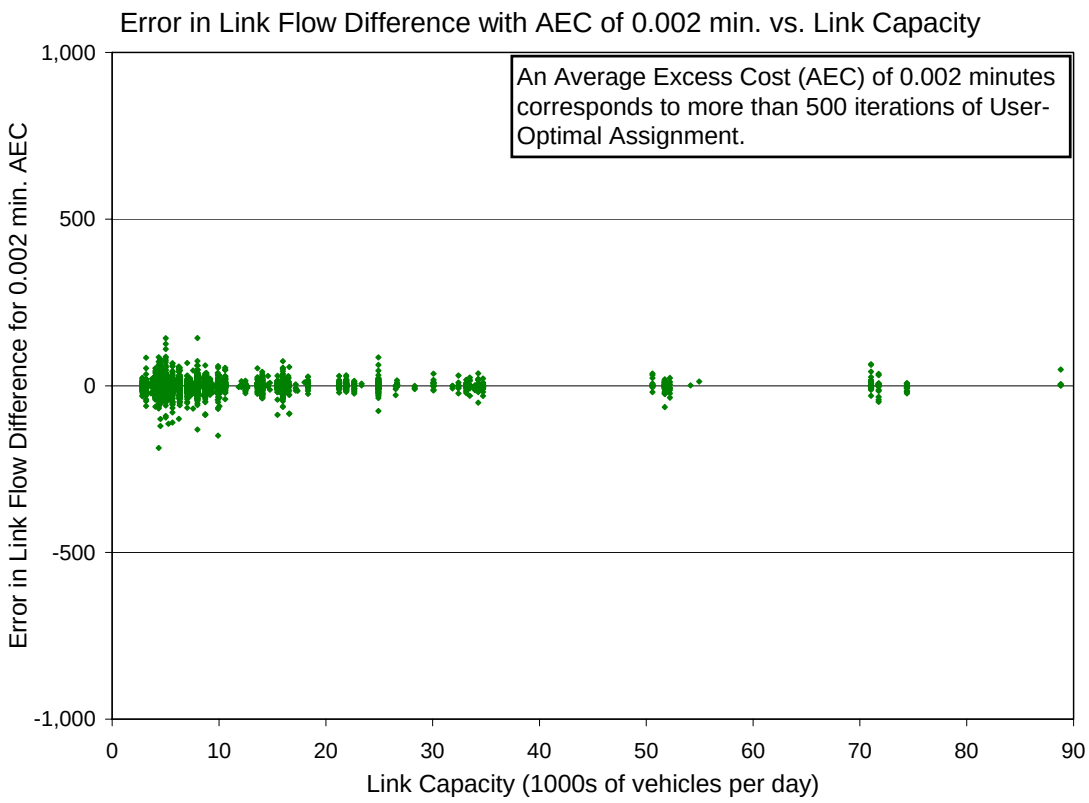

Error in Link Flow Difference with AEC of 0.002 min. vs. Link Capacity
An Average Excess Cost (AEC) of 0.002 minutes corresponds to more than 500 iterations of User-Optimal Assignment.
Error in Link Flow Difference for 0.002 min. AEC
1,000
500
0
-500
-1,000
0
10
20
30
40
50
60
70
80
90
Link Capacity (1000s of vehicles per day)

TABLE 1. Computational Effort for OBA and Frank-Wolfe Algorithms

Relative gap (%)	Equivalent number of Frank-Wolfe iterations	Computation effort for the algorithm (h)		Excess costs of origin-based solutions	
		Frank-Wolfe (*EMME/2*)	OBA	Total (h)	Average (min)
1.0 E + 01	6	0.18	0.96	384,067.9	1.6 E + 00
1.0 E + 00	25	0.64	1.53	28,254.7	1.2 E - 01
1.0 E - 01	92	2.31	2.35	3,129.5	1.3 E - 02
1.0 E - 02	534	13.28	3.41	338.1	1.4 E - 03
1.0 E - 03	>2000[1]	>50.00	5.80	30.1	1.3 E - 04
1.0 E - 04			8.04	3.2	1.4 E - 05
1.0 E - 05			10.82	0.3	1.2 E - 06

[1] Relative gap at 2000 iterations = 0.0024% (2.4E-03%)

TABLE 2. Delaware Valley Regional Planning Commission Road Network Indicators

Indicator	congested	free-flow[1]
Average travel time (min)	23.42	17.84
Average travel distance (km)	12.64	12.64
Space-mean-speed (kph)	32.41	42.54

Note: Total flow = 14,336,061 vehicles per day. Total vehicle hours of travel = 4,334,532 h per day.
[1] at user-equilibrium flows

Four additional maps show the same results at the regional level. From the first pair, we can observe where changes in link flow differences actually occur as a result of introducing the ramps. From the second pair, showing the errors in link flow differences, we can see that it is possible to achieve a solution that eliminates errors throughout the region covered by these 40,000 links.

Finally, two charts show how the errors in link flow differences over the entire region change from the solution with an average excess cost of 2 minutes (relative gap of 10%) to an average excess cost of 0.002 minutes (relative gap of 0.01%).

Table 1 provides additional details of the solution by the Origin-based Assignment algorithm and the Frank-Wolfe algorithm widely used in practice, as illustrated here by the EMME/2 software system.

3.6 Dynamic Route Choice on a Road Network

This subsection is based on a paper by Boyce, Lee and Ran (2001). For a review of related research, see Peeta and Ziliaskopoulos (2001).

This subsection has four objectives. First, we are concerned with predicting route choices of motorists in congested road networks, that is, networks in which travel times depend on

congested vehicle flows, which are prevalent in large urban areas. Second, we are interested in dynamic versions of the classic *traffic assignment problem*, as the problem of predicting motorists' route choices in a road network is traditionally known. In this context, *dynamic* means that each variable is time-specific. In fact, dynamic issues enter the problem in several ways:

1. route choices are based on route travel times from origins to destinations that depend on time-dependent link flows;
2. route travel times may refer to the network situation at an instant of time (*instantaneous*), such as the departure time, or the travel times experienced by travelers during their journeys (*actual* or *ideal*);
3. departure flow rates are time-dependent, and ultimately must be regarded as choices that depend on desired arrival times preceded by travel over a dynamically changing road network.

Third, we are interested in analytical representations of these complex phenomena. That is, we seek to formulate a system of equations and inequalities, or equivalent representations, which captures the essence of the physical and behavioral system of interest. The advantages of this approach are several:

1. such representations are specific and precise, although understanding their properties may require extensive analysis;
2. a mathematical formulation can be studied to determine if a solution exists, and whether or not it is unique and stable;
3. solution algorithms to well-formulated models can be devised and their convergence properties determined, both analytically and computationally.

Finally, we are concerned here with model validation; that is, we want to know whether these complex mathematical models predict the observed choices of routes, departure times, link flows, travel times, etc. Efforts to validate models are severely restricted at present by the lack of observations on phenomena of interest. Ironically, such data will likely be provided by the technology that has most motivated research on the dynamic traffic assignment problem: *intelligent transportation systems*. Dynamic models of travel behavior are needed to plan, design and operate these systems. Through their introduction, we can hope to obtain adequate data, presently unavailable, to validate and further refine the models themselves.

In this subsection, we seek to provide an introduction to an approach to this problem, to describe our approach and to discuss issues related to its successful implementation. The subsection is organized in the following way. First, we provide a short overview to indicate the three approaches to formulating the dynamic traffic assignment problem. Second, formulations of two versions of the problem are presented as an example of analytical formulations. Following a brief discussion of solution algorithms, we offer a brief conclusion.

Approaches to investigating the analytical dynamic traffic assignment problem can be grouped into three categories: mathematical programming, optimal control theory, and variational inequality. Mathematical programming has a long history in studying the traffic assignment

problem stemming from the first formulation of the static traffic assignment problem by Beckmann, McGuire and Winsten (1956); see Patriksson (1994) for the history of the static model. Many algorithms have been proposed and applied to solve mathematical programming-based dynamic traffic assignment models. However, in the context of traffic assignment, mathematical programming has an inherent technical limitation, which sometimes causes it to fail to provide a suitable description of traffic interactions and dynamics, namely the asymmetric nature of travel cost functions and time-dependent interaction of traffic flow and travel time. Optimization formulations generally include integrals of these link cost functions; if the function is asymmetric, its integral is path-dependent, and hence the formulation fails.

Optimal control theory by its nature is well suited for describing dynamic systems; hence, its application to modeling dynamic transportation networks is attractive. Several optimal control theory-based models use inflow as the control variable, whereas exit flow is considered to be a function of link flow. Although this formulation provides an explicit relationship between exit flow and link flow, it has several drawbacks:

1. if the exit flow function is concave, it is not possible to establish an optimal control model of the dynamic user-optimal traffic assignment problem with multiple origin-destination pairs;
2. exit flow functions may cause instantaneous flow propagation. The exit flow rate must be positive in order to satisfy the exit flow function; if the initial flow is zero, it causes unrealistic flow propagation;
3. no equivalent optimal control-based model exists if strict link capacities are considered.

Despite our contributions to this approach (Ran and Shimazaki, 1989a,b; Ran et al., 1993; Ran and Boyce, 1994), we choose not to emphasize it in this review because its limitations have been demonstrated.

Compared with mathematical programming and optimal control, variational inequality (VI) formulations provide a more attractive approach to formulating dynamic route choice problems. VI problems may be regarded as a generalization of constrained optimization problems, complementarity problems, and fixed point problems. Because of its more general properties and capabilities, VI has gained increasing attention during the last decade; see Nagurney (1999) for a review. In the context of dynamic network modeling, issues of asymmetric cost functions and capacitated assignment problems can be accommodated by the VI approach, assuming that certain monotonicity and convexity conditions are met.

3.5.1 Problem Formulation as a Variational Inequality

This section summarizes a dynamic route choice model, which can be extended to incorporate motorists' departure/arrival time choice jointly with route choice (Ran and Boyce, 1996); see also Chen (1999) for a related approach. In the next subsection, some basic definitions and dynamic network flow constraints are stated; Table 1 provides a list of definitions and symbols for convenient reference. Based on actual link travel times, the ideal dynamic user-optimal (DUO) route choice model is then formulated. These formulations are presented to illustrate the concepts and formulation of the dynamic traffic assignment problem using VI methods.

Table 1 Definitions and Notation

$x_a(t)$	number of vehicles on link *a* at time *t*
$x_{ap}^{rs}(t)$	number of vehicles on link *a* and route *p* between O-D pair *rs* at time *t*
$u_a(t)$	inflow rate of link *a* at time *t*
$v_a(t)$	exit flow rate of link *a* at time *t*
$v_{ap}^{rs}(t)$	exit flow rate of link *a* on route *p* between O-D pair *rs* at time *t*
$E_p^{rs}(t)$	cumulative number of vehicles arriving at destination *s* from origin *r* on route *p* by time *t*
$e_p^{rs}(t)$	arrival flow rate at destination *s* from origin *r* on route *p* at time *t*
$f^{rs}(t)$	departure flow rate from origin *r* toward destination *s* at time *t*
$A(j)$	set of links whose tail node is *j*
$B(j)$	set of links whose head node is *j*
$\tau_a(t)$	actual travel time over link *a* for flows entering link *a* at time *t*
$\bar{\tau}_a(t)$	estimated actual travel time over link *a* for flows entering link *a* at time *t*
$\eta_p^{rs}(t)$	travel time actually experienced over route *p* by motorists departing origin *r* toward destination *s* at time *t*
$\pi^{rs}(t)$	minimal travel time actually experienced by motorists departing origin *r* for destination *s* at time *t*
$\bar{\pi}^{rs}(t)$	estimated minimal travel time actually experienced by motorists departing origin *r* for destination *s* at time *t*
$\Omega_a^{rs}(t)$	perceived actual travel time for flows departing origin *r* toward destination *s* over route *p* at time *t*
$\Omega_a^{rj^*}(t)$	difference of the minimal travel time from *r* to *j* and the travel time from *r* to *j*, via the minimal travel time route from *r* to *i* and link *a* for motorists departing from origin *r* at time *t*
ξ_a^r	departure time interval for flow from zone *r* on link *a*

3.5.1.1 Network Flow Constraints

The constraint conditions necessary for the formulation of dynamic route choice models are defined first. In addition to flow conservation conditions, flow propagation constraints are emphasized. Other important constraints include link capacity and spillback constraints. The constraint set for a typical VI model is summarized as follows.

Relationship between link status and link flow variables:

$$\frac{dx_{ap}^{rs}}{dt} = u_{ap}^{rs}(t) - v_{ap}^{rs}(t) \quad \forall a, p, r, s; \tag{1}$$

where $x_{ap}^{rs}(t)$ is the number of vehicles on link *a* and route *p* between OD pair *rs* at time *t*,

$u_{ap}^{rs}(t)$ is the inflow rate into link *a* on route *p* between OD pair *rs* at time *t*, and

$v_{ap}^{rs}(t)$ is the exit flow rate from link *a* on route *p* between OD pair *rs* at time *t*.

$$\frac{dE_p^{rs}(t)}{dt} = e_p^{rs}(t) \quad \forall p, r, s \neq r; \tag{2}$$

where $E_p^{rs}(t)$ is the cumulative number of vehicles arriving at destination *s* from origin *r* on route *p* by time *t*, and

$e_p^{rs}(t)$ is the arrival flow rate at destination *s* from origin *r* on route *p* at time *t*.

<u>Flow conservation constraints:</u>

$$f^{rs}(t) = \sum_{a \in A(r)} \sum_{p} u_{ap}^{rs}(t) \quad \forall r, s; \tag{3}$$

where $f^{rs}(t)$ is the departure flow rate from origin *r* toward destination *s* at time *t* (given);

$$\sum_{a \in B(j)} v_{ap}^{rs}(t) = \sum_{a \in A(j)} u_{ap}^{rs}(t) \quad \forall j, p, r, s; j \neq r, s; \tag{4}$$

$$\sum_{a \in B(s)} \sum_{p} v_{ap}^{rs}(t) = e^{rs}(t) \quad \forall r, s; s \neq r; \tag{5}$$

where *A*(*j*) is the set of links whose tail node is *j* (after *j*), and *B*(*j*) is the set of links whose head node is *j* (before *j*); note that nodes *r* and *s* are specific cases of node *j*.

<u>Flow propagation constraints:</u>

$$x_{ap}^{rs}(t) = \sum_{b \in \tilde{p}} \{x_{bp}^{rs}[t + \tau_a(t)] - x_{bp}^{rs}(t)\} + \{E_p^{rs}[t + \tau_a(t)] - E_p^{rs}(t)\} \qquad \forall a \in B(j); j \neq r; p, r, s; \tag{6}$$

where $\tilde{p}$ denotes the subroute from node *j* to destination *s*, and $\tau_a(t)$ is the actual travel time over link *a* for flows entering link *a* at time *t*.

<u>Definitional constraints:</u>

$$\sum_{rsp} u_{ap}^{rs}(t) = u_a(t), \quad \sum_{rsp} v_{ap}^{rs}(t) = v_a(t), \quad \forall a; \tag{7}$$

$$\sum_{rsp} x_{ap}^{rs}(t) = x_a(t), \quad \forall a; \tag{8}$$

where $u_a(t)$ is the inflow rate into link *a* at time *t*, $v_a(t)$ is the exit flow rate from link *a* at time *t*, and $x_a(t)$ is the number of vehicles on link *a* at time *t*.

Nonnegativity conditions:

$$x_{ap}^{rs}(t) \geq 0, \quad u_{ap}^{rs}(t) \geq 0, \quad v_{ap}^{rs}(t) \geq 0 \quad \forall a, p, r, s; \tag{9}$$

$$e_p^{rs}(t) \geq 0, \quad E_p^{rs}(t) \geq 0 \quad \forall p, r, s; \tag{10}$$

Boundary conditions:

$$E_p^{rs}(0) = 0, \quad \forall p, r, s; \tag{11}$$

$$x_{ap}^{rs}(0) = 0, \quad \forall a, p, r, s; \tag{12}$$

3.5.1.2 Model Formulation

Dynamic route choice models can be formulated based on either actual or instantaneous travel times. The *instantaneous* link travel time at time *t* is defined as the travel time that would be experienced by motorists traversing a link if prevailing traffic conditions remain unchanged. The *actual* link travel time is the travel time over a link actually experienced by motorists. This time is also called the future or forecast time, since it is not observable at time *t*.

Let $\tau_a(t)$ be the actual travel time over link experienced by motorists entering link *a* at time *t*, which is assumed to depend on the number of vehicles $x_a(t)$, the inflow $u_a(t)$ and the exit flow $v_a(t)$ on link *a* at time *t*. It follows that:

$$\tau_a(t) = \tau_a[x_a(t),\ u_a(t),\ v_a(t)] \qquad \forall a \tag{13}$$

Similarly, the actual route travel time is the time over a route actually experienced by motorists. Define $\eta_p^{rs}(t)$ as the travel time actually experienced over route *p* by motorists departing origin *r* toward destination s at time *t*. Using a recursive formula, the route travel time $\eta_p^{rs}(t)$ can be computed for all allowable routes. Assume route *p* consists of nodes $(r, 1, 2, \cdots, i, \cdots, s)$. Let $\eta_p^{rs}(t)$ denote the travel time actually experienced over route *p* from origin *r* to node *i* by motorists departing origin *r* at time *t*. Then, a recursive formula for route travel time $\eta_p^{rs}(t)$ is:

$$\eta_p^{ri}(t) = \eta_p^{r(i-1)}(t) + \tau_a[t + \eta_p^{r(i-1)}(t)] \quad \forall p, r, i; i = 1, 2, \cdots, s; a = (i-1, i).$$

The formulation of the ideal DUO route choice problem is based on the underlying choice criterion that each motorist uses the route that minimizes his/her future (actual) travel time when departing from the origin to destination. Thus, for any OD pair under ideal DUO route choice conditions, motorists departing the origin at the same time must arrive at the destination at the same time (Wardrop, 1952).

Route-Time-Based Model

In this problem, the time-dependent origin-destination trip pattern is assumed to be known *a priori*. That is, the departure times of motorists are given. The ideal dynamic user-optimal (DUO) route choice problem is to determine the dynamic status of links and the inflow and exit flow variables at each instant of time resulting from motorists using minimal-time routes, given the network, the link travel time functions and the time-dependent OD departure rate requirements. Consider the following dynamic generalization of the conventional static user-optimal state.

> *Travel-Time-Based Ideal DUO State: If, for each OD pair at each instant of time, the* actual *travel times experienced by motorists departing at the same time are equal and minimal, the dynamic traffic flow over the network is in a travel-time-based* ideal *dynamic user-optimal state.*

If the actual link travel time $\tau_a(t)$ is determined, the minimal travel time $\pi^{rs}(t)$ actually experienced by motorists departing origin *r* for destination *s* at time *t* can be computed as $\pi^{rs}(t) = \min_p \left(\eta_p^{rs}(t)\right)$. Thus, $\pi^{rs}(t)$ is a functional of all link flow variables at time *w*:

$$\pi^{rs}(t) = \pi^{rs}[u(\omega), v(\omega), x(\omega) \mid \omega \geq t]$$

This functional is not available in closed form. Nevertheless, it can be evaluated when $u(\omega), v(\omega)$ and $x(\omega)$ are temporarily fixed. The route-time-based ideal DUO route choice conditions can then be defined as follows:

$$\eta_p^{rs^*}(t) - \pi^{rs}(t) \geq 0 \quad \forall p, r, s; \tag{14}$$

$$f_p^{rs^*}(t)[\eta_p^{rs^*}(t) - \pi^{rs^*}(t)] = 0 \quad \forall p, r, s; \tag{15}$$

$$f_p^{rs^*} \geq 0 \quad \forall p, r, s. \tag{16}$$

The asterisk in the above equations denotes that the flow variables are the optimal solutions under the route-time-based ideal DUO state. The equivalent variational inequality formulation of route-time-based ideal DUO route choice conditions (14)-(16) may be stated as follows:

> *The dynamic traffic flow pattern satisfying network constraint set* (*1*)-(*12*) *is in a route-time-based ideal DUO route choice state if and only if it satisfies the variational inequality problem:*

$$\int_0^T \sum_{rs} \sum_a \eta_p^{rs^*}(t)[f_p^{rs}(t) - f_p^{rs^*}(t)]dt \geq 0 \tag{17}$$

Solving variational inequality (17) is equivalent to solving ideal DUO route choice conditions (14)-(16). The properties of this formulation are described in Ran and Boyce (1996).

Link-Time-Based Model

Route-time-based VI models have an intuitive interpretation. With new solution methods, such as disaggregate simplicial decomposition (DSD) or gradient projection (GP), route enumeration is no longer required to obtain a solution to the route-time-based model. However, route-time-based models still require considerable computational effort for networks of realistic size. Therefore, a link-time-based VI model is proposed. The link-time-based ideal DUO route choice conditions given below imply the route-time-based ideal DUO route choice conditions.

The set of dynamic network constraints defined above apply to both the link-time-based VI model and the route-time-based VI model. The basic difference between the two models is that the link-time-based model is formulated using link-based flow variables instead of the route-based variables in the route-time-based model. Let $\Omega_a^{rj^*}(t)$ denote the difference of the minimal travel time from *r* to *j* and the travel time from *r* to *j* via minimal travel time route from *r* to *i* and link *a* for motorists departing from origin at time *t*:

$$\Omega_a^{rj^*}(t) = \pi^{ri^*}(t) + \tau_a[t + \pi^{ri^*}(t)] - \pi^{rj^*}(t) \quad \forall a, r; a = (i, j). \tag{18}$$

The link-time-based ideal DUO route choice conditions can then be stated as follows:

$$\Omega_a^{rj^*}(t) \geq 0 \qquad \forall a = (i, j), r; \tag{19}$$

$$u_a^{rs^*}[t + \pi^{ri^*}(t)]\,\Omega_a^{rj^*}(t) = 0 \qquad \forall a = (i, j), r, s; \tag{20}$$

$$u_a^{rs}[t + \pi^{ri^*}(t)] \geq 0 \qquad \forall a = (i, j), r, s. \tag{21}$$

Link-time-based ideal DUO route choice conditions (19)-(21) imply route-time-based ideal DUO route choice conditions (14)-(16). Then, the equivalent variational inequality formulation of link-time-based ideal DUO route choice conditions (19)-(21) may be stated as follows:

The dynamic traffic flow pattern satisfying constraints (1)-(12) is in a link-time-based ideal DUO route choice state if and only if it satisfies the variational inequality problem:

$$\int_0^T \sum_{rs} \sum_a \Omega_a^{rj^*}(t) \cdot \{u_a^{rs}[t + \pi^{ri^*}(t)] - u_a^{rs^*}[t + \pi^{ri^*}(t)]\}\,dt \geq 0 \tag{22}$$

Solving variational inequality (22) is equivalent to solving link-time-based ideal DUO route choice conditions (19)-(21). Further properties are given in Ran and Boyce (1996). The above models are limited in the sense that the motorists are assumed to be homogenous. One way to

relax this assumption is to stratify them into classes, in which each class has distinctly different characteristics. Each class of motorists is associated with a disutility or generalized cost function. Thus, the ideal DUO route choice conditions are defined for each class on the basis of travel disutilities instead of travel time only. Then, a multi-class DUO model can be formulated based on travel costs or disutilities instead of travel times.

3.5.2 Solution Algorithms

The link-time-based ideal DUO route choice model is selected as an example for describing a solution algorithm for solving this class of dynamic models. First, the continuous time formulation of the DUO route choice program is transformed into a discrete time VI formulation. Then, the VI model is reformulated as a discrete optimization problem under relaxation and solved using the linearization technique embedded in a relaxation procedure. In the relaxation procedure, the estimated link travel times are updated iteratively and the linearization technique is applied in each iteration to solve the resulting nonlinear programming problem (NLP). For the linearized subproblem, an expanded time-space network can be constructed so that the subproblem can be decomposed according to O-D pairs and can be viewed as a set of minimal-cost route problems. Flow propagation constraints representing the relationship between link flows and travel times are satisfied in the minimal-cost route search so that only flow conservation constraints for links and nodes remain.

To convert our continuous VI model into a discrete VI model, the time period [0,*T*] is subdivided into *N* small time increments. Each time increment is a unit of time. To simplify the formulation, we modify the estimated actual travel time on each link in the following way so that each estimated travel time is equal to a multiple of the time increment.

$$\bar{\tau}_a(n) = i \quad \text{if} \quad i - 0.5 \le \bar{\tau}_a(n) < i + 0.5,$$

where *i* is an integer and $0 \le i \le K$. The continuous VI problem is then reformulated as a discrete VI problem:

$$\sum_{n=1}^{N} \sum_{rs} \sum_{a} \Omega_a^{rj^*}(n) \cdot \{u_a^{rs}[n + \pi^{ri^*}(n)] - u_a^{rs^*}[n + \pi^{ri^*}(n)]\} \ge 0 \tag{23}$$

For each relaxation iteration, we temporarily fix: a) the actual travel time $\tau_a(n)$ in the link flow propagation constraints as $\bar{\tau}_a(n)$; b) the minimal travel times $\pi^{ri^*}(n)$ as $\bar{\pi}^{ri^*}(n)$ and $\pi^{rj^*}(n)$ as $\bar{\pi}^{rj^*}(n)$ for each link and each origin node. Then, the objective function of an optimization problem can be formulated, which is equivalent to the discrete VI under relaxation, as follows:

$$\min_{u,v,x,E} Z = \sum_k \sum_a \left\{ \int_0^{u_a(k)} \tau_a\left[x_a(k), \omega, v_a(k)\right] dw + \sum_r u_a^r(k)\left[\bar{\pi}^{ri}\left(\xi_a^r\right) - \bar{\pi}^{rj}\left(\xi_a^r\right)\right] \right\} \tag{24}$$

where the new time interval index k is defined as $k = \xi_a^r + \overline{\pi}^{ri}(\xi_a^r)$ with departure time interval ξ_a^r calculated by using the above equation once interval k, origin r and arrival node i are given. Note that both minimal travel times $\overline{\pi}^{ri}(\xi_a^r)$ and $\overline{\pi}^{rj}(\xi_a^r)$ are fixed temporarily at each relaxation iteration. In summary, in each relaxation step, we need to solve a discrete NLP consisting of objective function (24) and its associated constraints.

To develop an operational dynamic traffic assignment model for real-time applications, computational requirements pose a major challenge. Clearly, more efficient algorithms than the one outlined above are needed. To this end, the continuous variational inequality (VI) is first transformed into a discrete variational inequality. Then, a relaxation method is used to solve this discrete VI. During each relaxation, a nonlinear programming (NLP) problem is formulated and solved. Unlike the previous algorithm, which uses inflow, exit flow and number of vehicles as the basic variables, this new algorithm uses the inflow variable as the only independent variable to formulate the NLP. To solve this NLP using the Frank-Wolfe algorithm, a linear programming (LP) subproblem is formulated in each iteration.

One of the major contributions of this new algorithm (Ran et al, 2002) is that it solves the LP subproblem on a representation of the physical network rather than on an expanded time-space network. Thus, the new algorithm avoids the time-space network expansion and searches for shortest routes on the expanded network, the previous method for solving the analytical DTA model. Consequently, the computation time can be shortened substantially. Computational experience on the Sioux Falls network benchmarks the efficiency of this algorithm.

3.5.3 Conclusions and Comments on Recent Developments

In this review, we have tried to convey the basic concepts and exciting challenges that characterize analytical dynamic traffic assignment research at present. Except for the pioneering efforts of Merchant and Nemhauser, Friesz and Carey, no one had really thought about dynamics in travel choice modeling as recently as 15 to 20 years ago. Likewise, intelligent transportation systems were unknown.

Then everything changed, largely as a result of technological developments in computing and communications. Suddenly, ITS was possible, if not fully operational. As a result, dynamic travel choice models were suddenly in vogue and actually needed. Fortunately, scientific interest and understanding of the necessary mathematical constructs, optimal control theory and variational inequalities, were at hand although not widely understood. The result has been an exciting period of innovation and synthesis, which promises to continue.

Since the paper by Boyce, Lee and Ran was published in 2001, research has continued on analytical formulations of dynamic route choice models, as well as simulation-based approaches. It is more apparent today than in 2001 that simulation methods in combination with analytical models are likely to be successful in solving large-scale problems. Simulation methods can be used in place of link travel time functions (eqn. 13) and also to estimate flow propagation (eqn. 6). The use of these methods allows researchers to focus their efforts on model formulation and

solution methods, and to be less concerned with devising analytical functions for traffic flow relationships. Even so, care must be exercised to assure that the model properties are respected.

3.6 Transit Route Choice

The traditional approach to transit assignment is rather simple:

1. represent all transit submodes (bus, rapid transit, commuter rail, express bus) in on network;

2. find the minimal travel time route from origin zone *p* to destination zone *q* considering access, waiting, boarding, in-vehicle, and transfer times;

3. assign all transit trips from zone *p* to zone *q* to the minimal time route (all-or-nothing).

Even if all-or-nothing assignment to minimal time routes is regarded as adequate, the problem of representing transit networks is more complex than for auto route choice. Moreover, the use of only one route between each O-D pair is simplistic if the transit network is complex with several options for each trip. While only large cities in North America have such systems, in Europe and Asia, complex systems are more common. The transit assignment method developed by Speiss and Florian (1989) respond to this need for a more complex approach. If competing transit services are offered, the results may be simplistic if it is not used. The desired solution is an allocation of O-D flows over several transit routes which minimizes the total expected (mean) generalized cost including walking, access, waiting, boarding, in-vehicle, transfer and egress times plus fare.

References

Ahuja, R. K., J. B. Orlin, S. Pallottino and M. G. Scutellà; (2003) Dynamic Shortest Paths Minimizing Travel Times and Costs, *Networks* **41**, 197-205

Bar-Gera, H. (1999) *Origin-Based Algorithms for Transportation Network Modeling*, Technical Report #103**,** National Institute of Statistical Sciences, Research Triangle Park, NC. (Ph.D. thesis in civil engineering, University of Illinois at Chicago, 1999); downloadable file available at: http://www.niss.org/downloadabletechreports.html

Bar-Gera, H. (2002) Origin-Based Algorithm for the Traffic Assignment Problem, *Transportation Science* **36**, 398-417.

Beckmann, M., C. B. McGuire, and C. B. Winsten (1956) *Studies in the Economics of Transportation*, Yale University Press, New Haven.

Bothner, P. and W. Lutter (1982) Ein direktes Verfahren zur Verkehrsumlegung nach dem 1. Prinzip von Wardrop, Arbeitsbericht Nr. 1, Forschungsbereich: Verkehrssysteme, Universität Bremen, Germany.

Boyce, D., D.-H. Lee and B. N. Janson (1998) A Variational Inequality Model of an Ideal Dynamic User-Optimal Route Choice Problem, In *Transportation Networks: Recent Methodological Advances*, Bell, M.G.H., ed., Pergamon, Amsterdam, 289-302.

Boyce, D., B. Ralevic-Dekic and H. Bar-Gera (2004) Convergence of Traffic Assignments: How Much Is Enough? *ASCE Journal of Transportation Engineering* **130,** 49-55.

Boyce, D., D.-H. Lee and B. Ran (2001) Analytical Models of the Dynamic Traffic Assignment Problem, *Networks and Spatial Economics* 1, 377-390.

Chen, A. and D.-H. Lee (1999) Path-based Algorithms for Large Scale Traffic Equilibrium Problem: A Comparison Between DSD and GP, presented at the 78th Annual Meeting of the Transportation Research Board, Washington, DC.

Chen, H.-K. (1999) *Dynamic Travel Choice Models*. Springer-Verlag, Berlin.

Chen, H.-K, C.-W Chang and M.-S. Chang (1999) Comparison of Link-Based Versus Route-Based Algorithms in the Dynamic User-Optimal Route Choice Problem, *Transportation Research Record* **1678**, 242-250.

Frank, M. and P. Wolfe (1956) An Algorithm for Quadratic Programming, *Naval Research Logistics Quarterly* **3**, 95-110.

Gallo, G. and S. Pallottino (1988) Shortest Path Algorithms, *Annals of Operations Research* **13,** 3-79.

Kupsizewska, D. and D. Van Vliet (1999) 101 Uses for Path-based Assignment, Transport Planning Methods: *Proceedings of Seminar C* held at PTRC Planning and Transport Summer Annual Meeting, University of Sussex, UK.

Larsson, T. and M. Patriksson (1992) Simplicial Decomposition with Disaggregated Representation for the Traffic Assignment Problem. *Transportation Science* **26**, 4-17.

LeBlanc, L. J. (1975) An Algorithm for the Discrete Network Design Problem, *Transportation Science* **9**, 183-199

Lee, D.-H. (1996) Formulation and Solution of a Dynamic User-Optimal Route Choice Model on a Large-Scale Traffic Network, Ph.D. thesis in civil engineering, Univ. of Illinois, Chicago.

Lee, D.-H., A. Chen, and Y. Nie (2000) Computer Experiments and Statistical Analysis on Algorithms for Traffic Assignment Problems, Working Paper, Department of Civil Engineering, National University of Singapore, Singapore.

Nagurney, A. (1999) *Network Economics*, Kluwer Academic Publishers, Boston.

Patriksson, M. (1994) *The Traffic Assignment Problem*, VSP, Utrecht, The Netherlands.

Pallottino, S. and M. G. Scutellà (1998) Shortest Path Algorithms in Transportation Models: Classical and Innovative Aspects, *Equilibrium and Advanced Transportation Modelling*, P. Marcotte and S. Nguyen, eds., Kluwer, 245-281.

Peeta, S. and A. K. Ziliaskopoulos (2001) Foundations of Dynamic Traffic Assignment : The Past, the Present and the Future, *Networks and Spatial Economics* **1**, 233-265.

Ran, B. and D. Boyce (1994) *Dynamic Urban Transportation Network Models*, Springer, Berlin.

Ran, B. and D. Boyce (1996) *Modeling Dynamic Transportation Networks*, Springer, Berlin.

Ran B., D.-H. Lee and S. Shin (2002) New Algorithm for a Multiclass Dynamic Traffic Assignment Model, *ASCE Journal of Transportation Engineering* **128**, 323-335.

Ran, B. and Shimazaki, T. (1989a) A General Model and Algorithm for the Dynamic Traffic Assignment Problem, *Proceedings of Fifth World Conference on Transport Research*, Yokohama, Japan, 463-477.

Ran, B. and Shimazaki, T. (1989b) Dynamic User Equilibrium Traffic Assignment for Congested Transportation Networks, presented at the Fifth World Conference on Transport Research, Yokohama, Japan.

Spiess, H. and M. Florian (1989) Optimal Strategies: A New Assignment Model of Transit Networks, *Transportation Research* **21B**, 395-412.

Wardrop, J. G. (1952) Some Theoretical Aspects of Road Traffic Research, *Proceedings of the Institution of Civil Engineers, Part II* **1**, 325-378.

4. Mode Choice

If we view travel choices as a hierarchy, then route choice seems naturally to fall at the lowest level, since it is specific to a travel mode. It is less clear what choice should be considered next in the hierarchy. Since mode choices can be made on a daily basis, assuming a car is available, choice of mode may be regarded as the next choice in this hierarchy, as is assumed here. We will also use the mode choice model to introduce the concept of dispersion of choices. Dispersion refers to a departure from a strict, or deterministic, cost minimizing assumption. We say choices are dispersed to a higher cost alternative, if the choices are not strictly cost minimizing as we assumed was the case with auto route choice case. Actually, we could have introduced this concept in the route choice model, leading to the *stochastic route choice model*. Since that model has some inherent limitations, we are introducing the dispersion concept here.

4.1 Deterministic Mode Choice

First, let's extend the auto route choice model to include choice of mode. Recall our auto route choice formulation,

$$\min_{(\mathbf{h})} z(\mathbf{h}) = \sum_a \int_0^{f_a} c_a(x)dx$$

$$st: \sum_{r \in R_{pq}} h_r = d_{pq}, \qquad p \in P; q \in Q$$

$$h_r \geq 0, \qquad r \in R_{pq,} \qquad p \in P; q \in Q$$

$$\text{where } f_a \equiv \sum_{pq} \sum_{r \in R_{pq}} h_r \delta_r^a, \qquad a \in A$$

Up to now, all flows have been defined as vehicles per hour. Now add a subscript *m* to denote mode, with *a* representing the auto mode and *t* representing the transit mode. Even though we have used the subscript *a* to denote links, the use of *a* to represent the auto mode should be clear by its context. The addition of the subscript *a* to the OD pair *pq* denotes *person* flows per hour from zone *p* to zone *q* by auto, and likewise *t* for transit denotes *person* flows per hour. To be more definite, consider the following relaxation of the auto route choice formulation:

$$\min_{(\mathbf{h},\mathbf{d})} z(\mathbf{h},\mathbf{d}) = \eta \cdot \sum_a \int_0^{f_a} c_a(x)dx + \sum_{pq} c_{pqt} \cdot d_{pqt}$$

$$st: \qquad \sum_{r \in R_{pqa}} h_r = d_{pqa} / \eta, \qquad p \in P; q \in Q$$

$$\sum_m d_{pqm} = d_{pq}, \qquad p \in P; q \in Q$$

$$h_r \geq 0, \qquad r \in R_{pqa,} \qquad p \in P; q \in Q$$

$$d_{pqm} \geq 0, \; m \in M, \qquad p \in P; q \in Q$$

$$\text{where } f_a \equiv \sum_{pq} \sum_{r \in R_{pqa}} h_r \delta_r^a, \qquad a \in A$$

Note the following additions and changes in the model:

1. A parameter η is added to the objective function term representing auto costs; it represents the ratio of auto occupants to autos (persons per vehicle), and is called auto occupancy;

2. A second term is added to the objective function denoting total transit costs, where c_{pqt} is the fixed generalized cost of travel per person by transit from zone p to zone q, and d_{pqt} is the flow of persons from zone p to zone q by the transit mode t;

3. The right-hand-side of the first constraint has been divided by η to convert auto person flows d_{pqa} from zone p to zone q to auto vehicle flows on route r, h_r;

4. A constraint is added requiring that the variable OD flows by mode sum to the fixed total OD flow;

5. The set of routes by auto has been redesignated as R_{pqa}.

What this formulation states is find the allocation of fixed OD flows to the two modes, and within the auto mode to routes, so as to minimize a function of total generalized cost. Because we want the auto route flows to be User-Optimal, we do not minimize the total cost of auto travel, as we do for transit, but rather retain the sum of the integrals of the links cost functions. Let's examine the KKT optimality conditions to see how this problem formulation works.

The Lagrangean function is defined as follows:

$$L(\mathbf{h},\mathbf{d},\mathbf{u},\rho) = \eta \cdot \sum_a \int_0^{f_a} c_a(x)dx + \sum_{pq} c_{pqt} \cdot d_{pqt} - \sum_{pq} u_{pqa}\left(\sum_{r \in R_{pqa}} h_r - d_{pqa}/\eta\right) - \sum_{pq} \rho_{pq}\left(d_{pqa} + d_{pqt} - d_{pq}\right)$$

The KKT optimality conditions obtained by taking derivatives of this function with respect to h_r, d_{pqa} and d_{pqt} are:

$$\eta \cdot \sum_a c_a(f_a) \cdot \delta_r^a - u_{pqa} \geq 0, \qquad r \in R_{pqa}, p \in P, q \in Q$$

$$h_r\left(\eta \cdot \sum_a c_a(f_a) \cdot \delta_r^a - u_{pqa}\right) = 0, \qquad r \in R_{pqa}, p \in P, q \in Q$$

$$\frac{u_{pqa}}{\eta} - \rho_{pq} \geq 0, \qquad p \in P, q \in Q$$

$$d_{pqa}\left(\frac{u_{pqa}}{\eta} - \rho_{pq}\right) = 0, \qquad p \in P, q \in Q$$

$$c_{pqt} - \rho_{pq} \geq 0, \qquad p \in P, q \in Q$$

$$d_{pqt}\left(c_{pqt} - \rho_{pq}\right) = 0, \qquad p \in P, q \in Q$$

$$\sum_{r \in R_{pqa}} h_r - \frac{d_{pqa}}{\eta} = 0, \qquad p \in P, q \in Q$$
$$\sum_m d_{pqm} - d_{pq} = 0, \qquad p \in P; q \in Q$$
$$\mathbf{h} \geq \mathbf{0},\ \mathbf{d} \geq \mathbf{0}$$

Clearly, this is rather messy. Let's see what we can do to clarify matters. First, consider the case of $h_r > 0$, for some OD pair pq. Then, the travel cost, $c_{pqar} = u_{pqa} / \eta$, which is also true for all used auto routes from zone p to zone q. Next, consider the condition related to d_{pqa}. Since we assumed that $h_r > 0$, therefore $d_{pqa} > 0$, and $u_{pqa} / \eta = \rho_{pq}$, the equilibrium OD cost from zone p to zone q. If $\rho_{pq} = c_{pqt}$, then, $d_{pqt} \geq 0$; otherwise, $c_{pqt} > \rho_{pq}$, and $d_{pqt} = 0$. Without belaboring the analysis further, we draw the following conclusions about this formulation:

1. If OD flows occur by auto, then all used routes have equal cost, and no unused route has a lower cost;

2. The costs of the used auto routes not only determine the OD cost, but also determine whether transit has a sufficiently low cost to be used. If it does, then the OD cost by transit and auto are equal; that is, either there are no transit flows or the transit cost sets a maximum level for the auto costs for each OD pair.

3. If the transit costs are lower than can be achieved by auto, then all OD flows will move by transit. Hence, the solution is in some sense *all-or-nothing* with respect to mode.

4. If the auto occupancy η were not included in the objective function, then the auto OD cost would be different from the OD equilibrium cost by a factor equal to η. For consistency of the formulation, then, we see that the weight η is necessary in the objective function.

From this analysis, we can see that the formulation of the mode and auto route choice model as a deterministic cost minimization problem may be rather unrealistic, for example, we observe in Census data that both auto and transit flows occur between many, many OD pairs. Therefore, we explore an approach to relax the deterministic nature of the formulation through the addition of a certain constraint known as a dispersion or entropy function.

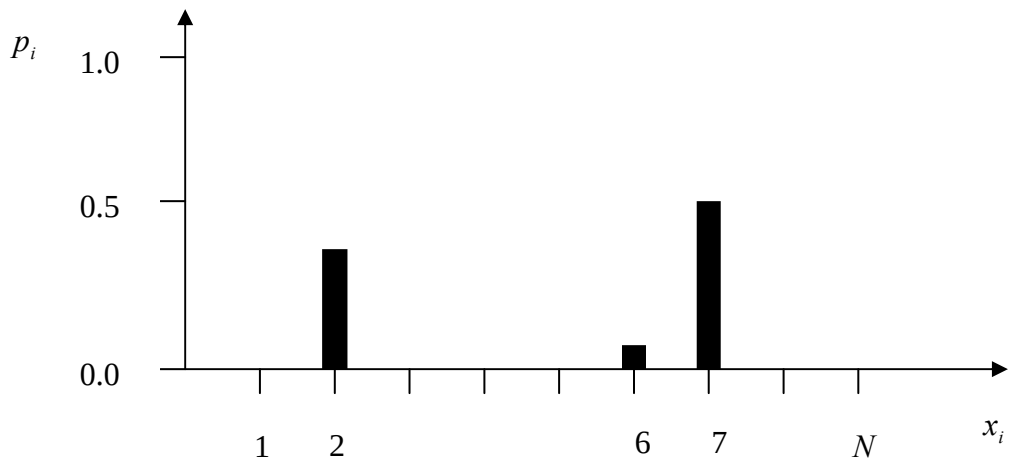
4.2 Dispersion Constraint

What we need to do is to define a constraint that will soften or blur the hard deterministic (all-or-nothing) character of the above model. Not surprisingly, there is a function available for this job, and it has a one-to-one correspondence with a well-known choice function, the logit function. The objective of this subsection is to introduce this function and its properties.

Suppose a probability is defined for each discrete value of a variable:

x_i = the discrete value observed for category i, i = 1, 2, …, N;
p_i = the probability that x_i occurs, or the observed relative frequency with which x_i occurs.

An example is shown by the following bar chart.



Discrete Probability Bar Chart

As we can see in the chart, there are three observed events:

event 2 occurs with observed relative frequency 0.4;
event 6 occurs with observed relative frequency 0.1;
event 7 occurs with observed relative frequency 0.5.

We would like to characterize these observed frequencies with a single measure indicating how dispersed or clustered they are. One possible response to this question is to compute the mean value of the observed response, and then to compute the variance. However, in this case, we would like a measure that doesn't depend on the values of the responses, because they are qualitative or nominal values such as mode or OD pair. The measure we propose is defined as: $S \equiv -\sum_i p_i \ln p_i$, where S is the measure of dispersion and p_i is the observed relative frequency of event i. This function is known as the *entropy function* in statistics and information theory.

For the case in question, the value of S is 0.9433483, if we agree that $0 \cdot \ln 0 = 0.0$. How should we interpret this value? Suppose we agree that the maximum dispersion that can be achieved in the observed events is for them all to have equal frequencies. If there are N possible events, and they have equal relative frequencies, then their relative frequency is $1/N$. Let's compute the dispersion of this case:

$$S_{\max} = -N \cdot \frac{1}{N} \ln\left(\frac{1}{N}\right) = -1 \cdot (\ln(1) - \ln(N)) = -(0 - \ln(N)) = \ln(N)$$

Therefore, the maximum dispersion is the natural logarithm of the number of categories. Note it is a positive number because we were careful to define S with a negative sign. Next, let's consider the minimum value that S can achieve when the events are completely clustered. That occurs if one event has relative frequency of 1.0; all other events have relative frequencies of 0.0.

$$S_{\min} = -(1.0)\cdot\ln(1.0) - (N-1)(0.0)\ln(0.0) = 0.0 - 0.0 = 0.0$$

Hence the range of S is 0.0 to $\ln(N)$. The function S can also be expressed in terms of the observed frequencies, rather than relative frequencies. The two versions are related as follows, where $T = \sum_i T_i$.

$$S = -\sum_i p_i \ln p_i = -\sum_i \left(\frac{T_i}{T}\right)\ln\left(\frac{T_i}{T}\right) = -\frac{1}{T}\sum_i T_i\left[\ln T_i - \ln T\right] = -\frac{1}{T}\sum_i T_i \ln T_i + \ln T$$

Therefore, $-\sum_i T_i \ln T_i = T\cdot S - T\ln T$. The sign of the left-hand side is negative, if all of the observed frequencies are greater than or equal to 1.

A constraint can be formed with the dispersion function in the following way. Generally, the cost minimizing solution is the least dispersed feasible solution, since the mode choices strictly take the least cost mode. By constraining the dispersion of choices to be higher than the minimum, some choices are shifted to the higher cost mode. The form of the constraint is $-\sum_{pqm} d_{pqm}\ln(d_{pqm}) \geq S_0$, where S_0 represents the dispersion of the observed modal choices. The form of this function is concave, as can be seen by plotting S vs. p_1, for the case of two probabilities, p_1 and p_2. Let's add this constraint to the problem defined in Section 4.1, but first simplify the problem definition to exclude route flows by fixing auto costs, as follows.

<u>Fixed Cost Assignment Problem with Deterministic Mode Choice</u>

$$\min_{(\mathbf{d})} z(\mathbf{d}) = \sum_{pqm} c_{pqm}\cdot d_{pqm}$$

$$st: \quad \sum_m d_{pqm} = d_{pq},\ p\in P; q\in Q$$

$$d_{pqm} \geq 0,\ m\in M, p\in P; q\in Q$$

where c_{pqm} = fixed cost of travel by mode m from zone p to zone q

d_{pqm} = travel flow by mode m from zone p to zone q

Adding the dispersion constraint, we obtain:

<u>Fixed Cost Assignment Problem with Stochastic Mode Choice</u>

$$\min_{(\mathbf{d})} z(\mathbf{d}) = \sum_{pqm} c_{pqm}\cdot d_{pqm}$$

$$st: \quad \sum_m d_{pqm} = d_{pq},\ p\in P; q\in Q$$

$$-\sum_{pqm} d_{pqm}\ln d_{pqm} \geq S_0$$

$$d_{pqm} \geq 0,\ m\in M, p\in P; q\in Q$$

The Lagrangean is:

$$L\left(\mathbf{d},\rho,\frac{1}{\mu}\right)=\sum_{pqm}c_{pqm}\cdot d_{pqm}-\sum_{pq}\rho_{pq}\left(d_{pqa}+d_{pqt}-d_{pq}\right)-\frac{1}{\mu}\left(-\sum_{pqm}d_{pqm}\ln d_{pqm}-S_0\right)$$

Note that in the Lagrangean, the entropy term is multiplied by (-1), so the overall term is convex. The optimality conditions are:

$$\begin{aligned}
&c_{pqa}-\rho_{pq}(+1)-\frac{1}{\mu}\left(-\ln d_{pqa}-1\right)=0\\
&c_{pqt}-\rho_{pq}(+1)-\frac{1}{\mu}\left(-\ln d_{pqt}-1\right)=0\\
&d_{pqa}+d_{pqt}-d_{pq}=0\\
&-\sum_{pqm}d_{pqm}\ln d_{pqm}\geq S_0\\
&\frac{1}{\mu}\left(-\sum_{pqm}d_{pqm}\ln d_{pqm}-S_0\right)=0
\end{aligned}$$

Since the person flows are the arguments of logarithmic functions, they all must be positive; otherwise, we would have $\ln(0)\Rightarrow-\infty$. Hence we can delete the complementary slackness constraints on d_{pqm}. Considering the first two conditions, we obtain:

$$\begin{aligned}
&\ln d_{pqm}=\mu\left(\rho_{pq}-c_{pqm}\right)-1,\text{ so}\\
&d_{pqm}=\exp\left(\mu\left(\rho_{pq}-c_{pqm}\right)-1\right),\qquad m\in M,\,p\in P,q\in Q
\end{aligned}$$

Summing the expressions for auto and transit,

$$\begin{aligned}
&\sum_m d_{pqm}=d_{pq}=\exp\left(\mu\rho_{pq}-1\right)\cdot\sum_m\exp\left(-\mu\cdot c_{pqm}\right)\\
&\Rightarrow\exp\left(\mu\cdot\rho_{pq}-1\right)=d_{pq}/\left(\sum_m\exp\left(-\mu\cdot c_{pqm}\right)\right)
\end{aligned}$$

Substituting this expression into the equation for d_{pqm}, we obtain:

$$d_{pqm}=d_{pq}\left(\frac{\exp\left(-\mu\cdot c_{pqm}\right)}{\sum_n\exp\left(-\mu\cdot c_{pqn}\right)}\right),\text{ or }d_{m|pq}=\left(\frac{\exp\left(-\mu\cdot c_{pqm}\right)}{\sum_n\exp\left(-\mu\cdot c_{pqn}\right)}\right),$$

where $d_{m|pq}\equiv d_{pqm}/d_{pq}=$ conditional probability of choosing mode *m*, given OD pair *pq*. Thus, the optimality conditions for this problem correspond to the multinomial logit model with cost sensitivity parameter μ, the reciprocal of the Lagrange multiplier of the dispersion constraint.

4.3 Combined Mode and Route Choice Model

Before proceeding to an examination of the properties of the logit model, let's synthesize the above results. The original formulation with the dispersion constraint inserted is:

$$\min_{(\mathbf{h},\mathbf{d})} z(\mathbf{h},\mathbf{d}) = \eta \cdot \sum_a \int_0^{f_a} c_a(x)dx + \sum_{pq} c_{pqt} \cdot d_{pqt}$$

$$st: \sum_{r \in R_{pqa}} h_r = d_{pqa} / \eta, \qquad p \in P; q \in Q$$

$$\sum_m d_{pqm} = d_{pq}, \qquad p \in P; q \in Q$$

$$-\sum_{pqm} d_{pqm} \ln(d_{pqm}) \geq S_0$$

$$h_r \geq 0, \qquad r \in R_{pqa,}\ p \in P; q \in Q$$

$$d_{pqm} > 0, \qquad m \in M, p \in P; q \in Q$$

$$\text{where } f_a \equiv \sum_{pq} \sum_{r \in R_{pqa}} h_r \delta_r^a, \qquad a \in A$$

The revised Lagrangean is defined as follows.

$$L(\mathbf{h},\mathbf{d},\mathbf{u},\rho) = \eta \cdot \sum_a \int_0^{f_a} c_a(x)dx + \sum_{pq} c_{pqt} \cdot d_{pqt} - \sum_{pq} u_{pqa} \left(\sum_{r \in R_{pqa}} h_r - d_{pqa} / \eta \right)$$
$$- \sum_{pq} \rho_{pq} \left(d_{pqa} + d_{pqt} - d_{pq} \right) - \frac{1}{\mu} \left(-\sum_{pqm} d_{pqm} \ln(d_{pqm}) - S_0 \right)$$

The optimality conditions derived by taking derivatives of this function are stated below.

$$\eta \cdot \sum_a c_a(f_a) \cdot \delta_r^a - u_{pqa} \geq 0, \qquad r \in R_{pqa}, p \in P, q \in Q$$

$$h_r \left(\eta \cdot \sum_a c_a(f_a) \cdot \delta_r^a - u_{pqa} \right) = 0, \qquad r \in R_{pqa}, p \in P, q \in Q$$

$$\left(\frac{u_{pqa}}{\eta} \right) - \rho_{pq} + \frac{1}{\mu} \left(\ln(d_{pqa}) + 1 \right) = 0, \qquad p \in P, q \in Q$$

$$c_{pqt} - \rho_{pq} + \frac{1}{\mu} \left(\ln(d_{pqt}) + 1 \right) = 0, \qquad p \in P, q \in Q$$

$$\sum_{r \in R_{pqa}} h_r - \frac{d_{pqa}}{\eta} = 0, \qquad p \in P, q \in Q$$

$$\sum_m d_{pqm} = d_{pq}, \qquad p \in P; q \in Q$$

$$-\sum_{pqm} d_{pqm} \ln\left(d_{pqm}\right) \geq S_0$$

$$\frac{1}{\mu}\left(-\sum_{pqm} d_{pqm} \ln\left(d_{pqm}\right) - S_0\right) = 0$$

$$\mathbf{h} \geq \mathbf{0},\ \mathbf{d} \geq \mathbf{0}$$

For routes with positive flow, $h_r > 0,\ r \in R_{pqa}$, the UO travel cost is:

$$\frac{u_{pqa}}{\eta} = \sum_a c_a\left(f_a\right) \cdot \delta_r^a \equiv c_{pqa}\text{; therefore,}$$

$$c_{pqa} - \rho_{pq} + \frac{1}{\mu}\left(\ln d_{pqa} + 1\right) = 0$$

$$c_{pqt} - \rho_{pq} + \frac{1}{\mu}\left(\ln d_{pqt} + 1\right) = 0$$

which are the same conditions found above for the case of fixed auto costs. Therefore, for the combined mode and route choice model, the mode choice function is given by:

$$d_{pqm} = d_{pq}\left(\frac{\exp\left(-\mu \cdot c_{pqm}\right)}{\sum_n \exp\left(-\mu \cdot c_{pqn}\right)}\right)$$

and the usual UO conditions hold for auto route costs. This completes the derivation of a model that combines mode and route choice models, which is therefore known as a *combined model*.

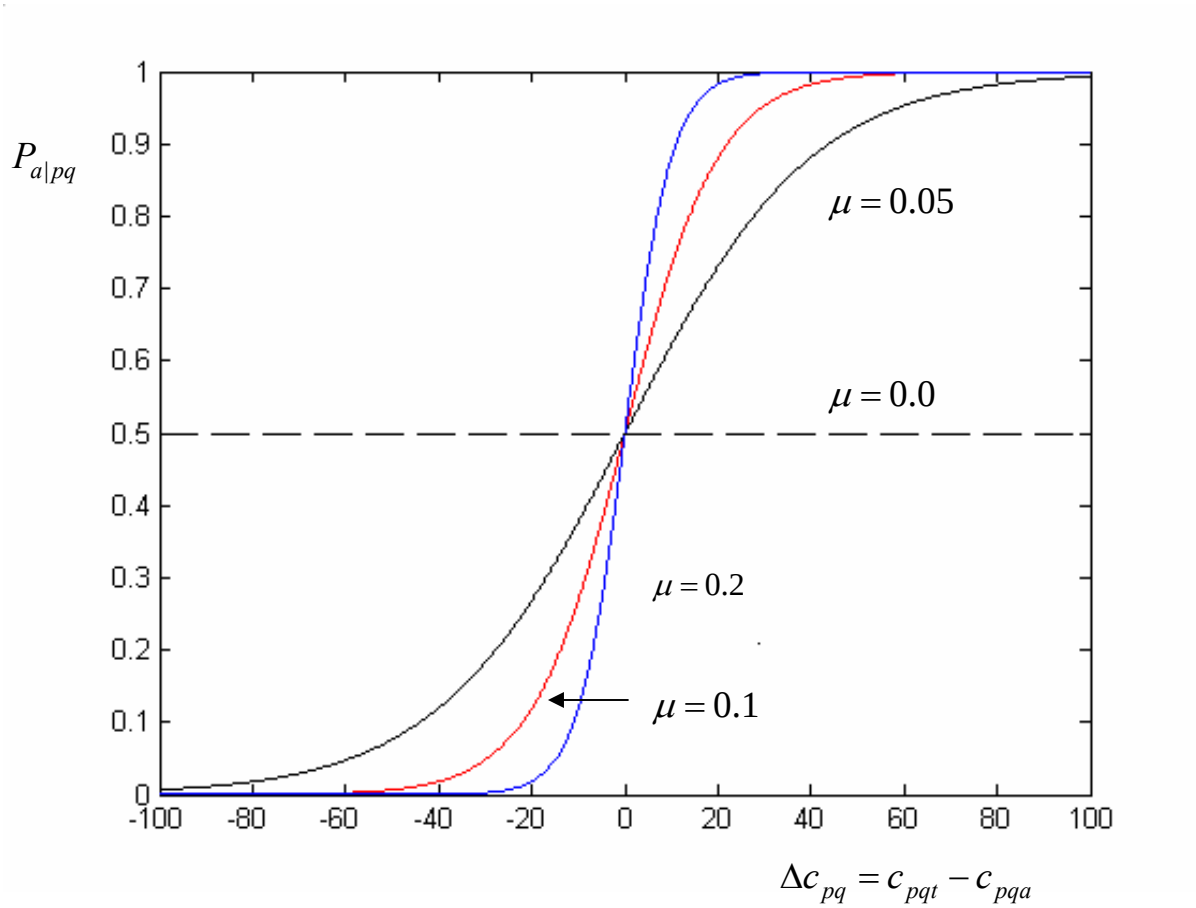
4.4 Properties of the Logit Model

Consider a model of choice of two modes, auto *a* and transit *t*:

$$\frac{d_{pqa}}{d_{pq}} \equiv P_{a|pq} = \frac{\exp\left(-\mu \cdot c_{pqa}\right)}{\exp\left(-\mu \cdot c_{pqa}\right) + \exp\left(-\mu \cdot c_{pqt}\right)} = \frac{1}{1 + \exp\left(-\mu \cdot \left(c_{pqt} - c_{pqa}\right)\right)}$$

Thus, probability of using mode *a*, $P_{a|pq}$, depends on $\Delta c_{pq} = \left(c_{pqt} - c_{pqa}\right)$. From the following figure, we can see the role of the dispersion parameter μ is to determine to what extent travelers prefer the *lower cost* mode. If $\mu = 0.2$, then the choice shifts from transit to auto within a small range of the cost difference Δc_{pq}. If $\mu = 0$, there is no preference between modes.

An important property and shortcoming of the Logit model concerns how the model responds to the addition of a new choice alternative. This property is known as the "independence of irrelevant alternatives" or IIA, property. Because the model is extremely simple, it treats each alternative mode independently without considering whether its attributes are similar to another



An Example of a Logit Function Applied to Mode Choice

mode; i.e. it ignores the correlation of modal attributes. Consider the ratio of two choice probabilities:

$$\frac{P_{a|pq}}{P_{t|pq}} = k = \frac{\exp(-\mu \cdot c_{pqa})}{\exp(-\mu \cdot c_{pqt})}$$

Now suppose a new mode t' is added. Since the denominator term $\sum_n \exp(-\mu \cdot c_{pqn})$ doesn't affect $P_{a|pq} / P_{t|pq} = k$, then the ratio of the mode choices remains constant when the new mode is added. For example,

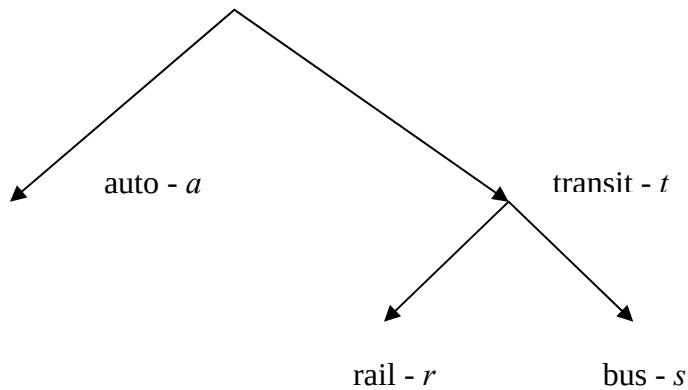
1. Before the new mode is added: $\begin{matrix} P_a + P_t = 1 \\ P_a / P_t = k \end{matrix}$, so, $P_t = \frac{1}{k+1}, P_a = \frac{k}{k+1}$.

2. After the new mode t' is added: $\begin{matrix} P_a + P_t + P_{t'} = 1 \\ P_a / P_t = k \end{matrix}$

Assume that mode t' has the same characteristics as mode t, so $P_t / P_{t'} = 1$, $P_t = P_{t'} = 1/(k+2)$ and $P_a = k/(k+2)$. Even though mode t' has the same characteristics as mode t, mode a loses a portion of its market to t', rather than the market of mode t being divided in half. For example, if $k = 3$, then $P_a = \frac{3}{4}$ and $P_t = \frac{1}{4}$ before mode t' is introduced. Afterwards, $P_a = \frac{3}{5}; P_t = P_{t'} = \frac{1}{5}$. Thus the market for t and t' is 40% compared to 25% for mode t alone. (We assume the service attributes of t and t' together did not change.) This property is often referred to as the "red bus/blue bus" problem since the color of a new mode should not affect mode choice.

4.5 Nested Logit Model and Composite Costs

One method for addressing the IIA property is to define a nested logit model. Consider two modes, a and t, and two submodes of mode t, r and s, as shown in the following figure.



Decision Tree for Two Modes and Two Submodes

Then,

$$P_{r|t} = \frac{\exp(-\theta \cdot c_r)}{\sum_{s \in R_t} \exp(-\theta \cdot c_s)}$$

where $P_{r|t}$ = conditional probability of using submode r

θ = dispersion parameter related to submode choice

R_t = set of all submodes of mode t

The *composite cost* of transit $\tilde{c}_t$ can then be defined by setting:

$$\exp(-\theta \cdot \tilde{c}_t) \equiv \sum_{s \in R_t} \exp(-\theta \cdot c_s)$$

Taking natural logs, we obtain $\tilde{c}_t = -\frac{1}{\theta} \ln\left(\sum_{s \in R_t} \exp(-\theta \cdot c_s) \right)$

$\tilde{c}_t$ is the expected (mean) cost experienced by an individual traveler drawn at random from the group using transit. Then, the nested mode choice model for submode r is:

$$P_{pqtr} = \left(\frac{\exp\left(-\mu \cdot \tilde{c}_{pqt}\right)}{\sum_{n} \exp\left(-\mu \cdot c_{pqn}\right)} \right) \cdot \left(\frac{\exp\left(-\theta \cdot c_{pqr}\right)}{\sum_{s \in R_t} \exp\left(-\theta \cdot c_{pqs}\right)} \right)$$

An alternative to using the nested formulation is a generalization of the dispersion (entropy) function, as follows:

$$S = -\sum_{pqm} d_{pqm} \ln\left(\frac{d_{pqm}}{Q_{pqm}} \right)$$

where $Q_{pqm} =$ prior probability of travel from zone *p* to zone *q* by mode *m*. The effect of adding the prior probability term may be determined as follows:

$$d_{pqm} = d_{pq} \left(\frac{Q_{pqm} \cdot \exp\left(-\mu \cdot c_{pqm}\right)}{\sum_{n} Q_{pqn} \cdot \exp\left(-\mu \cdot c_{pqn}\right)} \right)$$

Now, reconsider the addition of a new transit mode. Initially, $Q_{pqa} = Q_{pqt} = 0.5$, with the result that the prior probabilities cancel out. But suppose that $Q_{pqa} = 0.5$; $Q_{pqr} = Q_{pqs} = 0.25$. Then,

$$d_{pqa} = d_{pq} \left(\frac{0.5 \cdot \exp\left(-\mu \cdot c_{pqa}\right)}{0.5 \cdot \exp\left(-\mu \cdot c_{pqa}\right) + 0.25 \cdot \exp\left(-\mu \cdot c_{pqr}\right) + 0.25 \cdot \exp\left(-\mu \cdot c_{pqs}\right)} \right)$$

Now, if $c_{pqs} = c_{pqr}$,

$$d_{pqa} = d_{pq} \left(\frac{0.5 \cdot \exp\left(-\mu \cdot c_{pqa}\right)}{0.5 \cdot \exp\left(-\mu \cdot c_{pqa}\right) + 0.25 \cdot \exp\left(-\mu \cdot c_{pqr}\right) + 0.25 \cdot \exp\left(-\mu \cdot c_{pqr}\right)} \right)$$

$$= d_{pq} \left(\frac{0.5 \cdot \exp\left(-\mu \cdot c_{pqa}\right)}{0.5 \cdot \exp\left(-\mu \cdot c_{pqa}\right) + 0.5 \cdot \exp\left(-\mu \cdot c_{pqr}\right)} \right)$$

which is the intuitively expected result. Hence the “red bus-blue bus” problem can also be treated through the use of prior probabilities.

Exercises

1. Deterministic Mode Choice

The formulation of the *combined auto route choice and mode choice problem* takes the view that origin-destination flows by mode (d_{pqa}, d_{pqt}) are defined in persons per hour, but auto route choice flows (h_r) are defined in vehicles per hour, as shown in the following constraint:

$$\sum_{r \in R_{pqa}} h_r = d_{pqa} / \eta, \qquad p \in P; q \in Q$$

and $$f_a \equiv \sum_{pq} \sum_{r \in R_{pqa}} h_r \delta_r^a, \qquad a \in A$$

Clearly, this is one possibility, but another definition is equally plausible:

$$\sum_{r \in R_{pqa}} h_r = d_{pqa}, \qquad p \in P; q \in Q$$

and $$f_a \equiv \sum_{pq} \sum_{r \in R_{pqa}} h_r \delta_r^a / \eta, \qquad a \in A$$

where auto link flows are in vehicles per hour and auto route flows are in persons per hour.

Reformulate and solve the combined user-optimal auto route and mode choice problem stated in Section 4.1 with this second definition. Is the auto occupancy term (η) still needed in the objective function?

2. Regional Mode Choice Constraint

A desirable feature of a mode choice model is that it predicts the overall regional mode share correctly. That is, the ratio of the sum of the transit demand summed over all origin-destination pairs to total regional demand ought to equal an exogenous estimate of that demand derived from base year data.

$$\sum_{pq} d_{pqt} = M_t \cdot \sum_{pq} d_{pq}$$

where M_t = exogenous estimate of the regional transit share

To determine the implications of this requirement, add the above equation as a constraint to the combined auto route and mode choice problem stated in Section 4.3. What is the effect of the constraint on the modal cost structure?

Can you offer an interpretation of the new cost term in the model?

3. Mode Choice Parameter Estimation

I suggest you do this exercise in Excel; submit your plots in printed form.

Suppose a home-to-work trip by commuter railroad from the Chicago suburb of Evanston to the Chicago central area requires 60 generalized cost minutes; that is, the generalized cost in units of minutes, which consists of a weighted combination of travel time and fare, is c_{pqt} = 60. The generalized cost for the same trip by auto is c_{pqa} = 40 minutes.

Assuming that the mode choice is given by the following function, determine how the proportion of trips by auto (*a*) varies with the cost sensitivity parameter μ. That is, plot $P_{a|pq}$ vs. μ for values ranging from 0.0001 to 10, a possible range for μ. Use the log scale for plotting μ.

$$P_{a|pq} = \frac{\exp\left(-\mu \cdot c_{pqa}\right)}{\exp\left(-\mu \cdot c_{pqa}\right) + \exp\left(-\mu \cdot c_{pqt}\right)} = \frac{1}{1 + \exp\left(-\mu \cdot \left(c_{pqt} - c_{pqa}\right)\right)}$$

The Chicago area planning authority wants to introduce road prices that will cause more travelers to choose transit. The possibilities for tolls range from $1.00 to $4.00. Assuming that the travelers' mean value of time is $12.00 per hour, investigate what effect this range of tolls will have on their mode choice. To present your results, plot $P_{a|pq}$ vs. μ for each toll in increments of one dollar.

Explain what happens, according to the logit model, when the toll reaches the level that the generalized cost of travel by car exceeds that of transit.

5. Origin-Destination Choice (Trip Distribution)

The origin-destination choice (trip distribution) problem is to find the flow (persons/hour) from each origin zone to each destination zone, given the total flows leaving each origin and entering each destination, and the interzonal travel costs, as shown below.

Origin-Destination Flows

	1	2	.	.	q	.	.	Q	Total
1	d_{11}	d_{12}	…	…	d_{1q}	…	…	d_{1Q}	O_1
2	d_{21}	d_{22}	…	…	d_{2q}	…	…	d_{2Q}	O_2
.	…	…			…			…	…
.	…	…			…			…	…
p	d_{p1}	d_{p2}	…	…	d_{pq}	…	…	d_{pQ}	O_p
.	…	…			…			…	…
.	…	…			…			…	…
P	d_{P1}	d_{P2}	…	…	d_{Pq}	…	…	d_{PQ}	O_P
Total	D_1	D_2	…	…	D_q	…	…	D_Q	$d..$

Travel Costs

	1	2	.	.	q	.	.	Q
1	c_{11}	c_{12}	…	…	c_{1q}	…	…	c_{1Q}
2	c_{21}	c_{22}	…	…	c_{2q}	…	…	c_{2Q}
.	…	…			…			…
.	…	…			…			…
p	c_{p1}	c_{p2}	…	…	c_{pq}	…	…	c_{pQ}
.	…	…			…			…
.	…	…			…			…
P	c_{P1}	c_{P2}	…	…	c_{Pq}	…	…	c_{PQ}

where d_{pq} = flow from origin zone p to destination zone q (persons/hour)
c_{pq} = travel cost from origin zone p to destination zone q
O_p = flow departing from origin zone p
D_q = flow arriving at destination zone q
$d.. = \sum_p \sum_q d_{pq}$ = total flow departing from and arriving at all zones

The general form of the traditional model is $d_{pq} = A_p O_p B_q D_q f(c_{pq})$

where A_p is a factor related to origin zone p;

B_q is a factor related to destination zone q;

$f(c_{pq})$ is a non-increasing function of c_{pq} such as $\exp(-\beta \cdot c_{pq})$ or $(c_{pq})^{-\alpha}$.

Originally, such models were highly empirical and based on an analogy with the gravity law describing the attractive force between two masses, hence, the commonly used name, *gravity model*. The name comes from the fact that nearby zone pairs have higher flows than more separated zones. In the mid 1960's, several authors showed that such a functional form could be derived using statistical methods without reference to gravitational concepts from physics.

5.1 Model Derivation

Our derivation extends the fixed cost derivation of the mode choice model by relaxing the constraint that d_{pq} is known. The focus of this subsection is on the calculation of A_p and B_q. First, replace the constraint, $\sum_m d_{pqm} = d_{pq}, \quad p \in P; q \in Q$ with constraints defined on the origin and destination flows, O_p and D_q. The dispersion constraint is also generalized implicitly to describe dispersion to higher cost destinations as well as modes.

$$
\begin{aligned}
&\min_{(\mathbf{d})} z(\mathbf{d}) = \sum_{pqm} c_{pqm} \cdot d_{pqm} \\
&st: \quad \sum_q \sum_m d_{pqm} = O_p, \qquad p \in P; \\
&\qquad \sum_p \sum_m d_{pqm} = D_q, \qquad q \in Q \\
&\qquad -\sum_{pq} \sum_m d_{pqm} \ln d_{pqm} \geq S_0 \\
&\qquad d_{pqm} \geq 0, \qquad m \in M, p \in P; q \in Q
\end{aligned}
$$

For $d_{pqm} > 0,$ that is for positive flows between all origin-destination pairs and for all *available* modes, the solution is $d_{pqm} = \exp\left(\mu\left(\kappa'_p + \lambda'_q - c_{pqm}\right) - 1\right) \equiv \exp\left(\kappa_p - 1 + \lambda_q - \mu \cdot c_{pqm}\right)$ where κ'_p and λ'_q are Lagrange multipliers associated with the constraints on origin and destination flows, and μ is associated with the dispersion constraint, as in the mode choice model. We can express these parameters in terms of the flows as follows:

$$
\begin{aligned}
\sum_q \sum_m d_{pqm} = O_p &= \sum_q \sum_m \exp\left(\kappa_p - 1 + \lambda_q - \mu \cdot c_{pqm}\right) = \exp\left(\kappa_p - 1\right) \sum_q \left(\exp\left(\lambda_q\right) \sum_m \exp\left(-\mu \cdot c_{pqm}\right)\right) \\
\sum_p \sum_m d_{pqm} = D_q &= \sum_p \sum_m \exp\left(\kappa_p - 1 + \lambda_q - \mu \cdot c_{pqm}\right) = \exp\left(\lambda_q\right) \sum_p \left(\exp\left(\kappa_p - 1\right) \sum_m \exp\left(-\mu \cdot c_{pqm}\right)\right)
\end{aligned}
$$

Hence,

$$
\exp\left(\kappa_p - 1\right) = \frac{O_p}{\sum_q \left(\exp\left(\lambda_q\right) \sum_m \exp\left(-\mu \cdot c_{pqm}\right)\right)}; \quad \exp\left(\lambda_q\right) = \frac{D_q}{\sum_p \left(\exp\left(\kappa_p - 1\right) \sum_m \exp\left(-\mu \cdot c_{pqm}\right)\right)}.
$$

Substituting into the original expression, we obtain:

$$d_{pqm} = \frac{O_p D_q \exp\left(-\mu \cdot c_{pqm}\right)}{\sum_q \left(\exp\left(\lambda_q\right) \cdot \sum_m \exp\left(-\mu \cdot c_{pqm}\right)\right) \cdot \sum_p \left(\exp\left(\kappa_p - 1\right) \cdot \sum_m \exp\left(-\mu \cdot c_{pqm}\right)\right)}$$

Hence, we have not eliminated κ_p and λ_q, but only rearranged them, which must be true because they are functions of each other. But, we can see that we can define two sets of *balancing factors*:

$$A_p = \frac{1}{\sum_q \left(\exp\left(\lambda_q\right) \sum_m \exp\left(-\mu \cdot c_{pqm}\right)\right)}, p \in P;\ B_q = \frac{1}{\sum_p \left(\exp\left(\kappa_p - 1\right) \sum_m \exp\left(-\mu \cdot c_{pqm}\right)\right)}, q \in Q.$$

By substitution,

$$d_{pqm} = A_p O_p B_q D_q \exp\left(-\mu \cdot c_{pqm}\right)$$

Thus, the flow from zone p to zone q on mode m is directly proportional to the flow leaving zone p, O_p, the flow entering zone q, D_q, and a deterrence function defined on c_{pqm}, as well as balancing factors A_p and B_q. Moreover, A_p and B_q are functions of all destination flows and all origin flows, respectively, since

$$A_p = \frac{1}{\sum_q \left(B_q D_q \sum_m \exp\left(-\mu \cdot c_{pqm}\right)\right)}, p \in P;\ B_q = \frac{1}{\sum_p \left(A_p O_p \sum_m \exp\left(-\mu \cdot c_{pqm}\right)\right)}, q \in Q.$$

Moreover, by definition,

$$d_{pq} = \sum_m d_{pqm} = A_p O_p B_q D_q \sum_m \exp\left(-\mu \cdot c_{pqm}\right)$$

Define $\tilde{c}_{pq} = -\frac{1}{\mu} \ln\left(\sum_m \exp\left(-\mu \cdot c_{pqm}\right)\right)$. Then $d_{pq} = A_p O_p B_q D_q \exp\left(-\mu \cdot \tilde{c}_{pq}\right)$

More generally, this model may be extended to:

$$d_{pqm} = A_p O_p B_q D_q \exp\left(-\beta \cdot \tilde{c}_{pq}\right) \cdot \frac{\exp\left(-\mu \cdot c_{pqm}\right)}{\sum_n \exp\left(-\mu \cdot c_{pqn}\right)}$$

so that the mode choice is determined jointly with the origin-destination choice.

5.2 Computation of A_p and B_q

The following method is commonly used to compute A_p and B_q. The method is known to converge under very general conditions.

Step 0. set $A_p^1 = 1,\ p = 1,...P$; then $d_{pq}^1 = O_p B_q^1 D_q \exp\left(-\beta \cdot \tilde{c}_{pq}\right)$

Step 1. solve for $B_q^1,\ q = 1,...Q$:

$$\sum_p d_{pq} = D_q = B_q^1 D_q \sum_p O_p \exp\left(-\beta \cdot \tilde{c}_{pq}\right) \Rightarrow B_q^1 = \frac{1}{\sum_p O_p \exp\left(-\beta \cdot \tilde{c}_{pq}\right)}$$

Step 2. next solve for $A_p^2,\ p = 1,...P$:

$$\sum_q d_{pq} = O_p = A_p^2 O_p \sum_q B_q^1 D_q \exp\left(-\beta \cdot \tilde{c}_{pq}\right) \Rightarrow A_p^2 = \frac{1}{\sum_q B_q^1 D_q \exp\left(-\beta \cdot \tilde{c}_{pq}\right)}$$

Step 3. next solve for $B_q^2,\ q = 1,...Q$:

$$\Rightarrow B_q^2 = \frac{1}{\sum_p A_p^2 O_p \exp\left(-\beta \cdot \tilde{c}_{pq}\right)}$$

In general,

$$A_p^k = \frac{1}{\sum_q B_q^{k-1} D_q \exp\left(-\beta \cdot \tilde{c}_{pq}\right)},\ p = 1,...P;\ B_q^k = \frac{1}{\sum_p A_p^k O_p \exp\left(-\beta \cdot \tilde{c}_{pq}\right)},\ q = 1,...Q$$

Continue until

$$\frac{\left(A_p^k - A_p^{k-1}\right)}{A_p^k} < \varepsilon, \text{all } p; \qquad \frac{\left(B_q^k - B_q^{k-1}\right)}{B_q^k} < \varepsilon, \text{all } q.$$

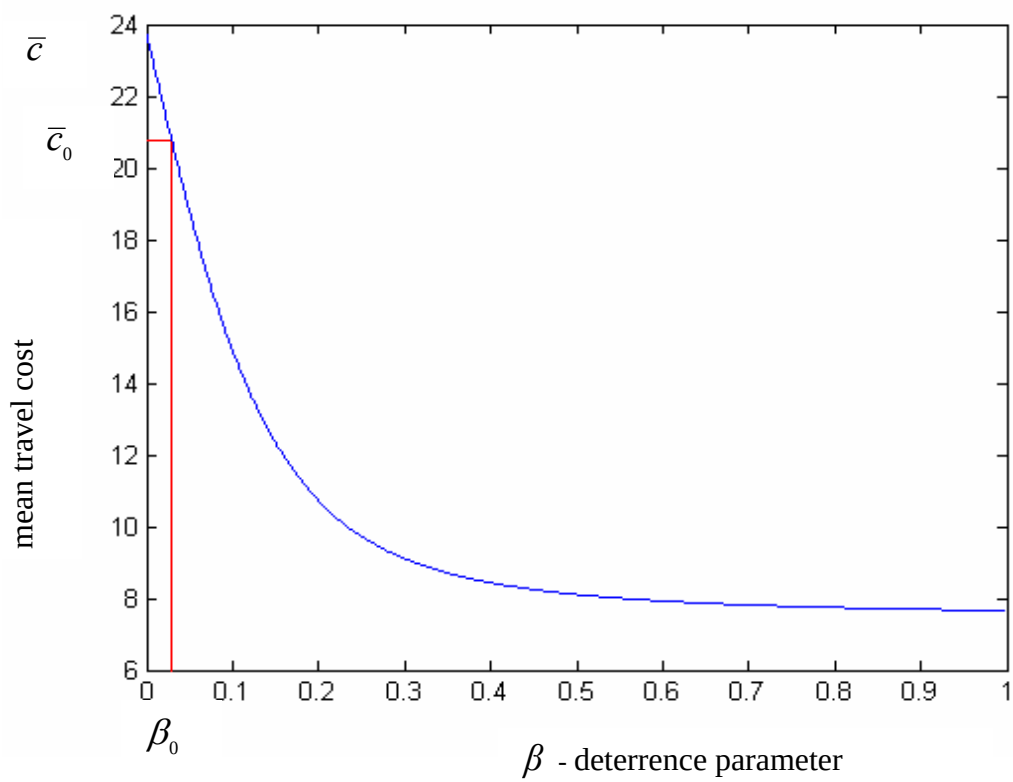
For 2000 zones and $\varepsilon = 10^{-6}$, the values of the $P + Q$ parameters are obtained in about 100 iterations. If the procedure does not converge, usually it means that the value of β is too large. In addition, for convergence it must be true that $\sum_p O_p = \sum_q D_q = d..$, the total flow in the region.

The use of origin and destination constraints is a rigorous and operational method for imposing assumed origin and destination flows on the model, or on any O-D flow matrix.

5.3 Calibration of Travel Deterrence Parameters

A remaining problem is to determine the appropriate value of the parameter β. While this parameter is related to the dispersion constraint, it is better chosen with respect to the mean cost, since this value $\bar{c}_0$ can be estimated from survey data: $\bar{c} = \frac{1}{d_{..}} \sum_{pq} d_{pq} \cdot \tilde{c}_{pq}$. By substituting values of β into the model and solving it, a curve of the shape shown below can be determined.

The figure shown above is based on the Sioux Falls trip matrix of LeBlanc et al (1975). The zone-to-zone travel costs used in preparing the figure are from the user-optimal assignment of LeBlanc's trip matrix to the Sioux Falls network. Using these travel costs, and the row and column sums of his trip matrix, the trip distribution model was solved for values of the parameter β, the mean costs computed, and plotted. The value of the mean travel cost for LeBlanc's matrix is 20.7, which has a corresponding value of β = 0.029. The value of the travel cost for β = 0.0 is 23.7. This is the travel cost that corresponds to the trip distribution function, $d_{pq} = \frac{O_p \cdot D_q}{d_{..}}$. From these results, we can see that interzonal flows in LeBlanc's matrix are close to being proportional to the row and column sums, which is not very representative of observed travel.



Mean Travel Cost vs. Deterrence Parameter

If an observed value of $\bar{c}_0$ is available from survey or census data, and the zone-to-zone travel costs $(\tilde{c}_{pq})$ are known, then the value of β_0 can be found by trial and error or by an iterative technique. Since the value of β also depends on the scale of $\bar{c}$ (minutes vs. hours, for example), a reasonable first guess is $\beta_1 = 1/\bar{c}_0$. Once $\bar{c}_1(\beta_1)$ is determined by solving the trip distribution model for β_1, then an iterative scheme is given by $\beta_n = \beta_{n-1} \cdot \bar{c}_{n-1}(\beta_{n-1})/\bar{c}_0$. Iterate until $\bar{c}_n(\beta_n) \approx \bar{c}_0$.

Since β depends on the mean observed generalized cost, it is important to know how travel times and distances are changing over time. Analysis of survey and census data for a sequence of years can be helpful in understanding how travel times and costs have changed, or not changed, whichever is the case.

5.4 Form of the Deterrence Function

Until now, only the negative exponential form of the deterrence function has been considered. This function follows from the entropy constraint, and is consistent with the logit function. Other functional forms, however, have been considered in the literature. See Erlander and Stewart (1990) for a definitive review and derivation of alternatives.

A simple way to derive one alternative deterrence function is to assume that origin-destination flows are determined by the natural logarithm of travel costs, rather than the travel costs directly. The travel behavior literature does offer some basis for such an assumption, as is discussed further below.

$$\begin{aligned}
&\min_{(\mathbf{d})} z(\mathbf{d}) = \sum_{pqm} \ln(c_{pqm}) \cdot d_{pqm} \\
&st: \quad \sum_q \sum_m d_{pqm} = O_p, \qquad p \in P \\
&\qquad \sum_p \sum_m d_{pqm} = D_q, \qquad q \in Q \\
&\qquad -\sum_{pq} \sum_m d_{pqm} \ln d_{pqm} \geq S_0 \\
&\qquad d_{pqm} \geq 0, \qquad m \in M, p \in P; q \in Q
\end{aligned}$$

In this case, the optimality condition becomes

$$d_{pqm} = \exp(\kappa_p + \lambda_q - \alpha \ln(c_{pqm})) = \exp(\kappa_p + \lambda_q) \cdot (c_{pqm})^{-\alpha}$$

This power function has a somewhat different shape than the negative exponential function, especially for small values of travel cost; of course, the function also approaches infinity as travel cost approaches zero.

These two functions may be combined to derive the Tanner (1961) function, which is noted for its flexibility.

$$\min_{(\mathbf{d})} z(\mathbf{d}) = \sum_{pqm} c_{pqm} \cdot d_{pqm} + \alpha \cdot \ln(c_{pqm}) \cdot d_{pqm}$$

$$\begin{aligned} st: \quad & \sum_q \sum_m d_{pqm} = O_p, \qquad p \in P \\ & \sum_p \sum_m d_{pqm} = D_q, \qquad q \in Q \\ & -\sum_{pq} \sum_m d_{pqm} \ln d_{pqm} \geq S_0 \\ & d_{pqm} \geq 0, \qquad m \in M, p \in P; q \in Q \end{aligned}$$

In this case, the resulting function is

$$d_{pqm} = \exp(\kappa_p + \lambda_q - \beta \cdot c_{pqm} - \alpha \cdot \ln(c_{pqm})) = \exp(\kappa_p + \lambda_q) \cdot \exp(-\beta \cdot c_{pqm}) \cdot (c_{pqm})^{-\alpha}$$

The Tanner function is the most general considered here, in that it includes the exponential and power functions as special cases. A drawback of the Tanner function is that there is no known way to derive it in the user-optimal framework. As applied in practice, the values of β and α have opposite signs. Thus, if the values of β and α are positive, the functions are either $d_{pqm} = \exp(\kappa_p + \lambda_q) \cdot \exp(-\beta \cdot c_{pqm}) \cdot (c_{pqm})^{+\alpha}$ or $d_{pqm} = \exp(\kappa_p + \lambda_q) \cdot \exp(+\beta \cdot c_{pqm}) \cdot (c_{pqm})^{-\alpha}$

Finally, Professor Dieter Lohse (2002) has proposed the EVA function:

$$d_{pqm} = (1 + c_{pqm})^{\left(-\gamma / \left(1 + \exp(\delta - \beta \cdot c_{pqm})\right)\right)}$$

He suggests this function leads to better fits to German data.

The properties of these four functions can be explored by computing their *elasticities*. The elasticity of a demand function is the proportional change in demand relative to a given proportional change in cost. That is,

$$e = \frac{\frac{\partial d}{d}}{\frac{\partial c}{c}} = \frac{\frac{\partial d}{\partial c}}{\frac{d}{c}}$$

Applying this procedure, we obtain the following results.

Name	Cost function	Elasticity
Negative exponential:	$d = \exp(-\beta \cdot c)$	$-\beta \cdot c$
Power:	$d = c^{-\alpha}$	$-\alpha$
Tanner:	$d = \exp(-\beta \cdot c) \cdot c^{-\alpha}$	$-\alpha - \beta \cdot c$
EVA	$d = (1+c)^{(-\gamma/(1+\exp(\delta-\beta\cdot c)))}$	$\frac{-\gamma \cdot c}{1+\exp(\delta-\beta\cdot c)} \cdot \left[\left(\frac{1}{1+c}\right) + \ln(1+c) \cdot \frac{\beta \cdot \exp(\delta-\beta\cdot c)}{1+\exp(\delta-\beta\cdot c)}\right]$

At the end of this section, these functions and their elasticities are plotted.

5.5 Extension to Tours

The above traditional treatment of origin-destination flows considers that travelers move from one origin to one destination. Actually, travel frequently consists of visits to a sequence of activities, or a tour. In this section results pertaining to representation of such tours are presented as an example of how to apply this approach to create new models. To provide an idea of the frequency of such tours for travel that begins at home in the morning peak period, consider the following table compiled from a household survey for the Tucson, Arizona region in 2003 by Ms. Yuhwa Lee, a transportation engineering graduate student at the University of Arizona.

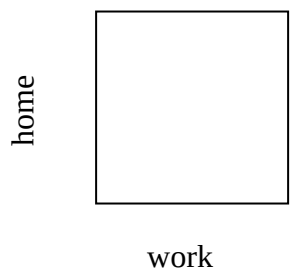
Origin	1st destination	2nd destination	Number
Home	Work	--	734
Home	School	--	655
Home	Serve passenger	--	120
Home	Recreation	--	72
Home	Shopping, etc.	--	82
Home	Serve passenger	Work	144
Home	Eat meal	Work	43
Home	Shopping	Work	34
Home	Other	Work	26
Home	Serve passenger	School	49
Home	Other	School	35
Home	Other longer tours		210
Home	All destinations		2204

We consider only the case of fixed travel costs; however, the model can be generalized to the case of endogenous travel costs. Also, to simplify the presentation, only one mode is considered. The basic idea of the model is to generalize the concept of a 2-way trip table shown in section 5.1 to an n-way table. Thus travel from home to an intermediate stop, followed by a travel to work or school can be represented as a 3-way table. In keeping with standard trip generation analysis, we assume that the number of trips per time period is forecast at the home origin by tour type. Typical morning tours are:

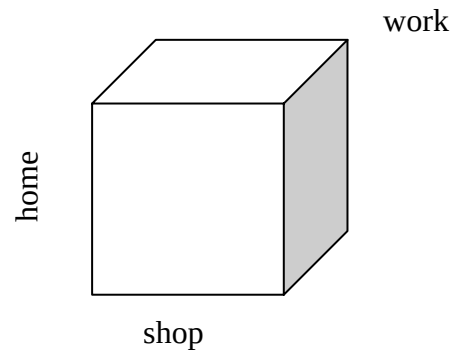
1. Home – work
2. Home – recreation – work
3. Home – serve_passenger – work
4. Home – eat_meal – school
5. Home – shop-personal_business – recreation

The concept of the 2-way vs. 3-way tables may be visualized as follows:

Two-way Flows:
home-work



Three-way Flows:
home-shop-work



We can formulate the three-way problem as follows for home-shop-work (*hsw*) travel:

$$\min_{(\mathbf{d})} \sum_{pq} c_{pq} d_{pq*}^{hsw} + \sum_{qr} c_{qr} d_{*qr}^{hsw} + \frac{1}{\alpha^{hs}} \sum_{pq} d_{pq*}^{hsw} \left(\ln d_{pq*}^{hsw} - 1\right) + \frac{1}{\alpha^{sw}} \sum_{qr} d_{*qr}^{hsw} \left(\ln d_{*qr}^{hsw} - 1\right)$$

multiplier

$$\text{st}: \sum_{q} d_{pq*}^{hsw} = O_p^{hsw}, \quad \forall p \qquad \kappa_p^{hsw}$$

$$\sum_{p} d_{pq*}^{hsw} = D_q^{s}, \quad \forall q \qquad \lambda_q^{s}$$

$$\sum_{r} d_{*qr}^{hsw} = O_q^{s} = D_q^{s}, \quad \forall q \qquad \eta_q^{s}$$

$$\sum_{q} d_{*qr}^{hsw} = D_r^{w}, \quad \forall r \qquad \rho_r^{w}$$

Note that the first two terms are equivalent to the total travel cost, $\sum_{pqr} c_{pqr} \cdot d_{pqr}^{hsw}$. Moreover, d_{pq*}^{hsw} denotes the *hsw* flows from zone *p* to zone *q*, which pertains to home-shop travel. Similarly, d_{*qr}^{hsw} pertains to the same flows from zone *q* to zone *r*. Thus zone *p* is associated with the home activity, zone *q* with the shop activity and zone *r* with the work activity.

Dispersion terms are included for each activity pair; origin and destination constraints govern all tours departing from residences in zone *p*, arriving at zone *q* for shopping, departing from shopping activities in zone *q*, and arriving at zone *r* for work. The optimality conditions are as follows:

$$\frac{\partial L}{\partial d_{pq*}^{hsw}} = c_{pq} + \frac{1}{\alpha^{hs}} \ln d_{pq*}^{hsw} - \kappa_p^{hsw} - \lambda_q^s = 0$$

$$\frac{\partial L}{\partial d_{*qr}^{hsw}} = c_{qr} + \frac{1}{\alpha^{sw}} \ln d_{*qr}^{hsw} - \eta_q^s - \rho_r^w = 0$$

$$d_{pq*}^{hsw} = \exp\left(\alpha^{hs}\left(\kappa_p^{hsw} + \lambda_q^s\right)\right) \cdot \exp\left(-\alpha^{hs} c_{pq}\right) = A_p^{hsw} \cdot O_p^{hsw} \cdot B_q^s \cdot D_q^s \cdot \exp\left(-\alpha^{hs} c_{pq}\right)$$

$$d_{*qr}^{hsw} = \exp\left(\alpha^{sw}\left(\eta_q^s + \rho_r^w\right)\right) \cdot \exp\left(-\alpha^{sw} c_{qr}\right) = A_q^s \cdot O_q^s \cdot B_r^w \cdot D_r^w \cdot \exp\left(-\alpha^{sw} c_{qr}\right)$$

Hence, we obtain a pair of origin-destination choice models for the tour *hsw* with cost sensitivity parameters related to home- shop and shop-work. Next, consider a formulation with two tours, *hw* and *hsw* with a common work destination constraint.

$$\min_{(\mathbf{d})} \sum_{pr} c_{pr} d_{pr}^{hw} + \frac{1}{\alpha^{hw}} \sum_{pr} d_{pr}^{hw}\left(\ln d_{pr}^{hw} - 1\right)$$

$$+ \sum_{pq} c_{pq} d_{pq*}^{hsw} + \sum_{qr} c_{qr} d_{*qr}^{hsw} + \frac{1}{\alpha^{hs}} \sum_{pq} d_{pq*}^{hsw}\left(\ln d_{pq*}^{hsw} - 1\right) + \frac{1}{\alpha^{sw}} \sum_{qr} d_{*qr}^{hsw}\left(\ln d_{*qr}^{hsw} - 1\right)$$

$$\text{st}: \sum_r d_{pr}^{hw} = O_p^{hw}, \quad \forall p \qquad \kappa_p^{hw}$$

$$\sum_q d_{pq*}^{hsw} = O_p^{hsw}, \quad \forall p \qquad \kappa_p^{hsw}$$

$$\sum_p d_{pq*}^{hsw} = D_q^s, \quad \forall q \qquad \lambda_q^s$$

$$\sum_r d_{*qr}^{hsw} = O_q^s, \quad \forall q \qquad \eta_q^s$$

$$\sum_p d_{pr}^{hw} + \sum_q d_{*qr}^{hsw} = D_r^w, \quad \forall r \qquad \rho_r^w$$

The first two terms of the objective function pertain to *hw* tours, while the last four terms pertain to *hsw* tours. Note that the last constraint at the workplace destination pertains to both tours. The optimality conditions for this formulation are as follows.

$$\frac{\partial L}{\partial d_{pr}^{hw}} = c_{pr} + \frac{1}{\alpha^{hw}} \ln d_{pr}^{hw} - \kappa_p^{hw} - \rho_r^w = 0$$

$$\frac{\partial L}{\partial d_{pq*}^{hsw}} = c_{pq} + \frac{1}{\alpha^{hs}} \ln d_{pq*}^{hsw} - \kappa_p^{hsw} - \lambda_q^s = 0$$

$$\frac{\partial L}{\partial d_{*qr}^{hsw}} = c_{qr} + \frac{1}{\alpha^{sw}} \ln d_{*qr}^{hsw} - \eta_q^s - \rho_r^w = 0$$

$$d_{pr}^{hw} = \exp\left(\alpha^{hw}\left(\kappa_p^{hw} + \rho_r^w\right)\right) \cdot \exp\left(-\alpha^{hw} c_{pr}\right)$$

$$d_{pq*}^{hsw} = \exp\left(\alpha^{hs}\left(\kappa_p^{hsw} + \lambda_q^s\right)\right) \cdot \exp\left(-\alpha^{hs} c_{pq}\right)$$

$$d_{*qr}^{hsw} = \exp\left(\alpha^{sw}\left(\eta_q^s + \rho_r^w\right)\right) \cdot \exp\left(-\alpha^{sw} c_{qr}\right)$$

The solution of these conditions is shown next.

$$\exp\left(\alpha^{hw} \cdot \kappa_p^{hw}\right) = O_p^{hw} / \sum_r \left(\exp\left(\alpha^{hw} \cdot \rho_r^w\right) \cdot \exp\left(-\alpha^{hw} c_{pr}\right)\right) \equiv A_p^{hw} \cdot O_p^{hw}$$

$$\exp\left(\alpha^{hs} \cdot \kappa_p^{hsw}\right) = O_p^{hsw} / \sum_q \left(\exp\left(\alpha^{hs} \cdot \lambda_q^s\right) \cdot \exp\left(-\alpha^{hs} c_{pq}\right)\right) \equiv A_p^{hsw} \cdot O_p^{hsw}$$

$$\exp\left(\alpha^{sw} \cdot \eta_q^s\right) = O_q^s / \sum_r \left(\exp\left(\alpha^{sw} \cdot \rho_r^w\right) \cdot \exp\left(-\alpha^{sw} c_{qr}\right)\right) \equiv A_q^s \cdot O_q^s$$

$$\exp\left(\alpha^{hs} \cdot \lambda_q^s\right) = D_q^s / \sum_p \exp\left(\alpha^{hs} \cdot \kappa_p^{hsw}\right) \cdot \exp\left(-\alpha^{hs} c_{pq}\right) \equiv B_q^s \cdot D_q^s$$

The constraint involving both tours requires some additional manipulation:

$$\begin{aligned} D_r^w &= \sum_p d_{pr}^{hw} + \sum_q d_{*qr}^{hsw} \\ &= \exp\left(\alpha^{hw} \cdot \rho_r^w\right) \cdot \sum_p \left(\exp\left(\alpha^{hw} \cdot \kappa_p^{hw}\right) \cdot \exp\left(-\alpha^{hw} c_{pr}\right)\right) \\ &+ \exp\left(\alpha^{sw} \cdot \rho_r^w\right) \cdot \sum_q \left(\exp\left(\alpha^{sw} \cdot \eta_q^s\right) \cdot \exp\left(-\alpha^{sw} c_{qr}\right)\right) \end{aligned}$$

We can include $\exp\left(\alpha^{hw} \cdot \rho_r^w\right)$ in the terms related to both tours as follows:

$$\begin{aligned} D_r^w &= \exp\left(\alpha^{hw} \cdot \rho_r^w\right) \cdot \left[\sum_p \left(\exp\left(\alpha^{hw} \cdot \kappa_p^{hw}\right) \cdot \exp\left(-\alpha^{hw} c_{pr}\right)\right)\right] \\ &+ \exp\left(\alpha^{hw} \cdot \rho_r^w\right) \cdot \left[\exp\left(\left(\alpha^{sw} - \alpha^{hw}\right) \cdot \rho_r^w\right) \cdot \sum_q \left(\exp\left(\alpha^{sw} \cdot \eta_q^s\right) \cdot \exp\left(-\alpha^{sw} c_{qr}\right)\right)\right] \end{aligned}$$

Then, we can solve for $\exp\left(\alpha^{hw} \cdot \rho_r^w\right)$ and define a balancing factor B_r^w for the terms in brackets:

$$\exp\left(\alpha^{hw}\cdot\rho_r^w\right)\equiv B_r^w\cdot D_r^w =$$
$$D_r^w \Big/ \left[\sum_p\left(\exp\left(\alpha^{hw}\cdot\kappa_p^{hw}\right)\cdot\exp\left(-\alpha^{hw}c_{pr}\right)\right)+\exp\left(\left(\alpha^{sw}-\alpha^{hw}\right)\cdot\rho_r^w\right)\cdot\sum_q\left(\exp\left(\alpha^{sw}\cdot\eta_q^s\right)\cdot\exp\left(-\alpha^{sw}c_{qr}\right)\right)\right]$$

Hence, the balancing factor B_r^w is a function of terms related to both tours.

Thus, using the same concepts as for 2-way tables, tours can be modeled as an n-way table, and various tables can have common origin and destination totals. Standard balancing factor methods appear to be applicable; however, this aspect needs further evaluation computationally. The next step in the development of this idea is to create a general formulation that will encompass many tours of various activities in a more automated manner.

Finally, let's check that these models can be derived with flow dependent travel costs. To this end, replace the two fixed costs terms in the formulation at the beginning of this section with the following expression, and add two sets of constraints:

$$\min_{(\mathbf{h},\mathbf{d})}\sum_a\int_0^{f_a}c_a(x)dx+\cdots \qquad \text{multiplier}$$
$$\text{st}:\sum_r\sum_{t\in R_{pqr}}h_{pqrt}^{hsw}=d_{pq*}^{hsw} \qquad \forall pq \qquad u_{pq}^1$$
$$\sum_q\sum_{t\in R_{pqr}}h_{pqrt}^{hsw}=d_{*qr}^{hsw} \qquad \forall qr \qquad u_{qr}^2$$
$$\text{where } f_a\equiv\sum_{pqrt}h_{pqrt}^{hsw}\cdot\delta_t^a$$

Taking partial derivatives of the Lagrangian function with respect to h_{pqrt}^{hsw}, we obtain:

$$\sum_a c_a\left(f_a\right)\cdot\delta_t^a-u_{pq}^1-u_{qr}^2\geq 0$$
$$h_{pqrt}^{hsw}\left(C_t-u_{pq}^1-u_{qr}^2\right)=0$$

Therefore, if $h_{pqrt}^{hsw}>0,\ C_t=u_{pq}^1+u_{qr}^2,$ which is equivalent to the fixed cost from *p* to *q* to *r*. The remainder of the analysis follows as above.

5.6 Extension to Time Intervals

Finally, we consider the choice of time interval of travel. The term *time interval* is used deliberately, since these are not really models of departure time in the dynamic sense. A preliminary question concerns the length of the time interval. The flow properties of this class of models are that flows exist from origin to destination for the duration of the period of interest. There is no notion of flow propagation, or movement of flow from origin to destination in the model. Some authors say that flows propagate instantaneously to make this point.

For this reason, any time interval within the period being modeled should be long enough to be sensible. Origin-destination travel times for large regions are as long as two hours, which is as long as a typical time period of interest. However, most flows have travel times less than 30 minutes, and this is considered to be a reasonable minimum value from a travel choice viewpoint. From an applications viewpoint, forecasts are needed for periods as short as 30 minutes. For example, in Singapore road tolls are changed at 30 minute intervals, and forecasts concerning their effect on travel choices are desired. Therefore, I consider that 30 minutes is a reasonable target for such a model. As a point of departure, I consider a single tour model with a time interval choice function which incorporates information concerning travel preferences among time intervals.

$$\min_{(\mathbf{d})} \sum_{pqt} c_{pqt} \cdot d_{pqt} + \frac{1}{\alpha} \sum_{pq} d_{pq*} \left(\ln d_{pq*} - 1 \right) + \frac{1}{\theta} \sum_{pq} \sum_{t} d_{pqt} \left(\ln \left(\frac{d_{pqt}}{d_{pq*}} \right) - 1 \right)$$

$$\begin{aligned}
st: \quad & \sum_{q} d_{pq*} = O_p, \quad \forall p && \kappa_p \\
& \sum_{p} d_{pq*} = D_q, \quad \forall q && \lambda_q \\
& \sum_{pq} d_{pqt} = p_t \cdot d_{...}, \quad \forall t && \beta_t \\
& \sum_{t} d_{pqt} = d_{pq*}, \quad \forall pq && \varepsilon_{pq}
\end{aligned}$$

The flow variable is now expanded to d_{pqt}, the flow from origin *p* to destination *q* in time interval *t*. The summation of these flow over all intervals is denoted as d_{pq*}. Two dispersion constraints are defined, one for O-D flows and another for flows within time intervals, with d_{pq*} as a prior. The optimality conditions related to d_{pqt} are:

$$\begin{aligned}
& \frac{\partial L}{\partial d_{pqt}} = c_{pqt} + \frac{1}{\theta} \cdot \ln \left(\frac{d_{pqt}}{d_{pq*}} \right) - \beta_t - \varepsilon_{pq} = 0 \\
& d_{pqt} = d_{pq*} \cdot \exp\left(\theta \varepsilon_{pq} \right) \cdot \exp\left(-\theta \left(c_{pqt} + \beta_t \right) \right) \\
& \exp\left(-\theta \varepsilon_{pq} \right) = \sum_{t} \exp\left(-\theta \left(c_{pqt} + \beta_t \right) \right) \equiv \exp\left(-\theta \cdot \tilde{c}_{pq} \right) \\
& d_{pqt} = d_{pq*} \cdot \frac{\exp\left(-\theta \left(c_{pqt} + \beta_t \right) \right)}{\sum_{t} \exp\left(-\theta \left(c_{pqt} + \beta_t \right) \right)}
\end{aligned}$$

The conditions related to d_{pq*} are:

$$\frac{\partial L}{\partial d_{pq*}} = \frac{1}{\alpha} \cdot \ln d_{pq*} - \frac{1}{\theta} - \kappa_p - \lambda_q + \varepsilon_{pq} = 0$$

$$d_{pq*} = \exp\left(\alpha \cdot \left(\kappa_p + \lambda_q + \varepsilon_{pq} - \frac{1}{\theta} \right) \right) = A_p O_p B_q D_q \exp\left(-\alpha \cdot \tilde{c}_{pq} \right)$$

The final result is:

$$d_{pqt} = A_p O_p B_q D_q \exp\left(-\alpha \cdot \tilde{c}_{pq} \right) \cdot \frac{\exp\left(-\theta \left(c_{pqt} + \beta_t \right) \right)}{\sum_t \exp\left(-\theta \left(c_{pqt} + \beta_t \right) \right)}$$

where $\tilde{c}_{pq} = -\frac{1}{\theta} \cdot \ln\left[\sum_t \exp\left(-\theta \left(c_{pqt} + \beta_t \right) \right) \right]$

The role of the parameter β_t is to represent the "bias" related to time interval preferences. These are similar to the transit bias parameter in the mode choice model. A related idea is to impose constraints related to when travelers must arrive at their destinations, such as the workplace. An example of such a constraint is $\sum_t d_{pqt} = \overline{p}_{qt} D_q$, where $\overline{p}_{qt}$ is the proportion of travelers who must arrive by the end of interval *t*. Substituting this set of constraints for the period bias constraint leads to a slightly different result:

$$d_{pqt} = A_p O_p B_q D_q \exp\left(-\alpha \cdot \tilde{c}_{pq} \right) \cdot \frac{\exp\left(-\theta \left(c_{pqt} + \beta_{qt} \right) \right)}{\sum_t \exp\left(-\theta \left(c_{pqt} + \beta_{qt} \right) \right)}$$

In this case the parameters β_{qt} are period and destination specific. A system of such constraints could represent various constraints faced by travelers during the morning peak period with regard to obligations at home, school, meetings and work. The determination of the parameter values requires further investigation.

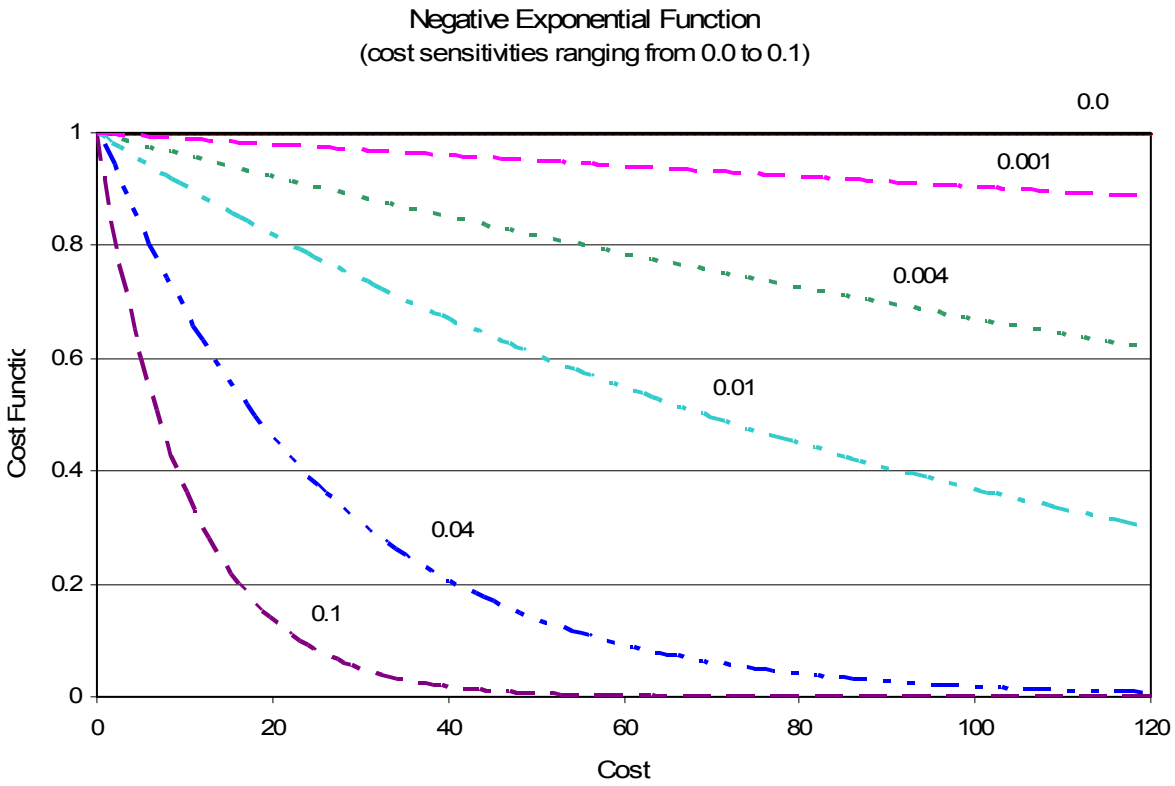
References

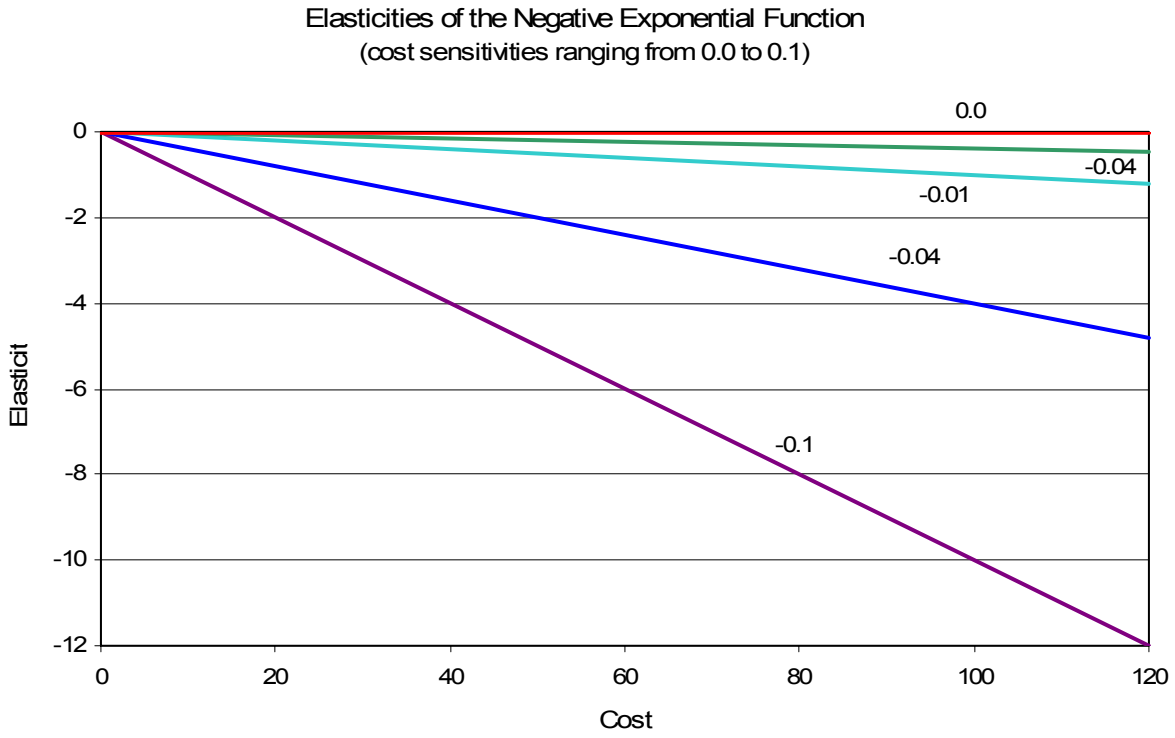
Erlander, S. and N. F. Stewart (1990) *The Gravity Model in Transportation Analysis*, VSP, Utrecht.

LeBlanc, L. J., E. K. Morlok and W. P. Pierskalla (1975) An Efficient Approach to Solving the Road Network Equilibrium Assignment Problem, *Transportation Research* **9**, 309-318.

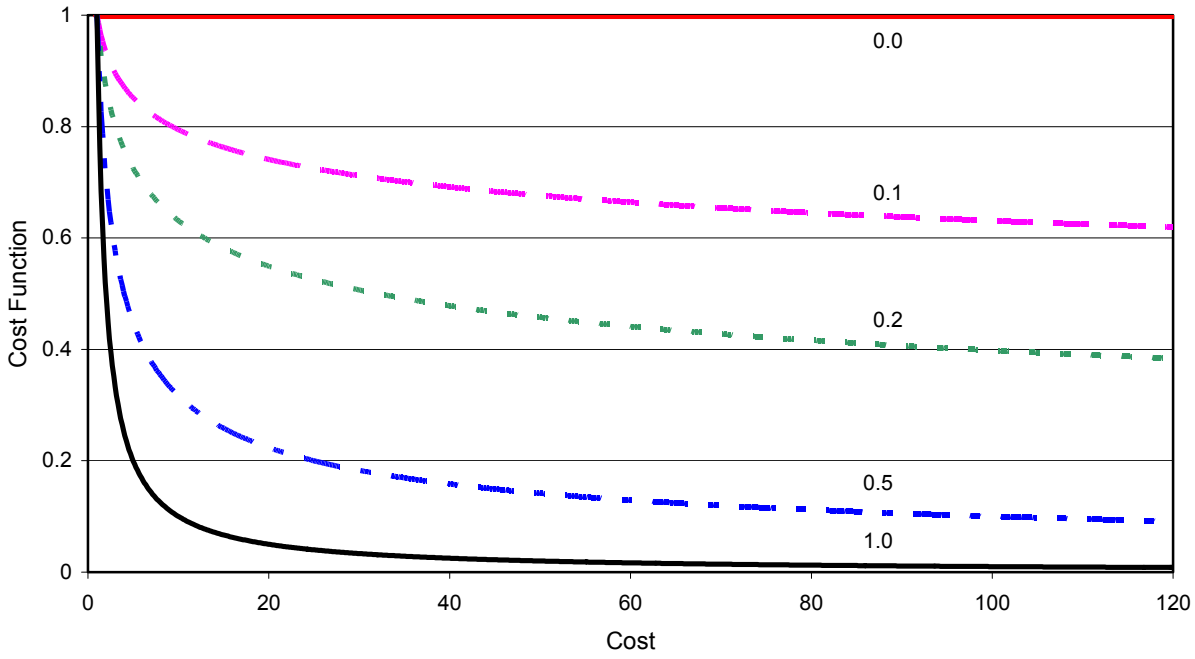
Lohse, D., H. Teichert and B. Dugge (2002) Modelling Travel Demand with Nonlinear Equation Systems, Working Paper, Institute for Transportation Planning and Road Traffic, University of Technology Dresden, Dresden, Germany.

Tanner, J. C. (1961) Factors Affecting the Amount of Travel, Road Research Technical Paper No. 51, Her Majesty's Stationery Office, London.

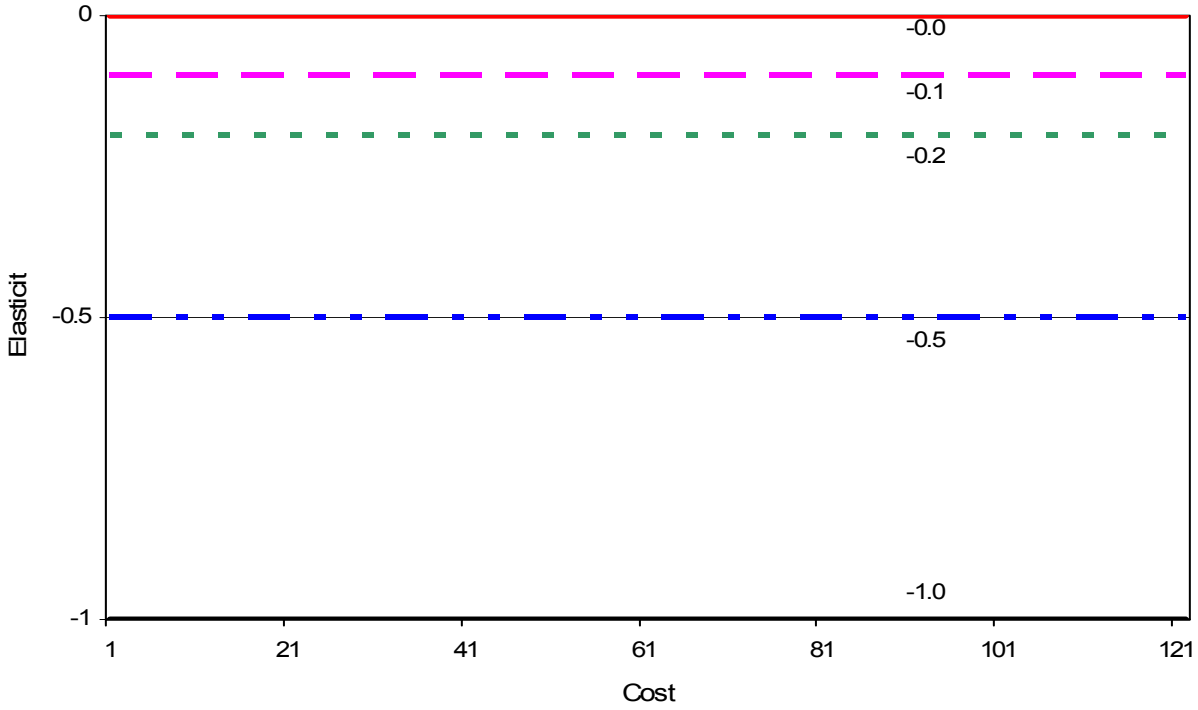

Negative Exponential Function
(cost sensitivities ranging from 0.0 to 0.1)
0.0
0.001
0.004
0.01
0.04
0.1
Cost Functic
Cost
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100
120

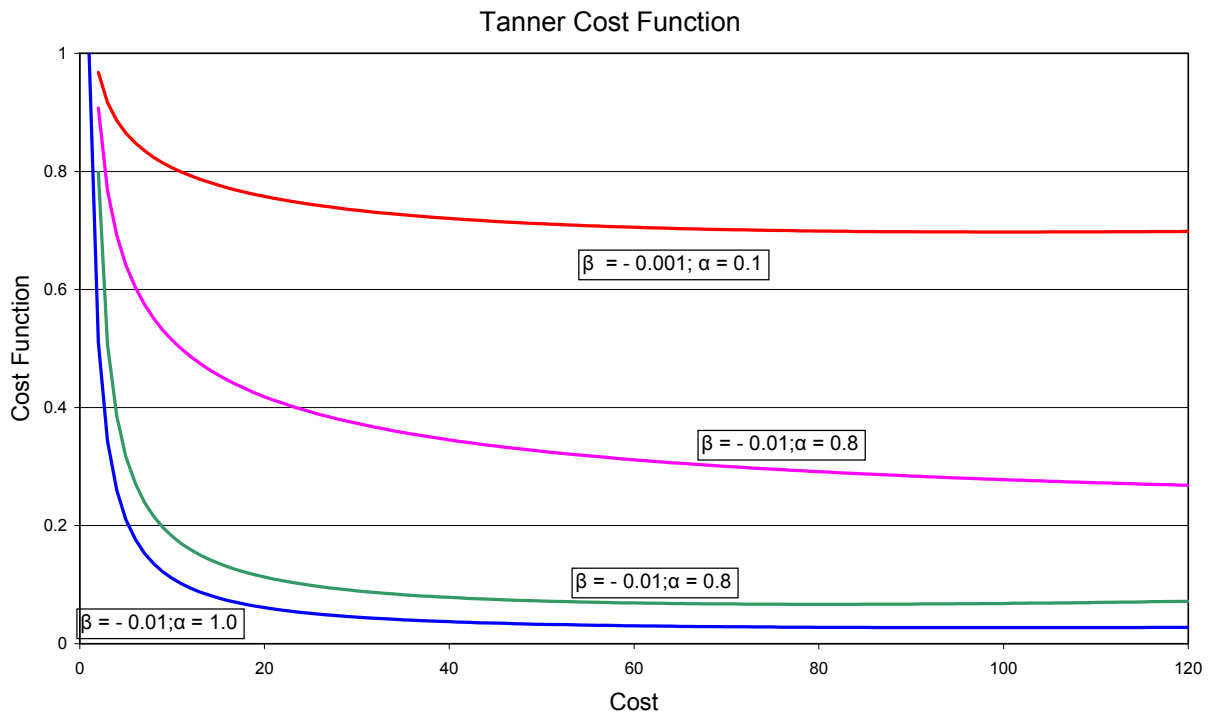

Elasticities of the Negative Exponential Function
(cost sensitivities ranging from 0.0 to 0.1)
0.0
-0.04
-0.01
-0.04
-0.1
Elasticit
Cost
0
-2
-4
-6
-8
-10
-12
0
20
40
60
80
100
120

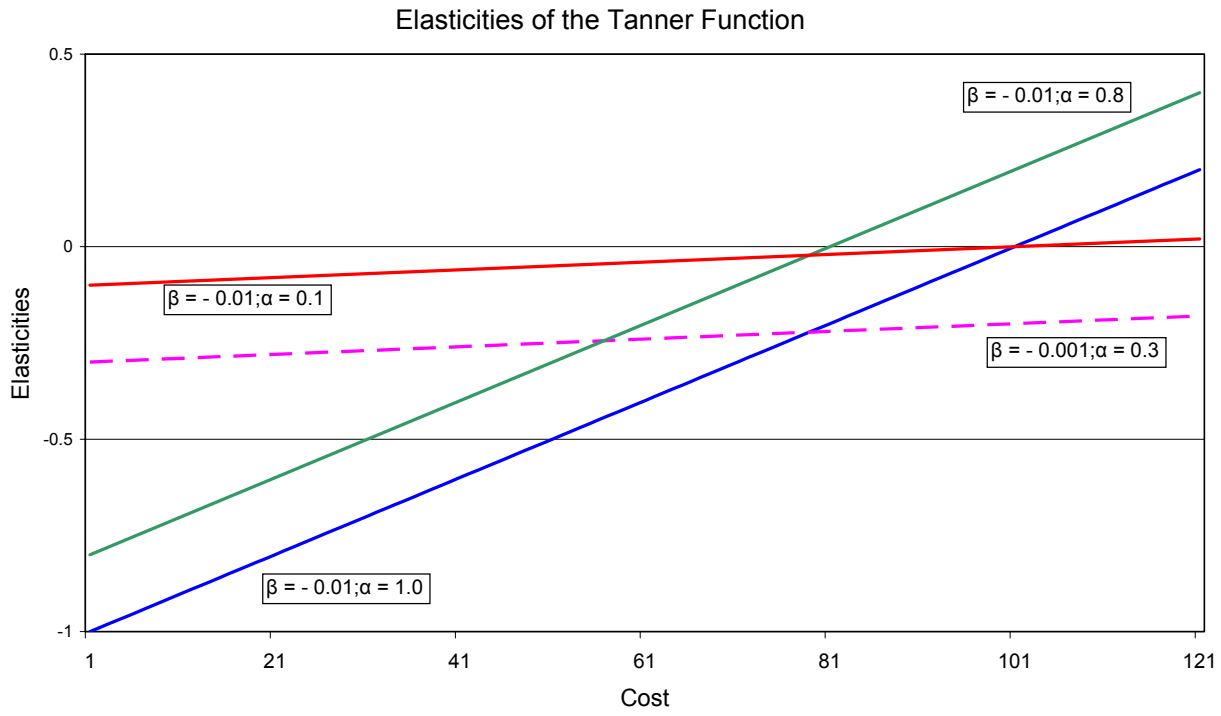
Power Cost Function
(cost sensitivities ranging from 0.0 to 1.0)



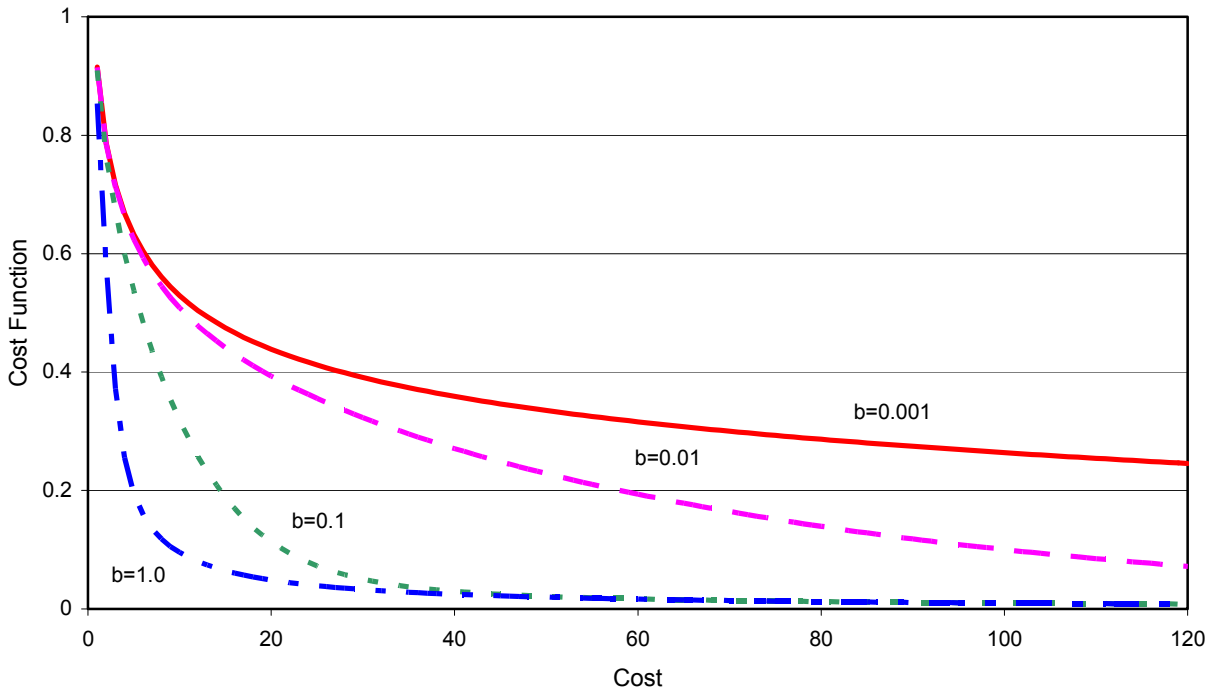
Elasticities of the Power Function
(cost sensitivities ranging from 0.0 to 1.0)



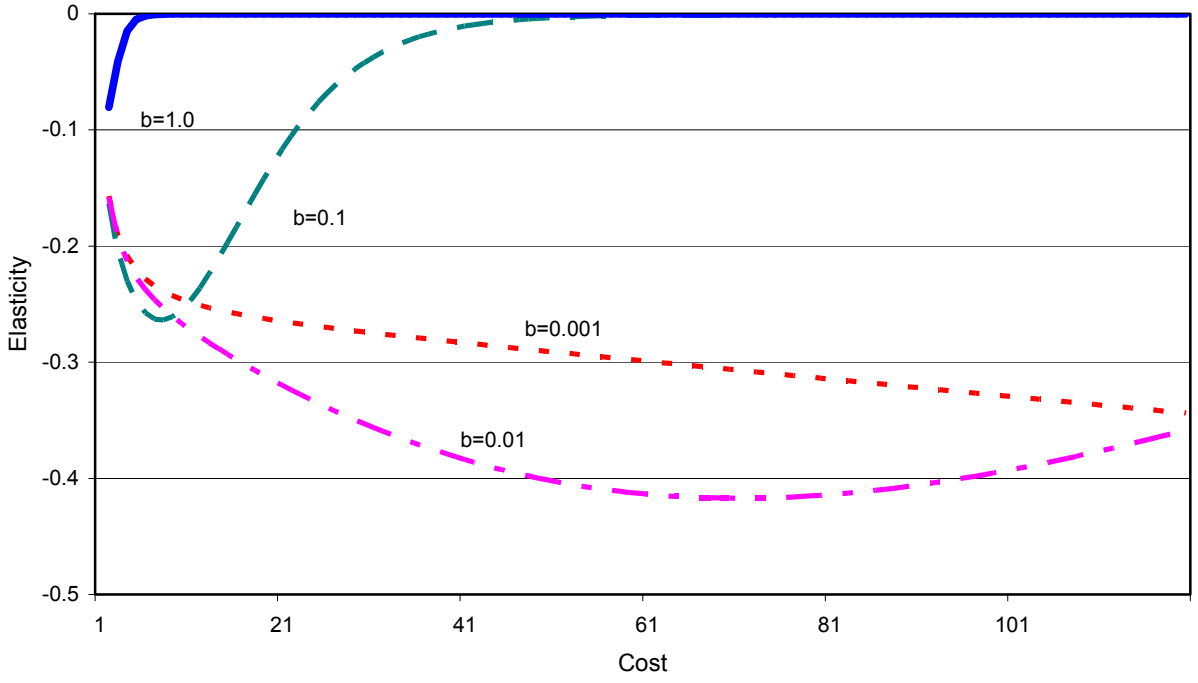

Tanner Cost Function
Cost Function
Cost
β = - 0.001; α = 0.1
β = - 0.01;α = 0.8
β = - 0.01;α = 0.8
β = - 0.01;α = 1.0
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100
120


Elasticities of the Tanner Function
Elasticities
Cost
β = - 0.01;α = 0.8
β = - 0.01;α = 0.1
β = - 0.001;α = 0.3
β = - 0.01;α = 1.0
0.5
0
-0.5
-1
1
21
41
61
81
101
121

EVA Cost Function
(cost sensitivities b (beta) ranging from 0.001 to 1.0)



Elasticities of the EVA Cost Function
(cost sensitivities b (beta) ranging from 0.001 to 1.0)



Exercises

1. Derivation of Balancing Factors

Consider the following optimality condition from the derivation of the origin-destination choice model in Section 5.1: $d_{pqm} = \exp(\kappa_p + \lambda_q - \mu \cdot c_{pqm})$. Apply the following origin and destination constraints to this function, by substituting for d_{pqm}.

$$\sum_q \sum_m d_{pqm} = O_p, \qquad p \in P;$$
$$\sum_p \sum_m d_{pqm} = D_q, \qquad q \in Q$$

Write out the constraints without using summation signs, factor out common terms, and prove that the following results are correct:

$$O_p = \exp(\kappa_p) \sum_q \left(\exp(\lambda_q) \sum_m \exp(-\mu \cdot c_{pqm}) \right)$$
$$D_q = \exp(\lambda_q) \sum_p \left(\exp(\kappa_p) \sum_m \exp(-\mu \cdot c_{pqm}) \right)$$

2. Singly Constrained Trip Distribution Model

The doubly-constrained model formulated in Section 5 is one of a family of O-D choice models. Another class of this family of models is the singly-constrained model.

Consider the following scenario. Households choose their shopping locations on the basis of the size of shops, measured by floor area, and the generalized cost from their home to the shopping areas. The number of households per zone is given, as is the floor area per shopping zone S_q. Prior to consideration of travel costs, the expected number of trips per time period is proportional to S_q/P, where P is the number of residential zones.

To represent the dispersion of trips given the prior information, we consider the following generalized dispersion function: $-\sum_{pq} \sum_m d_{pqm} \ln\left(\frac{d_{pq}}{S_q / P} \right) \geq S_0$

Assume that travel is constrained only by the trips per period leaving the origin zone p:

$$\sum_q d_{pq} = O_p, \qquad p \in P$$

Formulate and solve a shopping location choice model for this situation. To simplify the derivation, you may assume that mode choice is not represented in the model. In other words, delete the subscript m. Discuss the relation of the resulting model to the logit model introduced in Section 4.

6. Combined Models and Feedback

6.1 Problem Statement

It is unusual for the solution algorithm of a mathematical problem to be regulated by an Act of the U. S. Congress and a decree of a U. S. court resulting from a civil lawsuit. To a substantial extent, that is what has occurred with regard to the Congressional and judicial mandates to solve the four-step travel forecasting procedure "with feedback." In particular, see the Intermodal Surface Transportation Efficiency Act of 1991. In these notes, we seek to understand the issue of feedback, and the related combined model of origin-destination, mode and route choice. After examining the basic concepts, we consider a combined model implementation and a recommended feedback procedure.

The four-step procedure views the representation of the choice of where to travel, by what mode to travel, and by what route to travel as three separate, but loosely related choices which are dependent on travel times and costs prevailing in the highway and transit networks. The fourth dimension of travel choice, how often to travel, is generally regarded as independent of network conditions, and assumed to depend on traveler and household characteristics; accordingly, it will not be considered further here except as a source of control totals for the remaining travel choices.

A principal difficulty with solving the sequential procedure is that the travel times and costs on the road network are not fixed and known, but rather are the result of solving an equilibrium model, as has been seen in our consideration of trip assignment (route choice) models. In the four-step procedure, however, these generalized travel costs are not just needed for the route choice model, but are also essential inputs to the O-D choice and mode choice models, known traditionally as trip distribution and mode split models. Since the solution of these models also determines the outcome of the route choice model, a circularity exists, which has come to be known in the urban transportation planning field as "solving the models with feedback." Stated intuitively, the problem is to bring the generalized travel costs, which are inputs to the travel choice models and outputs of the route choice models, into agreement.

There is at least one other difficulty, namely that the definitions of travel times and costs in the three choice models are not always consistent. For example, auto and transit travel times may be the basis for the route choice models, modal travel times and operating costs/fare the basis for the mode choice model, but only auto travel times the basis for the O-D choice model. While it is not imperative for these travel times and costs to be the same in each model, they should be defined consistently.

This problem is as old as the four-step procedure itself. Moreover, it was the subject of considerable study in the early 1970s by a British applied mathematician, John D. Murchland, and an Australian graduate student at University College London, Suzanne P. Evans. She successfully solved the problem of "combining trip distribution and trip assignment" in her Ph.D. thesis completed in 1973, published two papers on her findings and left the field. A somewhat parallel effort by Michael Florian at the University of Montreal produced a related, but less

useful result. Because of the term coined by Murchland and Evans, these formulations came to be known as *combined models*.

When the problem of "solving the four-step procedure with feedback" is viewed from the standpoint of "combined models", we believe that the conceptual issues are at once clear and straightforward. However, our experience is that the combined model framework is not as compelling to others as it is to us. This situation inspires us to try again to communicate these results effectively.

6.2 Formulation of the Combined Model

We begin our formulation with the auto route choice model with the objective function augmented to include transit costs:

$$\min_{(\mathbf{h})} z(\mathbf{h}) = \eta \cdot \sum_a \int_0^{f_a} c_a(x)dx + \sum_{pq} c_{pqt} \cdot d_{pqt}$$

$$\text{s.t.} \sum_{r \in R_{pq}} h_r = d_{pqa} / \eta, \qquad p = 1,...P; q = 1,...Q$$

$$\mathbf{h} \geq \mathbf{0}$$

$$\text{where } f_a \equiv \sum_{r \in R_{pqa}} h_r \delta_r^a, \qquad a \in A$$

where d_{pqm} is given for each mode, and the auto occupancy η is given.

To this objective function and route flow conservation constraint, we add constraints requiring that the auto and transit flows from zone p to zone q sum to the fixed O-D flows d_{pq}. Then, we define a dispersion constraint that relates the dispersion of flows to the higher cost mode to the observed dispersion. Finally, we have constraints requiring all variables to be nonnegative.

$$d_{pqa} + d_{pqt} = d_{pq}, \qquad p = 1,...P; q = 1,...Q$$

$$-\sum_{pqm} d_{pqm} \ln d_{pqm} = S_0$$

$$d_{pqm} > 0, \qquad p = 1,...P; q = 1,...Q, m = a, t.$$

In this formulation, we have relaxed some constraints on the auto route choice model, added a term in the objective function relating to transit costs, and added conservation of flow constraints pertaining to mode choice. The result is a combined mode and route choice model with the following properties:

1. The route choice conditions are similar to conditions in the UO route choice model;

2. The mode choice conditions correspond to the familiar logit model of mode choice:

$$d_{pqm} = \frac{d_{pq} \exp\left(-\mu \cdot c_{pqm}\right)}{\sum_{n \in \{a,t\}} \exp\left(-\mu \cdot c_{pqn}\right)}$$

where c_{pqm} is the generalized travel cost of mode *m*, and c_{pqa} is the UO auto cost.

3. To add O-D choice to the model, we relax the constraint that d_{pq} is fixed, and add the following constraints:

$$\sum_{qm} d_{pqm} = O_p, \qquad p = 1,...P$$
$$\sum_{pm} d_{pqm} = D_q, \qquad q = 1,...Q$$

These constraints are the familiar row and column sum constraints from the O-D choice (trip distribution) model, which state that the origin-destination-mode flows must sum to the specified origin and destination totals.

4. The optimality conditions of this model are the same as before for route choice, plus

$$d_{pqm} = A_p O_p B_q D_q \exp\left(-\mu \cdot c_{pqm}\right)$$

where A_p and B_q are balancing factors that insure the origin and destination constraints are satisfied.

In summary, from this formulation, we can see that the way to *combine* O-D and mode choice with route choice is to augment the route choice problem with the constraints pertaining to O-D and mode choice. This was the first major insight of S. P. Evans. Her second fundamental result was to show that the Frank-Wolfe linearization method could be generalized to a partial linearization method to solve this problem formulation, and to prove that the algorithm does converge to the desired solution. The mathematical result is presented in detail in Patriksson (1994), as well as in Evans (1976). This algorithm is stated in the next section.

6.3 Evans Algorithm for the Combined Model

The fundamental insight of Evans was to realize that it is only necessary to linearize that part of the objective function pertaining to variables which cannot be solved directly from the optimality conditions. Since optimality conditions are available that define origin-destination and mode choice, as described above, only the auto route choice portion of the objective function is linearized, as in the traffic assignment problem with fixed demand. Once the O-D-mode and route choice variables have been updated, then a line search is used to find a step length or weight to drive the solution towards the optimum. The problem to be solved by the algorithm may be stated as follows.

TAP-ODMC

$$\min_{(\mathbf{d},\mathbf{h})} z(\mathbf{d},\mathbf{h}) = \eta \cdot \sum_a \int_0^{f_a} c_a(x)dx + \sum_{pq} c_{pqt} \cdot d_{pqt} + \frac{1}{\mu} \cdot \sum_{pqm} d_{pqm} \ln d_{pqm}$$

$$\text{st}: \sum_{r \in R_{pq}} h_r = d_{pqa} / \eta, \qquad p = 1,...P; \qquad q = 1,...Q$$

$$d_{pqa} + d_{pqt} = d_{pq}, \qquad p = 1,...P; \qquad q = 1,...Q$$

$$\sum_{qm} d_{pqm} = O_p, \qquad p = 1,...P$$

$$\sum_{pm} d_{pqm} = D_q, \qquad q = 1,...Q$$

$$d_{pqm} > 0, \qquad p = 1,...P; \qquad q = 1,...Q, \qquad m = a,t$$

$$h_r \geq 0, \qquad r \in R_{pq}, \qquad p = 1,...P; \qquad q = 1,...Q$$

$$\text{where } f_a \equiv \sum_{r \in R_{pqa}} h_r \delta_r^a, \qquad a \in A$$

The objective function has been augmented to include the only nonlinear term of the Lagrange function, $-\sum_{pqm} d_{pqm} \ln d_{pqm}$, and its Lagrange multiplier $1/\mu$. All of the constraints are linear. The main problem variables are the origin-destination-mode flows (d_{pqm}) and the auto link flows (f_a); the subproblem variables are (e_{pqm}) and (g_a) respectively. The Evans algorithm is stated as follows.

1. Initialization - compute an initial solution $(d^1_{pqm}),(f^1_a)$ for origin-destination-mode flows and link flows; a common procedure is to set road link flows to zero, compute the links costs and the shortest route costs, solve for the O-D-mode flows, and assign the O-D auto flows to their shortest routes.

2. Update link costs: $c_a = c_a(f^k_a)$ where the iteration index *k* is initially 1.

3. Compute shortest routes from each origin *p* to each destination *q* to obtain (c^k_{pqa}).

4. Solve the origin-destination-mode choice model for the subproblem flows (e^k_{pqm}).

5. Perform an all-or-nothing assignment of the auto O-D flows (e^k_{pqa}) to the shortest routes from *p* to *q*, yielding link flows (g^k_a); that is, increase the flow on each link of the shortest route by the O-D-auto mode flow.

6. Compute the Relative Gap and test for convergence:

$$\text{Gap}^k = \eta \cdot \sum_a c_a\left(f_a^k\right)\cdot\left(g_a^k - f_a^k\right) + \sum_{pq} c_{pqt}\left(e_{pqt}^k - d_{pqt}^k\right) + \frac{1}{\mu}\sum_{pqm} e_{pqm}^k \ln\left(e_{pqm}^k\right) - \frac{1}{\mu}\sum_{pqm} d_{pqm}^k \ln\left(d_{pqm}^k\right)$$

$$\begin{aligned}
\text{LB}^k &= \eta \cdot \sum_a \int_0^{f_a} c_a(x)dx + \sum_{pq} c_{pqt} d_{pqt} + \frac{1}{\mu}\sum_{pqm} d_{pqm}^k \ln\left(d_{pqm}^k\right) + \text{Gap}^k \\
&= \eta \cdot \sum_a \int_0^{f_a} c_a(x)dx + \sum_{pq} c_{pqt} d_{pqt} + \frac{1}{\mu}\sum_{pqm} d_{pqm}^k \ln\left(d_{pqm}^k\right) + \eta \cdot \sum_a c_a\left(f_a^k\right)\cdot\left(g_a^k - f_a^k\right) \\
&\quad + \sum_{pq} c_{pqt}\left(e_{pqt}^k - d_{pqt}^k\right) + \frac{1}{\mu}\sum_{pqm} e_{pqm}^k \ln\left(e_{pqm}^k\right) - \frac{1}{\mu}\sum_{pqm} d_{pqm}^k \ln\left(d_{pqm}^k\right) \\
&= \eta \cdot \sum_a \int_0^{f_a} c_a(x)dx + \eta \cdot \sum_a c_a\left(f_a^k\right)\cdot\left(g_a^k - f_a^k\right) + \sum_{pq} c_{pqt} e_{pqt} + \frac{1}{\mu}\sum_{pqm} e_{pqm}^k \ln\left(e_{pqm}^k\right)
\end{aligned}$$

$$\text{BLB} = \max_k\left(\text{LB}^k\right)$$

Is the Relative $\text{Gap}^k \equiv \dfrac{\text{Gap}^k}{\text{BLB}} < \varepsilon$? If YES, STOP; otherwise go to Step 7.

7. Perform a line search to determine the optimal step length (weight) λ^k:

$$\min_{(\lambda^k)} z(\lambda^k) = \eta \cdot \sum_a \int_0^{f_a^{k+1}} c_a(x)dx + \sum_{pq} c_{pqt} \cdot d_{pqt}^{k+1} + \frac{1}{\mu} \cdot \sum_{pqm} d_{pqm}^{k+1} \ln d_{pqm}^{k+1}$$

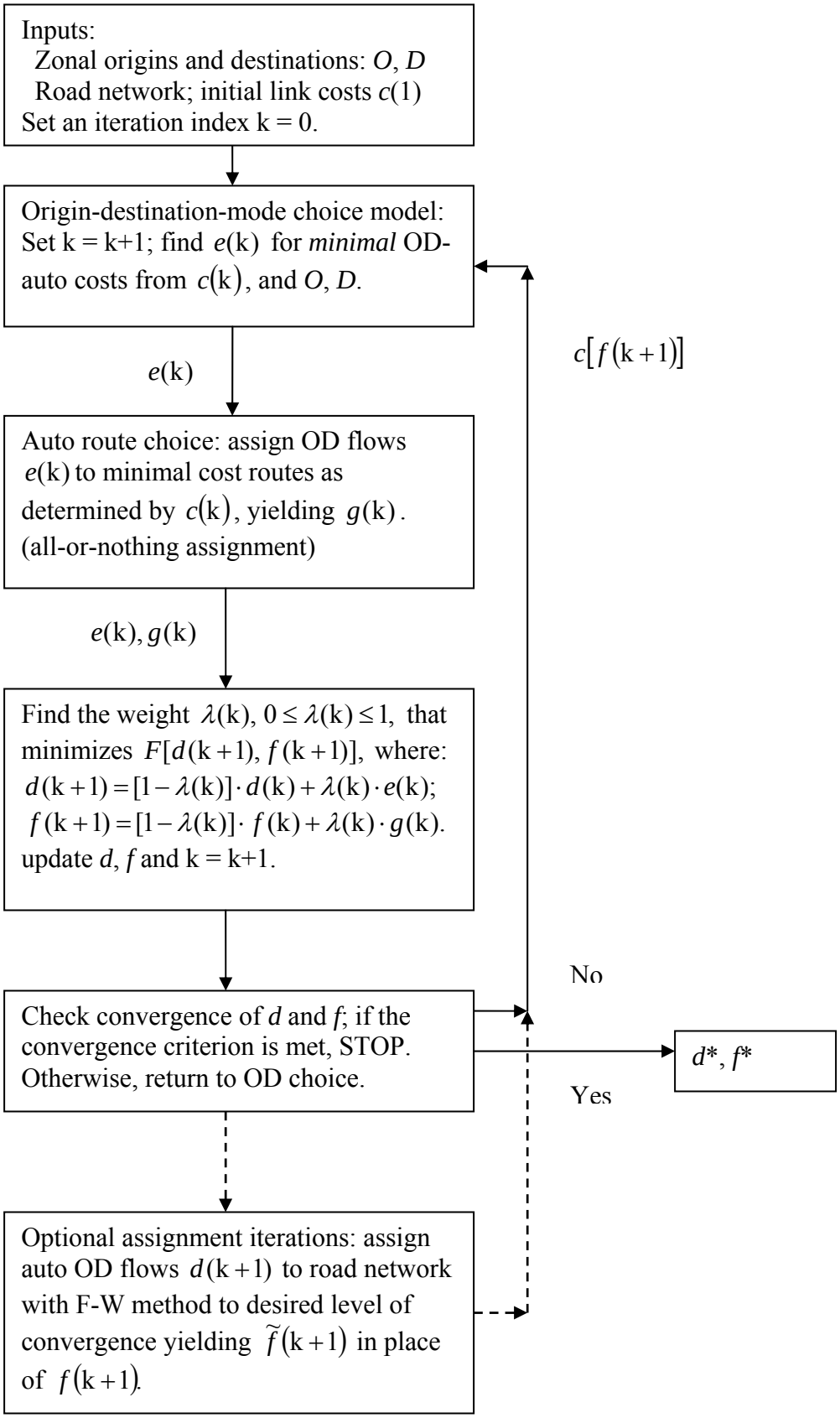
where $f_a^{k+1} = \left(1-\lambda^k\right)\cdot f_a^k + \lambda^k \cdot g_a^k$ and $d_{pqm}^{k+1} = \left(1-\lambda^k\right)\cdot d_{pqm}^k + \lambda^k \cdot e_{pqm}^k$

8. Update the O-D-mode and link flows:

$$d_{pqm}^{k+1} = \left(1-\lambda^k\right)\cdot d_{pqm}^k + \lambda^k \cdot e_{pqm}^k$$
$$f_a^{k+1} = \left(1-\lambda^k\right)\cdot f_a^k + \lambda^k \cdot g_a^k$$

9. Retest the updated value of the objective function for convergence.

The algorithm is illustrated by the following simplified flowchart.



A Generalized Evans Algorithm

Comments on the verbal statement of the algorithm:

1. Step 3 involves applying a shortest route algorithm to each origin node *p*.

2. Step 4 consists of solving an origin-destination-mode choice model using a doubly-constrained gravity model with mode-specific costs.

3. Step 5 performs an all-or-nothing assignment of O-D flow by auto to the road network.

4. According to Lundgren and Patriksson (1998), the solution of the subproblem provides a lower bound on the optimal value of $z(\mathbf{d},\mathbf{h})$. Feasible solutions to **TAP-ODMC** provide upper bounds on $z(\mathbf{d},\mathbf{h})$. The difference between the upper and lower bounds may be regarded as the Gap. The Gap/BLB defines a Relative Gap, which may be used to monitor convergence of the solution.

5. Step 7 finds the optimal step length λ^k.

6. Step 8 updates the O-D-mode flows and the auto link flows.

6.4 Case Study – An Application of the Combined Model to the Chicago Region

A research team led by Boyce and Bar-Gera (2003) applied a multiple-class version of the combined model to data for the Chicago region from the 1990 Census and household survey.

TAP-CM

$$\begin{aligned}
\min T(\mathbf{d},\mathbf{h}) \equiv & \sum_a \int_0^{f_a} t_a(f_a) + \sum_{la}\left(\gamma_2^l \cdot f_a^l \cdot k_a + \gamma_4^l \cdot f_a^l \cdot s_a\right) + \sum_{lpq} \frac{\gamma_3^l}{\nu^l} \cdot d_{pq,au}^l \cdot w_{pq,au} \\
& + \sum_{lp} \frac{d_{pp,au}^l}{\nu^l}\left(t_{pp,au} + \gamma_4^l \cdot s_{pp,au}\right) + \sum_{lpq}\left(\frac{d_{pq,tr}^l}{\nu^l}\left(\gamma_5^l \cdot t_{pq,tr} + \gamma_6^l \cdot k_{pq,tr} + \gamma_7^l \cdot w_{pq,tr} + \gamma_8^l\right)\right) \\
& + \sum_{lpq} \frac{1}{\eta^l \nu^l} \cdot d_{pq}^l\left(\ln d_{pq}^l - 1\right) + \sum_{lpqm} \frac{1}{\mu^l \nu^l} \cdot d_{pqm}^l\left(\ln\left(\frac{d_{pqm}^l}{d_{pq}^l}\right) - 1\right)
\end{aligned}$$

$$st: \quad \sum_{r \in R_{pq}} h_r^l = \frac{d_{pq,au}^l}{\nu^l}, \; p,q \in Z; l \in L.$$

$$\sum_{r \in R_{pq}} h_r^k = K_{pq}, \; p,q \in Z.$$

$$\sum_m d_{pqm}^l = d_{pq}^l, \; p,q \in Z; l \in L.$$

$$\sum_q d_{pq}^l = O_p^l, \; p \in Z; l \in L.$$

$$\sum_p d_{pq}^l = D_q^l, \; q \in Z; l \in L.$$

$$h_r^l \geq 0, r \in R_{pq},\ p,q \in Z;\ l \in L.$$

$$h_r^k \geq 0, r \in R_{pq},\ p,q \in Z.$$

where $$f_a \equiv \sum_l f_a^l + f_a^k \equiv \sum_{lr} h_r^l \delta_r^a + \sum_r h_r^k \delta_r^a,\ a \in A.$$

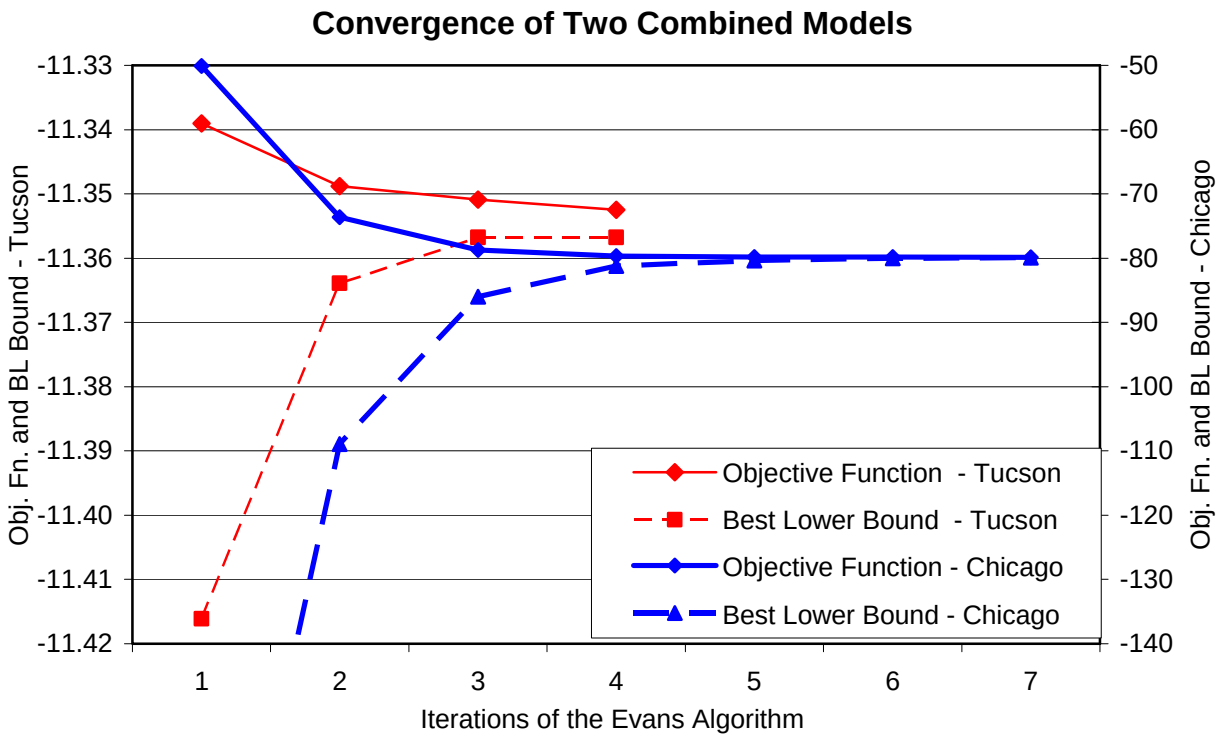
Additional definitions used in the above formulation are:

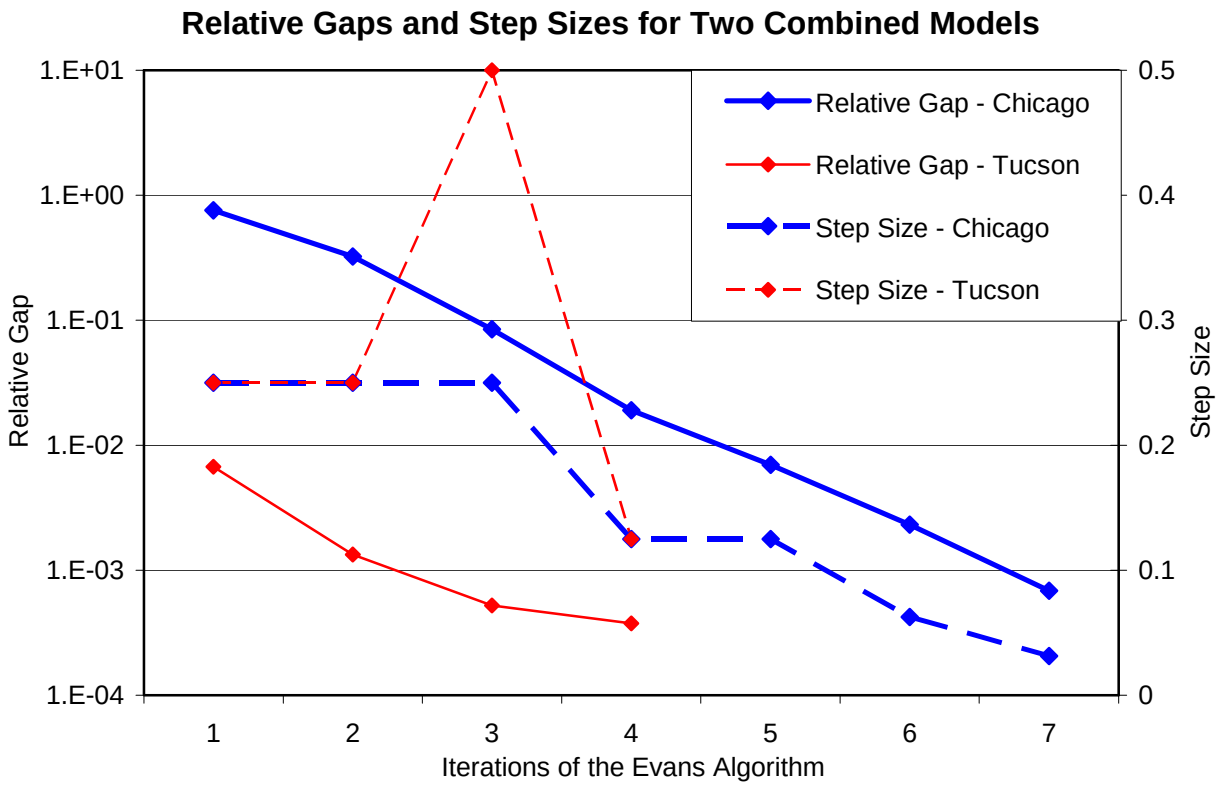
$t_a(f_a) =$ in-vehicle auto travel time on road link *a*, a function of total vehicle flow f_a (min)
k_a = vehicle toll per auto equivalent unit on link *a* or parking fee at the terminal link (cents)
s_a = length of link *a* (miles)
$\delta_r^a =$ 1, if link *a* belongs to route *r*, and 0 otherwise
$t_{pp,au} =$ average in-vehicle auto travel time within zone *p* (minutes); fixed, exogenous value
$s_{pp,au} =$ average length of auto trip within zone *p* (miles); fixed, exogenous value
$w_{pq,au} =$ out-of-vehicle travel time by auto for travel from zone *p* to zone *q* (minutes)
$t_{pq,tr} =$ in-vehicle transit travel time from zone *p* to zone *q* (minutes)
$k_{pq,tr} =$ transit fare from zone *p* to zone *q* (cents)
$w_{pq,tr} =$ out-of-vehicle travel time by transit from zone *p* to zone *q* (minutes)
$\gamma_1^l = 1,$ the coefficient associated with auto in-vehicle travel time for class *l*
$\gamma_2^l =$ the coefficient associated with auto monetary cost for class *l*
$\gamma_3^l =$ the coefficient associated with auto out-of-vehicle travel time for class *l*
$\gamma_4^l =$ the coefficient associated with auto travel distance for class *l*
$\gamma_5^l =$ the coefficient associated with transit in-vehicle travel time for class *l*
$\gamma_6^l =$ the coefficient associated with transit fare for class *l*
$\gamma_7^l =$ the coefficient associated with transit out-of-vehicle travel time for class *l*
$\gamma_8^l =$ the coefficient associated with transit bias for class *l*
$\mu^l, \eta^l =$ mode and OD cost sensitivity parameters for class *l*

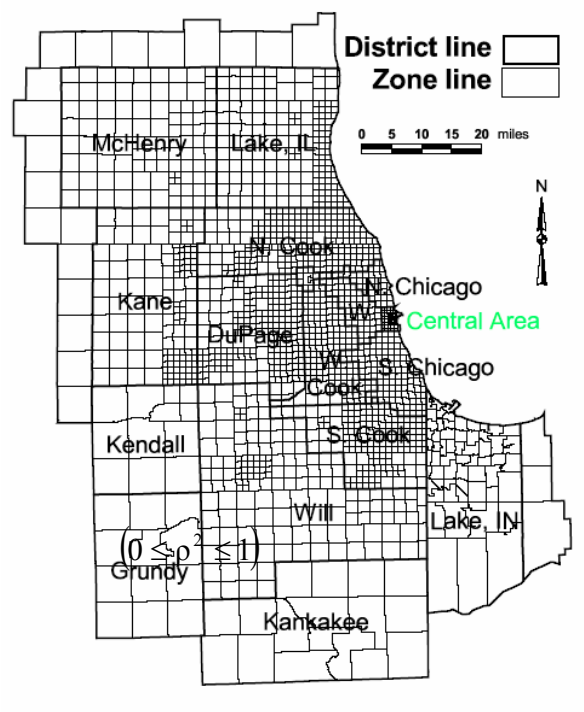
The ODM choice model implied by this formulation is:

$$d_{pqm}^l = A_p^l O_p^l B_q^l D_q^l \exp\left(-\eta^l \tilde{c}_{pq}^l\right) \cdot \frac{\exp\left(-\mu^l c_{pqm}^l\right)}{\sum_m \exp\left(-\mu^l c_{pqm}^l\right)}$$

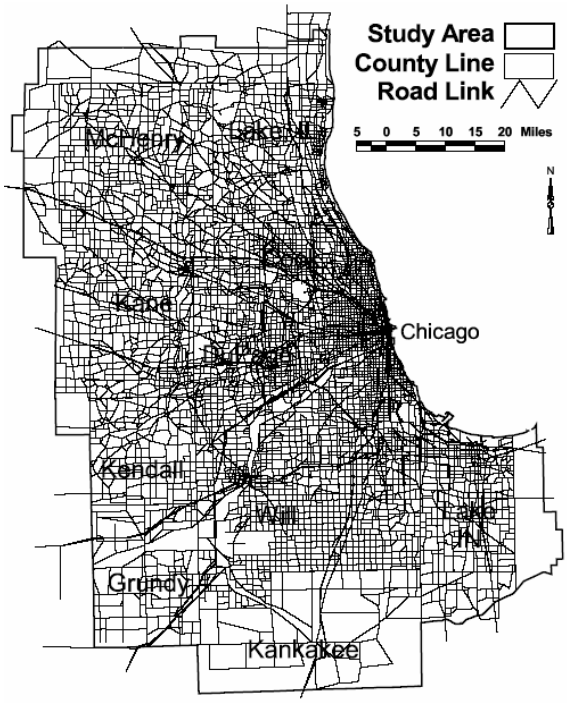
This model was estimated with data from the 1990 household travel survey (HHTS) conducted by the Chicago Area Transportation Study. Details of the estimation procedure are described by Boyce and Bar-Gera (2003). The convergence of the solution to the Chicago model is shown on the next two charts followed by two maps of the region and the model parameters and performance measures in Tables 1 and 2. For comparison with a model of a less congested region, results are also shown for Tucson, Arizona.


Convergence of Two Combined Models
-11.33
-11.34
-11.35
-11.36
-11.37
-11.38
-11.39
-11.40
-11.41
-11.42
Obj. Fn. and BL Bound - Tucson
-50
-60
-70
-80
-90
-100
-110
-120
-130
-140
Obj. Fn. and BL Bound - Chicago
Objective Function - Tucson
Best Lower Bound - Tucson
Objective Function - Chicago
Best Lower Bound - Chicago
1
2
3
4
5
6
7
Iterations of the Evans Algorithm


Relative Gaps and Step Sizes for Two Combined Models
1.E+01
1.E+00
1.E-01
1.E-02
1.E-03
1.E-04
Relative Gap
0.5
0.4
0.3
0.2
0.1
0
Step Size
Relative Gap - Chicago
Relative Gap - Tucson
Step Size - Chicago
Step Size - Tucson
1
2
3
4
5
6
7
Iterations of the Evans Algorithm



District and Zone Map for Chicago Region



Road Network Map for Chicago Region

TABLE 1: Estimated Values of Coefficients for Two Multiclass Travel Choice Models

Travel Class	Home-to-Work Travel		Other Travel	
Generalized Cost Coefficients (units)	Auto	Transit	Auto	Transit
In-vehicle Travel Time (gc/min)	1.0	0.25	1.0	0.28
Out-of-vehicle Travel Time (gc/min)	0.0	0.90	0.47	0.46
Monetary Cost (gc/cents)	0.049	0.084	0.0003	0.052
Length of Route (gc/miles)	0.15	NA	0.05	NA
Transit Bias (gc)	NA	0.00	NA	0.09
Cost Sensitivity Coefficients				
Simultaneous O-D-Mode	0.14		0.25	
Nested Logit				
O-D	0.13		0.28	
Mode	0.15		0.20	
Goodness-of-Fit				
Simultaneous	0.19		0.33	
Nested Logit	0.21		0.35	

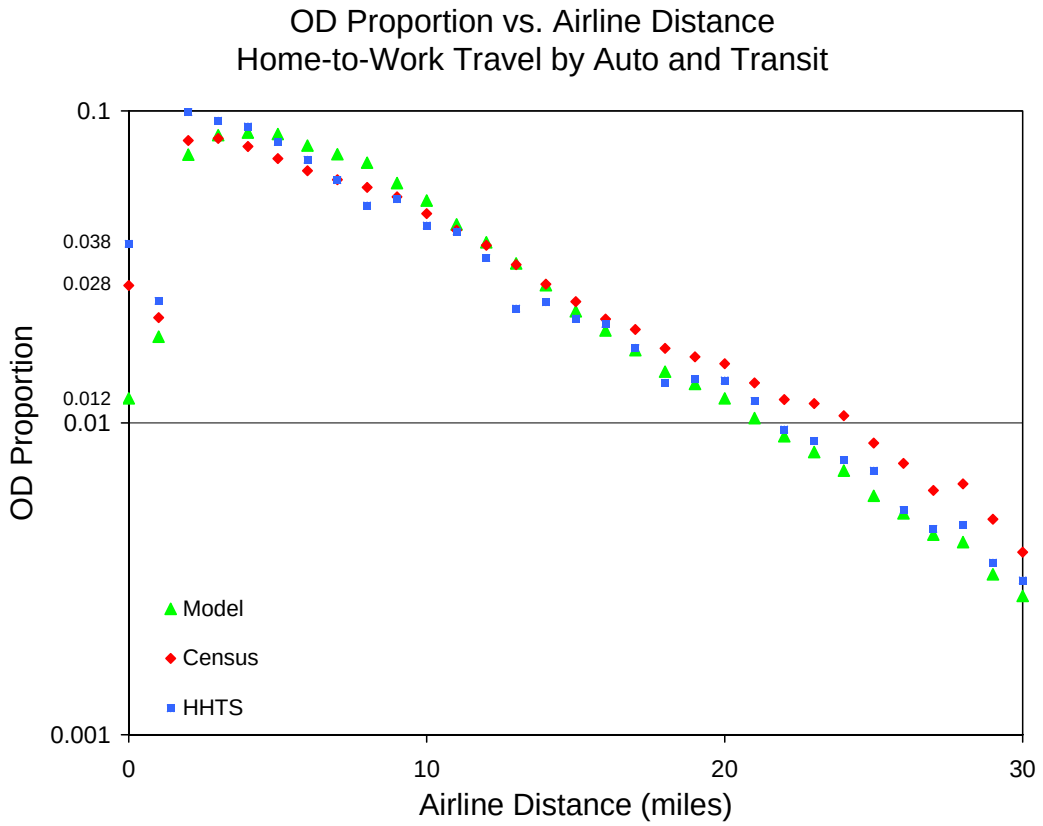
gc: generalized cost of travel in auto in-vehicle minutes
NA: not applicable
Note: All estimated values are rounded to two significant digits.

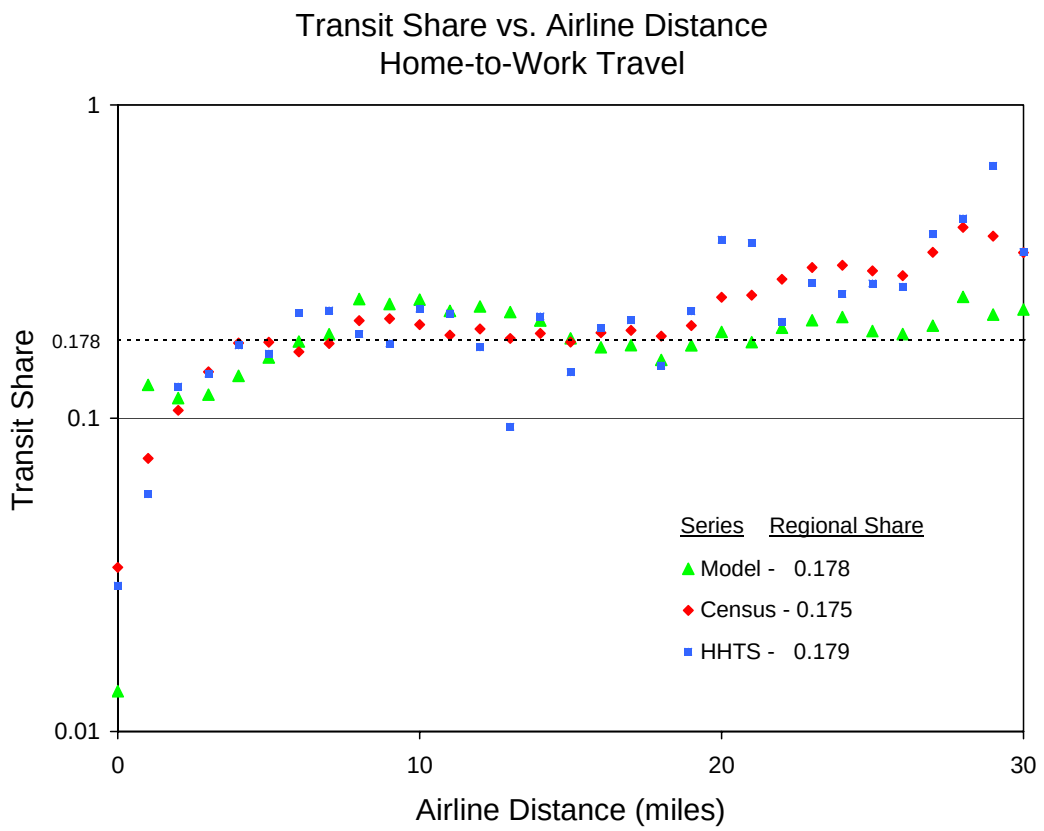
TABLE 2: Summary Means for the Multinomial Logit and Nested Logit Models

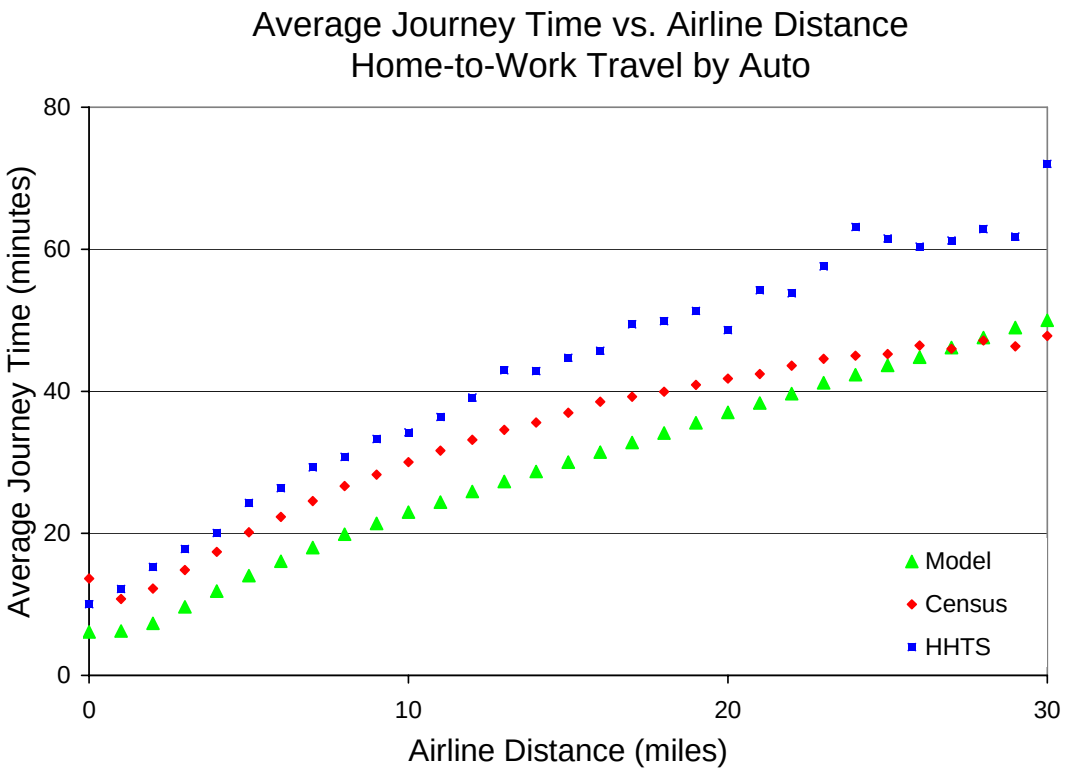
Class		Home-to-Work Travel			
Mode		Auto		Transit	
Model		Multi-Logit	Nested Logit	Multi-Logit	Nested Logit
Variable	unit				
In-vehicle time	minutes	16.1	17.7	25.5	24.9
Out-of-vehicle time	minutes	2.7	2.7	12.8	13.1
Total travel time	minutes	18.8	20.4	38.3	38.0
Monetary cost	dollars	0.31	0.34	1.46	1.46
Length	miles	10.8	11.5	NA	NA
Generalized cost	minutes	20.1	21.3	30.0	30.2
Observed gen. cost	*minutes*	*18.4*	*18.6*	*29.7*	*29.7*
Regional share	percent	83.7	83.7	16.3	16.3
Observed reg.share	*percent*	*83.8*		*16.2*	
Central Area share	percent	34.8	35.0	65.2	65.0
Observed CA share	*percent*	*36.*		*64.*	
Class		Other Travel			
Mode		Auto		Transit	
Model		Multi-Logit	Nested Logit	Multi-Logit	Nested Logit
Variable	unit				
In-vehicle time	minutes	8.9	8.4	16.3	14.5
Out-of-vehicle time	minutes	2.5	2.6	15.5	14.0
Total travel time	minutes	11.3	10.9	31.9	28.5
Monetary cost	dollars	0.15	0.14	1.32	1.28
Length	miles	5.4	4.9	NA	NA
Generalized cost	minutes	11.8	11.2	18.6	17.2
Observed gen. cost	*minutes*	*10.3*	*10.4*	*18.9*	*18.9*
Regional share	percent	82.7	83.0	17.3	17.0
Observed reg.share	*percent*	*82.8*		*17.2*	
Central Area share	percent	40.8	51.3	59.2	48.7
Observed CA share	*percent*	*41.*		*59.*	

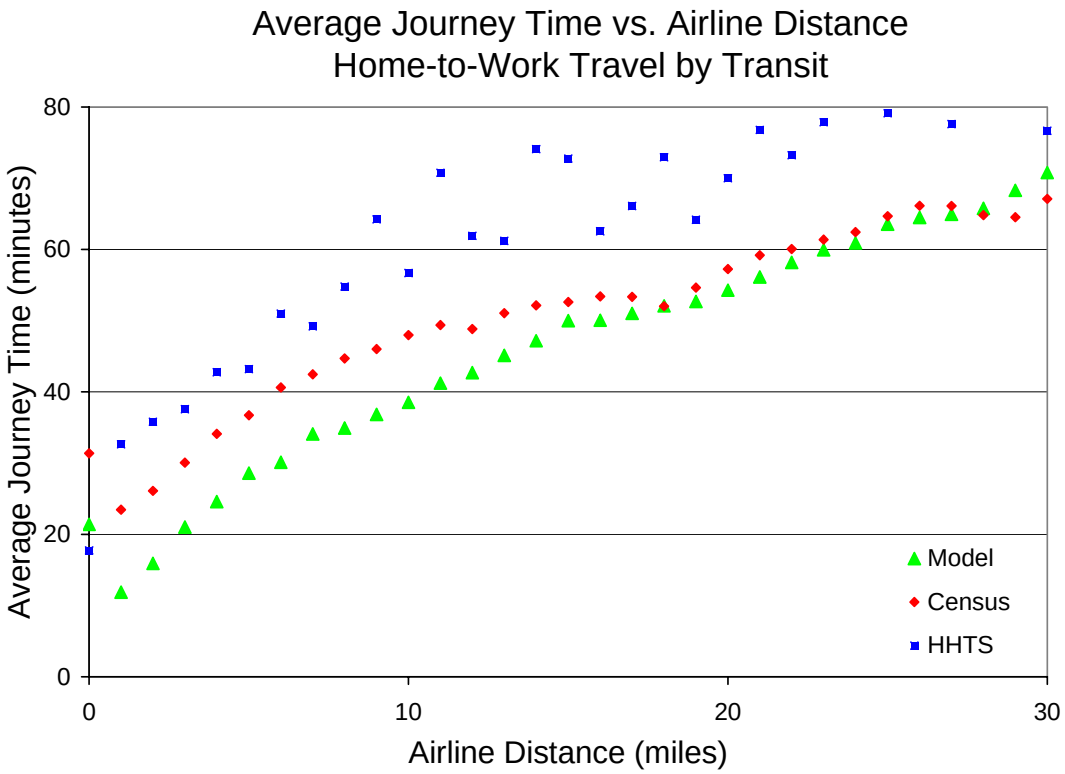
NA – not applicable
Italicized rows denote observed values.
Observed generalized costs are different for the Simultaneous and Nested Logit models because of differences in model coefficients that determine these values.

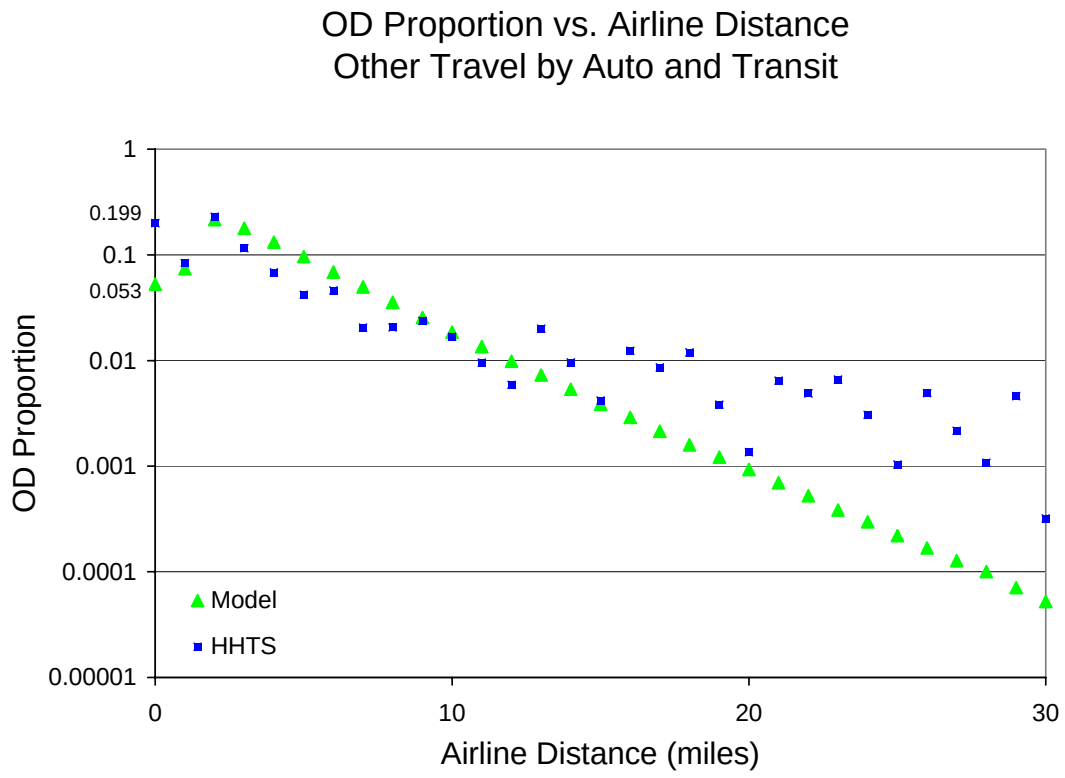
The Evans algorithm was used in the estimation procedure, as well as for solving the final model for validation studies. The validation studies are shown in the following figures. The first four figures for Home-to-Work Travel compare the Census validation data and the model predictions aggregated by airline distance from origins to destinations. The data on which the model was estimated is also shown. The following four figures for Other Travel show only the predictions and the data used in the estimation, since Census data are not available for this class of travel. The last two sets of figures compare the predictions and Census data for district-to-district flows.

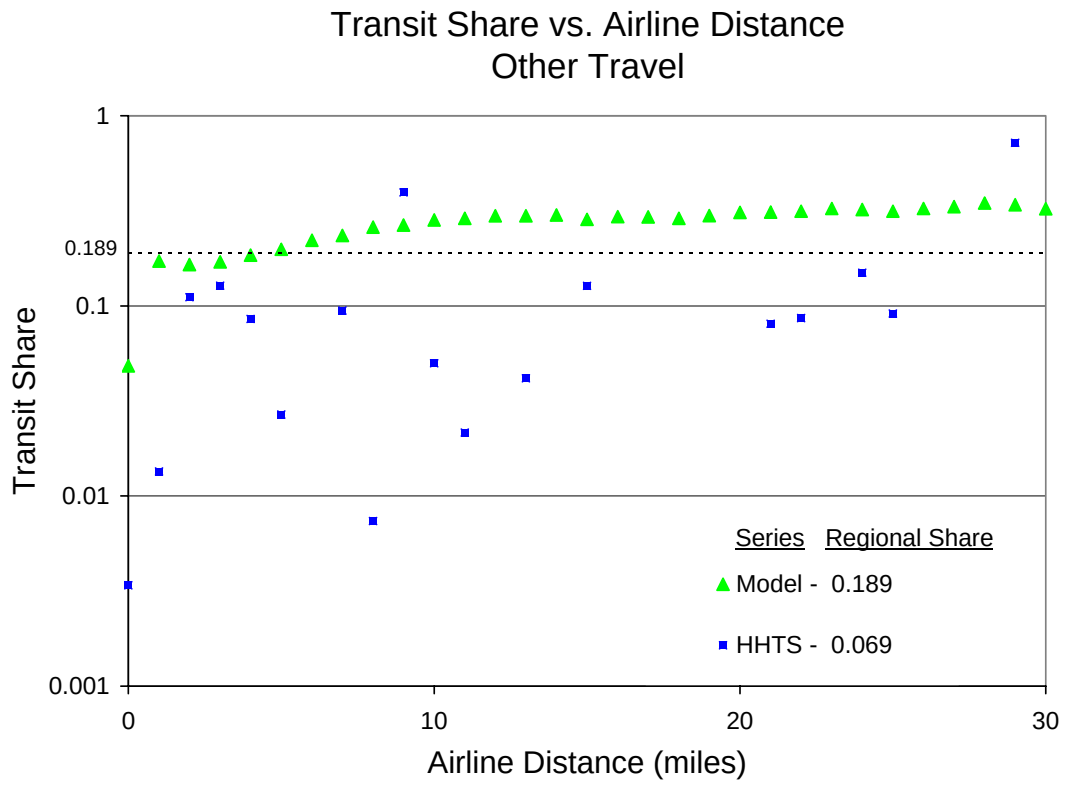

OD Proportion vs. Airline Distance
Home-to-Work Travel by Auto and Transit
OD Proportion
0.1
0.038
0.028
0.012
0.01
0.001
0
10
20
30
Airline Distance (miles)
Model
Census
HHTS

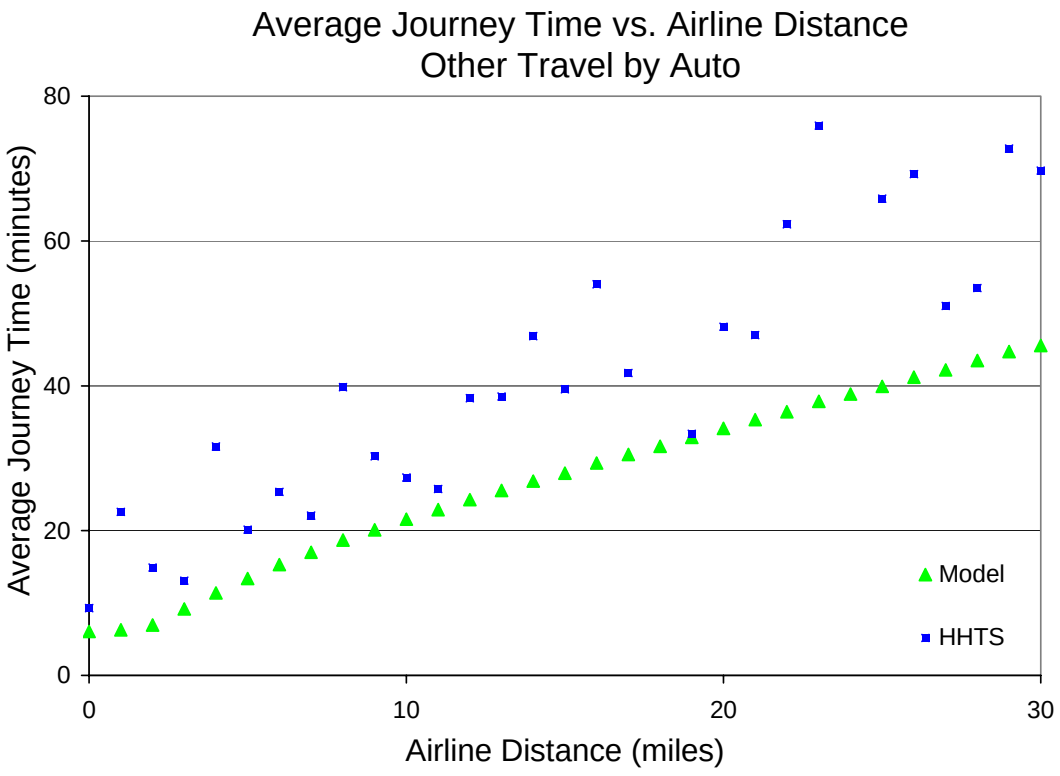

Transit Share vs. Airline Distance
Home-to-Work Travel
Transit Share
1
0.178
0.1
0.01
0
10
20
30
Airline Distance (miles)
Series Regional Share
Model - 0.178
Census - 0.175
HHTS - 0.179

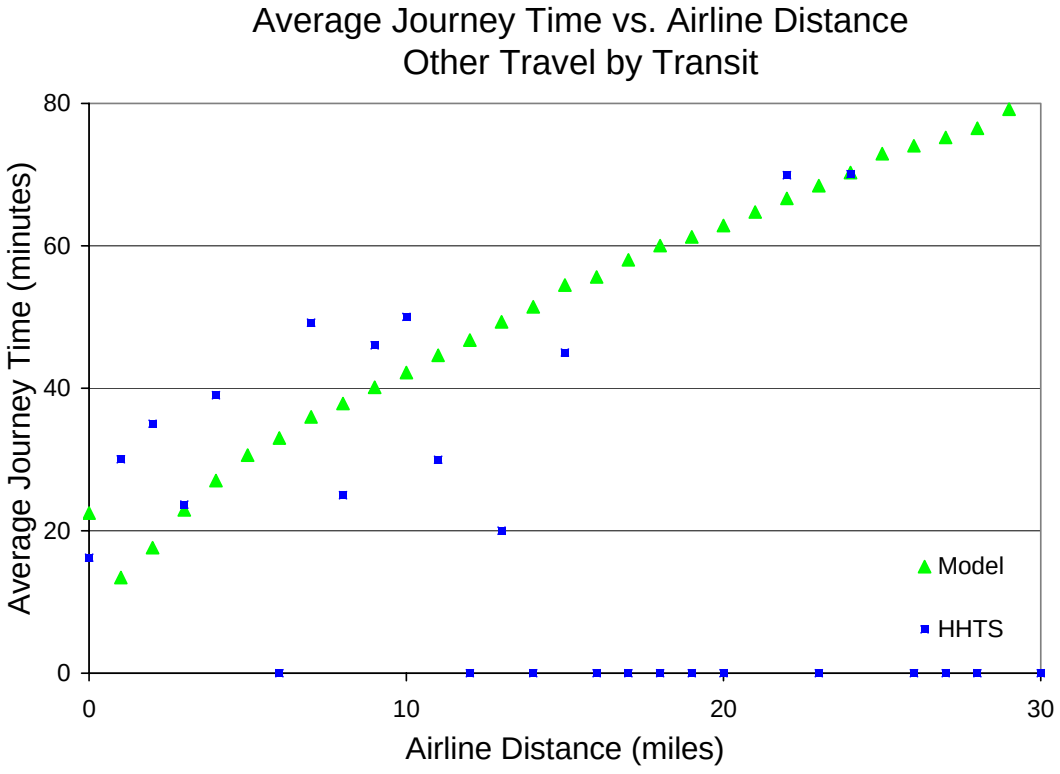

Average Journey Time vs. Airline Distance
Home-to-Work Travel by Auto
Average Journey Time (minutes)
Airline Distance (miles)
0
20
40
60
80
0
10
20
30
Model
Census
HHTS

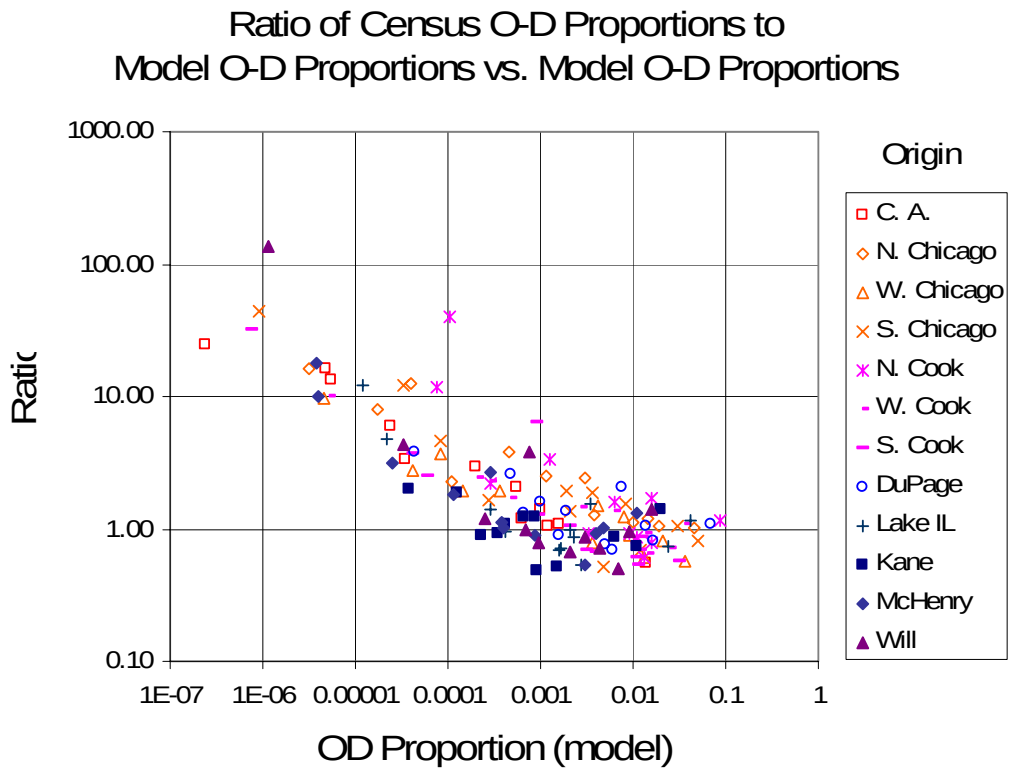

Average Journey Time vs. Airline Distance
Home-to-Work Travel by Transit
Average Journey Time (minutes)
Airline Distance (miles)
0
20
40
60
80
0
10
20
30
Model
Census
HHTS

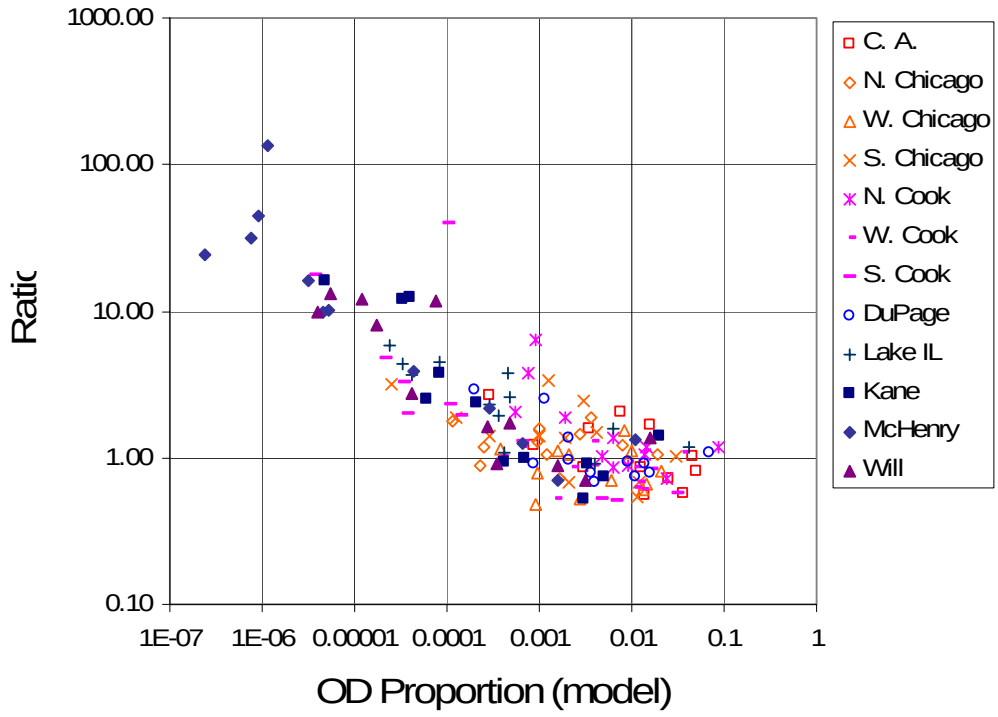

OD Proportion vs. Airline Distance
Other Travel by Auto and Transit
OD Proportion
1
0.199
0.1
0.053
0.01
0.001
0.0001
0.00001
Model
HHTS
0
10
20
30
Airline Distance (miles)

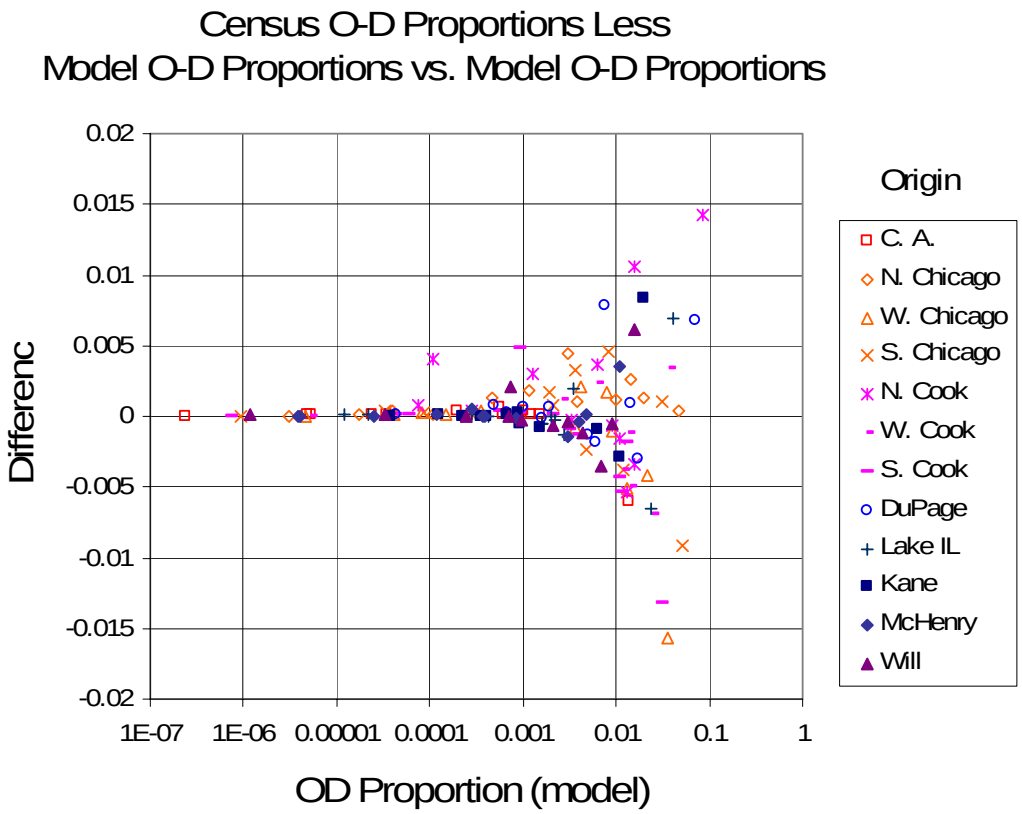

Transit Share vs. Airline Distance
Other Travel
Transit Share
1
0.189
0.1
0.01
0.001
Series Regional Share
Model - 0.189
HHTS - 0.069
0
10
20
30
Airline Distance (miles)

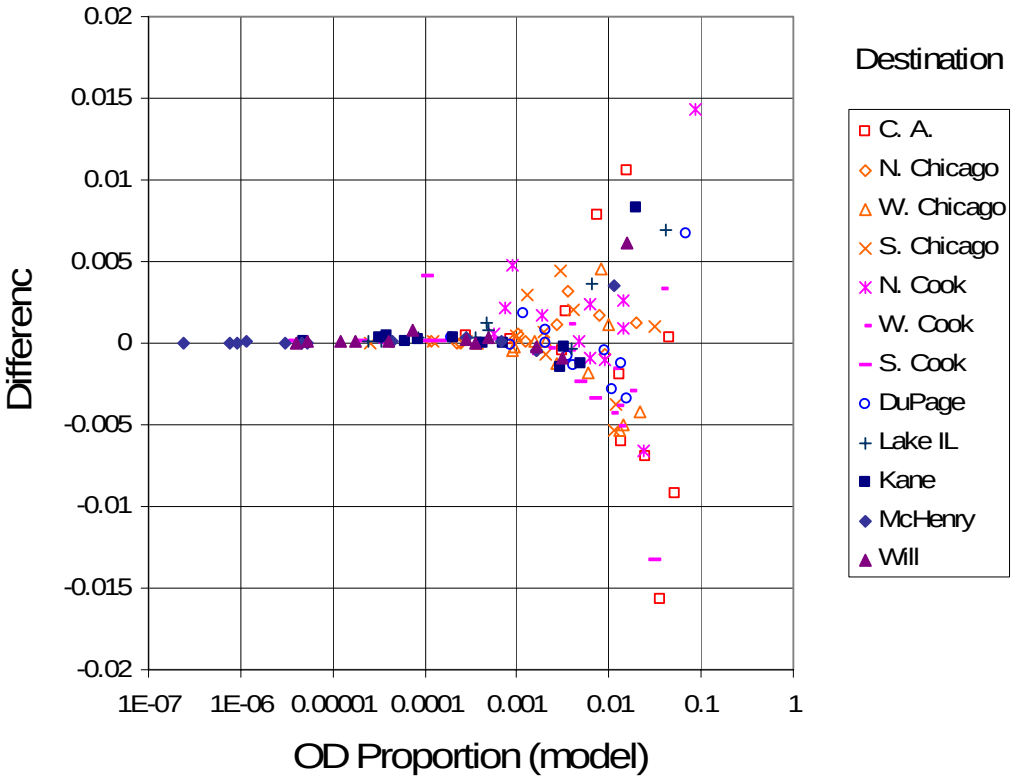

Average Journey Time vs. Airline Distance
Other Travel by Auto
Average Journey Time (minutes)
Airline Distance (miles)
0
20
40
60
80
0
10
20
30
Model
HHTS


Average Journey Time vs. Airline Distance
Other Travel by Transit
Average Journey Time (minutes)
Airline Distance (miles)
0
20
40
60
80
0
10
20
30
Model
HHTS


Ratio of Census O-D Proportions to
Model O-D Proportions vs. Model O-D Proportions
1000.00
100.00
10.00
1.00
0.10
Ratio
1E-07
1E-06
0.00001
0.0001
0.001
0.01
0.1
1
OD Proportion (model)
Origin
C. A.
N. Chicago
W. Chicago
S. Chicago
N. Cook
W. Cook
S. Cook
DuPage
Lake IL
Kane
McHenry
Will


1000.00
100.00
10.00
1.00
0.10
Ratio
1E-07
1E-06
0.00001
0.0001
0.001
0.01
0.1
1
OD Proportion (model)
C. A.
N. Chicago
W. Chicago
S. Chicago
N. Cook
W. Cook
S. Cook
DuPage
Lake IL
Kane
McHenry
Will


Census O-D Proportions Less
Model O-D Proportions vs. Model O-D Proportions
Differenc
0.02
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
-0.02
1E-07
1E-06
0.00001
0.0001
0.001
0.01
0.1
1
OD Proportion (model)
Origin
C. A.
N. Chicago
W. Chicago
S. Chicago
N. Cook
W. Cook
S. Cook
DuPage
Lake IL
Kane
McHenry
Will

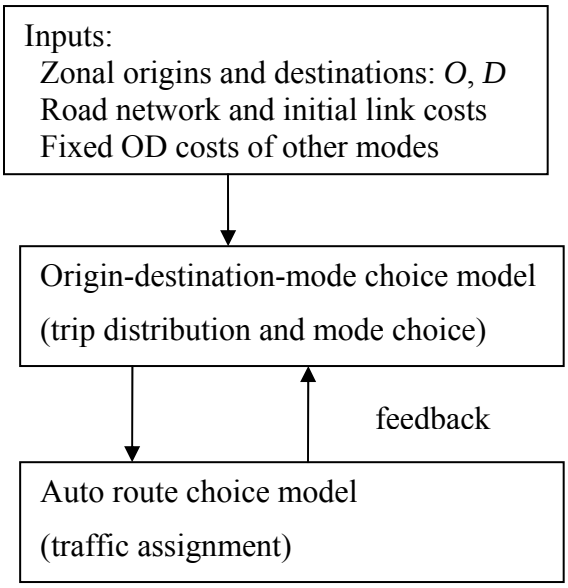

Differenc
0.02
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
-0.02
1E-07
1E-06
0.00001
0.0001
0.001
0.01
0.1
1
OD Proportion (model)
Destination
C. A.
N. Chicago
W. Chicago
S. Chicago
N. Cook
W. Cook
S. Cook
DuPage
Lake IL
Kane
McHenry
Will

6.5 Solving the Sequential Travel Forecasting Procedure with Feedback

The seemingly simple notion of solving the Sequential Procedure (also known as the four-step procedure) with feedback is actually quite a difficult problem, both conceptually and computationally. The objective of this paper is to set forth the issues related to this matter in terms that are familiar and understandable to practitioners. Since a rigorous, detailed understanding requires the use of mathematical notation and relationships, the use of words and flowcharts to describe these matters is an immediate challenge. An alternative to solving the Sequential Procedure with feedback is to solve an integrated or combined model that represents the equivalent choices. Although we consider the solution of an integrated model to be superior, this approach is not available to many practitioners using existing travel forecasting software systems. For a review of the status of these integrated models, see Boyce and Bar-Gera (2004).

6.5.1 The Feedback Criterion

The first question that must be addressed and settled concerns what is the intended result of solving the procedure with feedback. That is, how do we know when the desired solution has been achieved? To discuss this question, we consider a version of the Sequential Procedure in which there are multiple modes, but only one mode has costs that increase with flow; the others have fixed costs. Therefore, only one route choice (assignment) problem is solved. (For a model with congested costs also on the transit modes, see de Cea et al (2003).) Furthermore, we assume that origin and destination (production and attraction) inputs to the trip distribution model are fixed. The initial solution of the origin-destination-mode (ODM) choice model results in a trip table, which depends upon the initial road link costs and the fixed inputs. Assignment of the auto trip table to the road network results in an array of link flows and their associated link costs. If this link cost array is equal to the initially assumed link costs, this solution may be said to be in equilibrium with regard to ODM choice and route choice. This condition would be met, for example, if the link costs were fixed. The basic idea is illustrated by the following flowchart.



Although the above statement may be considered to be acceptable, it does contain an implicit assumption, namely that the origin-destination-auto costs can be determined from the link costs. One way in which these costs can be determined is to assume that origin-destination costs are equal to the costs of minimal cost routes. As we shall see later, other possibilities might be assumed. If we assume in addition that the initial link costs are based on zero link flows, then minimal cost routes would be suitable, since there are no route flows in the network. Assuming that the equality of link costs is not immediately satisfied, we require a condition that is independent of the initially assumed link costs. Two ways to state this condition are:

1. The usual procedure is to assume that the trip table is determined by an array of road link costs $c(f)$, which depends on a link flow array f. If this trip table is assigned to the road network by an unspecified method, and results in link flows $\hat{f}$ with costs $c\left(\hat{f}\right)$ that are equal to those on which the trip table was based, $c(f)$, then we can say that this Sequential Procedure is in equilibrium.

2. A more general statement would include the solution of the ODM choice step as well as the traffic assignment step. Suppose the ODM choice step depends on arrays of previously derived link flows, which in an unspecified manner, produces a trip table d. Suppose that trip table d is assigned to the road network by an unspecified method, resulting in a link flow array, and we use that array and the ODM choice model to produce a new trip table $\hat{d}$, which depends directly on that array, as well as other fixed inputs common to both trip tables. We can now ask whether these two trip tables, d and $\hat{d}$, are equal. If they are, then the Sequential Procedure is in equilibrium. In this second case, we have also not made any assumption about how either model is solved. However, we do claim that if the equilibrium link flow array can produce a trip table equal to the current trip table, then we have found an equilibrium solution. We return to this question in 6.5.7.

As a result of these two statements, we now have an understanding of what criteria must be satisfied to achieve the condition that is generally called, "solving the Sequential Procedure with feedback." Next we consider procedures for generating trip tables and link flow arrays that satisfy these conditions.

6.5.2 Types of Feedback

Several ways to solve the Sequential Procedure with feedback have been defined and applied. We categorize them generally as follows:

1. Procedures that do not involve averaging of either trip tables or link flow arrays. For short, we describe these as naïve feedback methods, without intending to be pejorative.

2. Procedures based on the formulation of the Sequential Procedure as a convex programming problem or a fixed point problem. In these special cases, algorithms can be constructed and their convergence to satisfy the above criterion proved. Examples are algorithms proposed by Evans (1976) and Bar-Gera and Boyce (2003).

3. In cases in which a convex programming problem is not known to exist, one can appeal to methods related to optimization problems, and apply the Method of Successive Averages (MSA) or even Constant Weights.

6.5.3 A Naïve Feedback Procedure

Discussions with numerous practitioners regarding how they solve the Sequential Procedure with feedback led to the following statement, as represented in Fig. 1:

1. Solve the ODM choice model by using an array of initial link costs $c(0) = c[f(0)]$ to determine the costs of the minimal cost routes, resulting in the trip table $d(1)$; the initial link flows $f(0)$ are often assumed to be zero.

2. Assign trip table $d(1)$ to the road network, resulting in link flows $f(1)$. Compute the link costs $c(1) = c[f(1)]$ that depend upon $f(1)$, and solve for the minimal cost routes.

3. Using the link cost array $c(1)$, re-solve the ODM choice model and re-assign the auto part of the resulting trip table $d(2)$ to the road network, resulting in $f(2)$ and $c[f(2)]$. In some variants, link speeds are used in place of link costs; moreover, costs or speeds from successive feedback *loops* are sometimes averaged in an attempt to dampen oscillations.

4. Repeat step 3 a predetermined number of times n. The final trip table $d(n)$ and link flow array $f(n)$ are the answer to the Sequential Procedure.

What conclusions can one draw about this feedback procedure? First, since the final result depends upon congested link flows and link costs, it is likely to be preferred to the initial solution based on the assumed link flows. Second, the procedure does not invoke any test to determine whether the final trip table $d(n)$ is in equilibrium with the final link costs $c(n)$, so one cannot claim that the procedure has reached equilibrium. Third, what one can conclude is that the ODM choice and route choice models have been solved several times with different assumptions about the input link flow array, $f(0)$, $f(1)$, $f(2)$, … $f(n-1)$? One cannot conclude anything about the validity of these assumptions, or about the equilibrium solution.

6.5.4 Procedures Based on Integrated (or Combined) Optimization Formulations

The first successful attempt to devise a travel forecasting model, by Martin Beckmann and his co-authors (1956), proposed a single, rather general model of origin-destination choice and route choice. That is, they did not propose to solve a sequence of models, but included both OD choices and route choices over the network in a single, integrated model. Beckmann did not devise a way to solve the model; indeed, he has stated that he never imagined that it would be solved for large networks and zone systems (Boyce, 2004).

His model was so advanced for its time that virtually no one understood its significance. Nearly 20 years passed before a connection was identified between Beckmann's formulation and the Sequential Procedure devised by early practitioners about the same time. The person who noticed this connection most clearly was Suzanne Evans in her 1973 Ph.D. thesis; she devised a

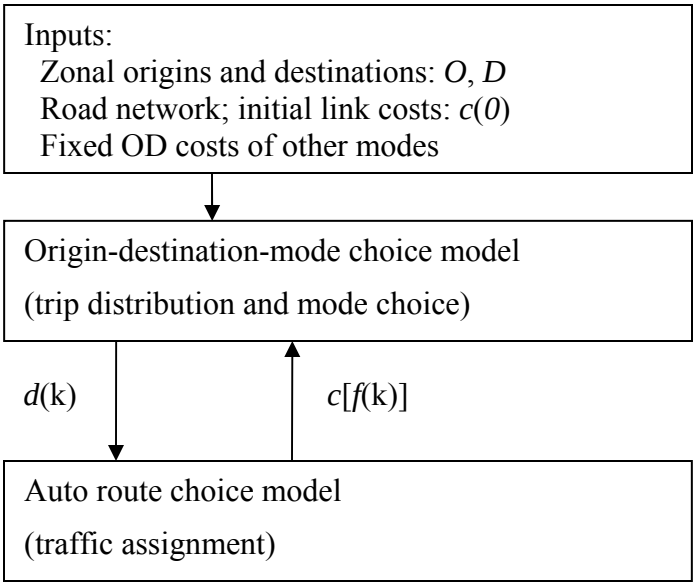


Fig. 1 A Naïve Feedback Procedure

procedure for solving the problem, called an algorithm, which has since been shown to be somewhat practical for large networks (Boyce and Bar-Gera, 2003, 2004). By then, however, the Sequential Procedure was so ingrained in the thinking of both academics and practitioners that hardly anyone noticed that Beckmann's model, and its descendents, could now be solved.

From the vantage point of formulation of Beckmann and the algorithm of Evans, the answer to the feedback question is immediately apparent. But first, one should try to understand the implications of his formulation. The unknown variables of the formulation are trip tables d and link flow arrays f, just as in the above representation of the Sequential Procedure. Beckmann's formulation was stated as a constrained optimization problem, which simply means the minimization of a function of d and f, subject to some definitional constraints. Let us denote this function as $F(d, f)$, which we assume to be convex, meaning that it is generally U-shaped and therefore has a unique minimum. The definitional constraints are that the OD-auto flow equals the sum of the route flows for each OD pair, and that the route flows are greater than or equal to zero. Through the analysis of this problem, one learns some facts about the conditions that describe the desired equilibrium, where "e" following a variable symbol in parentheses denotes the optimal, or equilibrium, solution:

1. The equilibrium trip table $d(\mathrm{e})$ is determined by the ODM choice model based on link costs $c(\mathrm{e})$, determined by the equilibrium link flows $f(\mathrm{e})$, as well as the given zonal origins and destinations, O and D, assuming the functional form of the model corresponds to this framework. The primary candidates in this regard are the several variants of the logit model. Thus, the solution of Beckmann's formulation contains the statement of the equilibrium condition for ODM choices, which the Sequential Procedure is intended to fulfill.

2. The equilibrium route choices that determine $f(\mathrm{e})$ have the user-equilibrium properties often attributed to Wardrop, which were independently described and investigated by Beckmann:

for each OD pair, at equilibrium all used routes have equal and minimal travel costs. Therefore, the OD table is based on travel costs corresponding to minimal cost routes on which all flows occur.

3. In a practical sense, the problem must be solved in terms of real-valued variables (sometimes called floating pointing arithmetic) and not in terms of integers. This requirement is necessary for achieving the desired convergence of the solution, since restricting the variables to integers is impractical. The resulting values may be readily interpreted as average weekday flows, and do not diminish the validity of the model. Moreover, integer-valued representations of the flows may be extracted by drawing a sample from the flows.

6.5.4.1 The Evans Algorithm

Evans proposed an algorithm for solving this model, which looks very similar to the Sequential Procedure, but with some critical differences. By understanding her solution algorithm, one may gain certain insights into how to solve properly the Sequential Procedure with feedback. A generalized version of her algorithm is represented by the flowchart in Fig. 2. Before discussing the algorithm, note the initial solution ($d(1)$, $f(1)$) is found by setting $\lambda(1) = 1$, so that $d(1) = \hat{d}(1)$ and $f(1) = \hat{f}(1)$. The trip table d and link flow array f are called the main-problem variables, whereas $\hat{d}, \hat{f}$ are sub-problem, or auxiliary, variables, meaning that they are found by holding the main-problem variables temporarily fixed. The Evans Algorithm differs from the Naïve Procedure in three important respects. First, an averaging step is added, with a weight (also called a step size) determined by a *line search* based on an objective function F defined on $d, \hat{d}, f$ and $\hat{f}$. This averaging step is crucial in finding the equilibrium solution, and provides a strong hint that some kind of averaging of the *solution variables* is necessary to solve the Sequential Procedure with feedback. Second, a convergence check is added, which provides a well-defined stopping criterion. Third, the route choice model (assignment) is specified as a simple assignment of the total OD-auto flow to the minimal cost route for each OD pair, commonly known as all-or-nothing (AON) assignment. The averaging step assures the convergence of the algorithm.

In general terms, then, the Evans Algorithm consists of solving a sequence of trip tables and AON assignments, which are averaged together so as to drive the solution to the equilibrium defined by minimizing the objective function F[d, f]. Note that d as well as f is required for the averaging, even though only the link flow array f may be required as an output. If the trip table is not needed for other analysis purposes, it is discarded following the termination of the solution algorithm. This point is important for comparing the Evans Algorithm with other solution procedures.

As noted, the Evans Algorithm finds the solution in terms of link flows as well as OD flows. Route flows are not determined; therefore, the assignment portion of the algorithm may be termed *link-based*. Evans's assignment procedure is effectively the same as the link-based algorithm for solving the assignment problem with a fixed trip table, also proposed in the early 1970s based on Frank and Wolfe (1956).

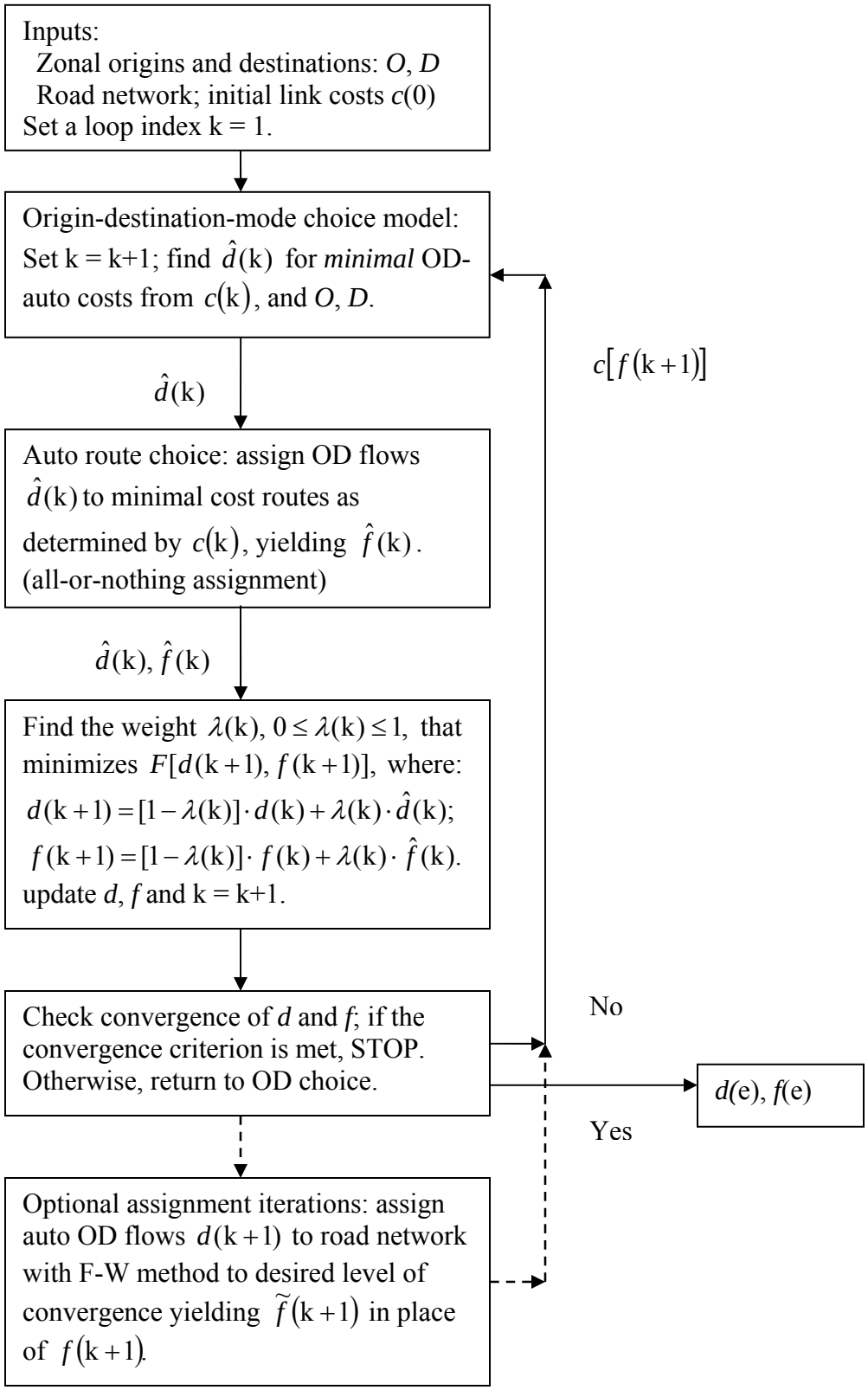


Fig. 2 A Generalized Evans Algorithm

Although the Evans Algorithm may be considered to be useful for practice, its convergence is relatively slow; however, it is definitely preferred to the Naïve Procedure. Reasons for the poor performance of the algorithm are the slow convergence of the assignment method plus the need to adjust the trip table following each all-or-nothing assignment and line search.

Practitioners generally do not perform a single AON assignment of $\hat{d}(k)$, the trip table found at loop *k*. Instead, they perform a user-equilibrium assignment of $\hat{d}(k)$ using the Frank-Wolfe (F-W) method. Based on limited experimentation, we can suggest an alternative procedure that may avoid specific convergence problems related to this approach: a) perform the averaging step of the Evans Algorithm yielding $d(k+1)$ and $f(k+1)$; b) perform a user-equilibrium assignment of the auto portion of the main-problem trip table $d(k+1)$ to the road network to the desired level of convergence yielding $\tilde{f}(k+1)$; c) replace $f(k+1)$ with $\tilde{f}(k+1)$ and proceed to the OD-mode choice step. For $k = 1$, this option simply replaces the AON assignment with sufficient iterations of user-equilibrium assignment to reach the desired convergence level. Subsequent iterations may refine this solution more quickly than performing only AON assignments. The optional additional assignment step is shown near the bottom of the Evans Algorithm in Fig. 2.

6.5.4.2 Origin-Based Assignment Algorithm

Following Evans's contribution, two *route-based* algorithms were proposed, tested and implemented for assigning fixed trip tables to large networks: Bothner and Lutter (1978) and Larsson and Patriksson (1992). These algorithms were the first attempts to represent the solution of the assignment problem in terms of route flows rather than link flows.

In the late 1990s a new assignment algorithm, called Origin-Based Assignment (OBA), was invented by Hillel Bar-Gera. This algorithm was initially motivated by the idea of representing the route flow solution in a more efficient way. Initially implemented for solving the route choice problem for a fixed trip table, Bar-Gera later adapted his algorithm to solve an integrated model of OD, mode and route choice. The algorithm was implemented on the large-scale zone system and road network of Chicago, showing remarkably precise and fast convergence, as compared with the Evans Algorithm and the Sequential Procedure with feedback; see the charts in Bar-Gera and Boyce (2003). A flowchart for his algorithm is shown as Fig. 3.

By adopting a different representation of the assignment on a subnetwork defined for each origin zone, Bar-Gera's Origin-Based Assignment Algorithm computes the *average* travel cost for each OD-auto pair without explicitly storing route flows. It accomplishes this feat by solving the assignment problem in terms of origin-specific approach proportions, the proportion of flow from a given origin to a node via a given approach link. The approach proportion representation has two important implications for this algorithm:

1. A previous assignment can be updated when the trip table is adjusted to reflect changes in the OD auto costs;

2. The OD-auto choices are based on the average OD auto cost, rather than on the minimal OD auto cost, as in the Evans Algorithm and the Sequential Procedure with feedback.

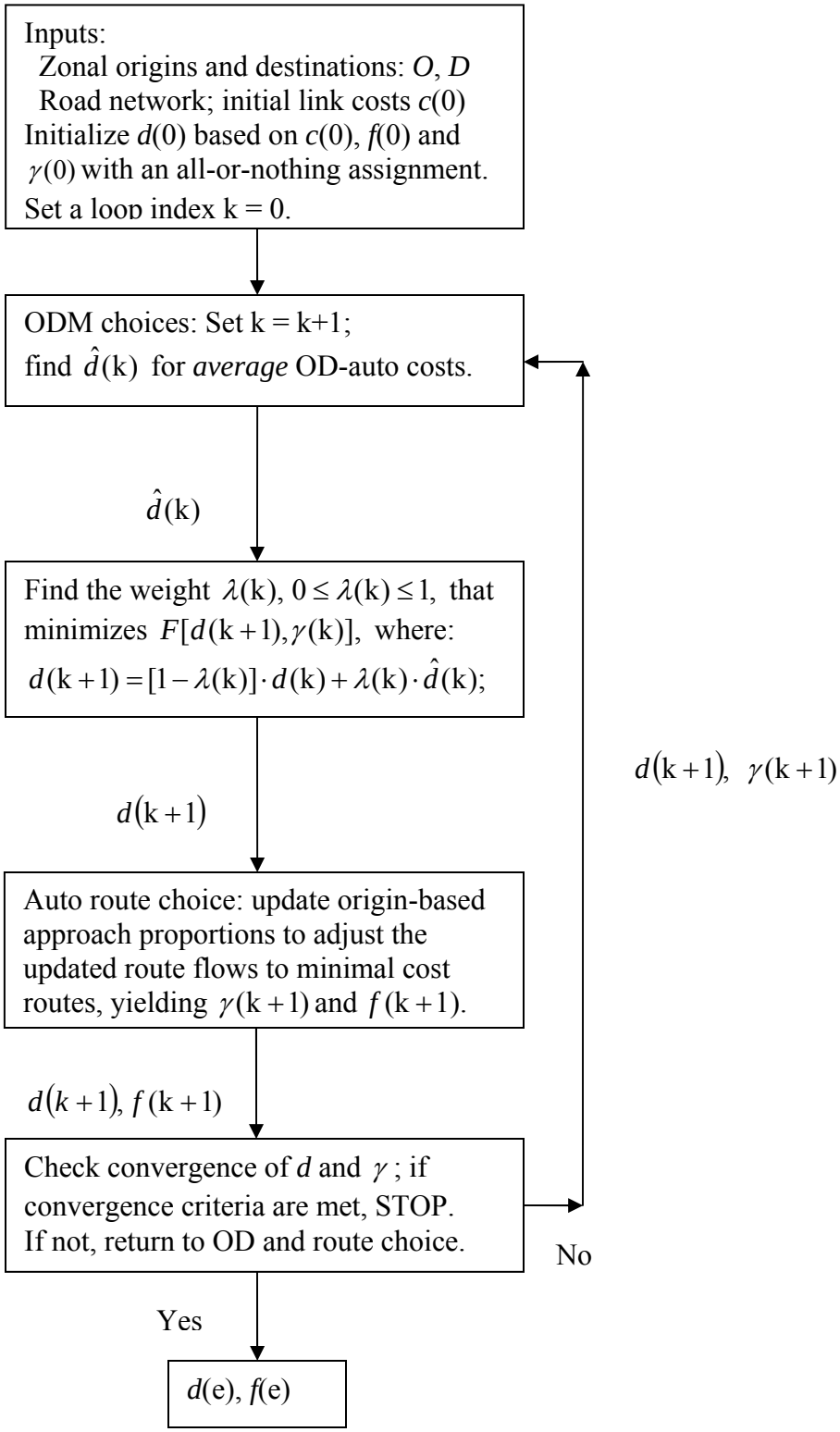


Fig. 3 The Origin-Based Assignment Algorithm for an Integrated Travel Choice Model

These advantages not only permit the algorithm to converge faster, but allows it to approach much closer to the true equilibrium. These advantages may also be true for some route-based algorithms, depending upon how they are implemented.

The solution procedure is initialized by finding the trip table *d* based on the assumed link costs. The trip table is then assigned so as to place all OD-auto flows on user-equilibrium routes of equal and minimal cost, initially in an approximate manner. Based on this initial assignment and its associated OD-auto costs, the ODM choice function is applied to update the trip table, holding fixed the route proportions $\gamma(k)$. Thus the computation of $\hat{d}(k)$, $\lambda(k)$ and $d(k+1)$ are based on $\gamma(k)$. After this update is completed, new route proportions are determined. During the update of route proportions from $\gamma(k)$ to $\gamma(k+1)$, $d(k+1)$ is held fixed.

Several important distinctions between the Origin-Based Assignment Algorithm and the Evans Algorithm should be noted. First, the solution proceeds by refining the trip table and the associated route proportions. The link flows do not enter into the computations, except in determining the link costs. Second, the auto route proportions replace the auto link flows as the principal variable representing flows on the network. Because these are stored implicitly and compactly as origin-based approach proportions, the algorithm quickly reaches a precise solution, as compared with procedures based on link flows.

6.5.4.3 Monitoring the Convergence

The description of the Evans and Origin-Based Assignment Algorithms is incomplete without specifying the convergence criteria. Although the criteria are related for the two algorithms, they are somewhat different. Moreover, the definitions for these algorithms provide guidance for monitoring the convergence of related procedures. Two convergence criteria are needed, one for the trip table and another for the link flow array.

These criteria are often combined into a single criterion called the Relative Gap. Feasible solutions to convex optimization problems have an associated Lower Bound, which is defined on $\hat{d}$, f and $\hat{f}$ for the Evans Algorithm and $\hat{d}, \gamma$ and $\hat{\gamma}$ for the Origin-Based Assignment Algorithm. The Relative Gap (*RG*) is formed from the lowest objective function value (*LOF*), which is always from the current loop, and the Best Lower Bound (*BLB*), which may be from a previous loop: $RG = (LOF - BLB)/BLB$. Because this paper does not provide details in equation form, the specification here is limited to verbal descriptions. Detailed specifications may be found in Bar-Gera and Boyce (2003). The Lower Bound contains terms related to the above variables, which can be given an intuitive interpretation. For the trip table variable, the Objective Function and Lower Bound corresponding to the logit function contain entropy terms defined on the trip table.

For the trip table, we consider a simple criterion alluded to at the outset of the paper. This criterion is the *Total Misplaced Flow*, which is the sum of the absolute differences of zone-to-zone ODM flows in the main problem and subproblem solutions, d and $\hat{d}$. If these two tables are equal, then the algorithm has converged with regard to the trip table.

For the route and link flow portion of the problem, we begin with a statement based on the intuitive notion of the *Total Excess Cost*. As noted, at user-equilibrium all used routes have equal and minimal travel cost. Therefore, there is no excess cost in the network in the sense that for each OD pair, no used route has a higher cost than the minimal cost route. For non-equilibrium solutions, however, flows with excess costs are present on some routes. Therefore, for each OD pair, we determine the difference between each used route's cost and the minimal cost, weighted by the used route flow, and sum over all used routes for that OD pair to obtain the OD-auto excess cost. If we sum this quantity for all OD pairs, then we obtain the Total Excess Cost in the network.

This quantity is directly computed by the Origin-Based Assignment Algorithm, as well as route-based algorithms. Remarkably, it can also be computed for any feasible link-based flow solution as well. To do so, one computes the link travel costs associated with the feasible link flow solution, and performs an all-or-nothing assignment, given these costs. The total travel cost of these two solutions may be computed separately by multiplying each link flow times the link cost, and summing over all links. The total cost of the feasible solution less the total cost of the all-or-nothing solution, evaluated as stated, is the Total Excess Cost. At equilibrium, the Total Excess Cost equals zero because the minimal OD costs in the all-or-nothing solution are the same as the equal and minimal route costs in the equilibrium solution.

These two criteria, one for trip tables and one for link flows, provide intuitive and useful measures for evaluating any feasible solution, that is any solution that satisfies the origin inflows and destination outflows, as well as that the sum of route flows equal OD-auto flows for each OD pair. To my knowledge, these criteria are not generally applied by practitioners.

6.5.5 Procedures for Integrated Models Not Having an Optimization Formulation

The above explanation is well and good for the limited case of formulations with an equivalent convex optimization formulation. But what about other formulations, and especially procedures applied in practice?

For about 25 years it has been generally understood that the weights yielded by the line search step of the above algorithms can be replaced by predetermined weights, based on the properties of the Method of Successive Averages (MSA). This method becomes useful when no objective function F(d, f) is available. The weights $\lambda(\mathrm{k})$ generally applied with MSA are 1/k, where k is the loop number. The sequence of sub-problem weights generated by this method, then, are 1, 1/2, 1/3, 1/4, 1/5, … . The application of this sequence can be proven to converge to the equilibrium solution, but the convergence may be quite slow.

Recent computational experiences revealed that using a Constant Weight instead of the decreasing weights of MSA may yield a substantially improved result (Bar-Gera and Boyce, 2006). In this case, the values of (1/k) in Fig. 4 are replaced by a constant weight between 0 and 1; typically a value near 0.5 was found to work well, but a range of weights should be tried to determine the best one for a given network and scenario. In the experiments reported in the next section, both MSA weights and Constant Weights were applied.

A flowchart illustrating this procedure for link-based, route-based and origin-based user-equilibrium assignment is shown as Fig. 4. In this procedure, only the successive trip tables are averaged with the MSA weights, and the resulting auto trip tables are assigned to the road network to determine the link flows and minimal or average OD auto travel costs needed for feedback. In this scheme, convergence of the feedback procedure is checked with the trip table only; convergence of assignments is checked at the assignment step within each feedback loop.

6.5.6 Findings of Preliminary Computational Experiments

Computational experiments to compare Naïve Feedback, and feedback with the Method of Successive Averages and Constant Weights were performed with a model implemented by Wolfgang Scherr, PTV America, from the Capital District Transportation Committee, Albany, NY, in the PTV software system VISUM. The model has five classes of travelers with a total peak period OD flow of 313,082 persons per hour.

The Sequential Procedure for this model was solved using the Naïve Feedback procedure, with a Relative Gap (convergence level) for the Assignment step of 0.0001. The model was also solved using the procedure described in Section 6.5.5 for models not having an optimization formulation. Two solutions were obtained, one using the Method of Successive Averages, and the other using a Constant Weight (CW) equal to 0.6 for d and 0.4 for $\hat{d}$. In both cases, the convergence of the Assignment step was set to 0.0001.

The convergence of the three solution methods were compared using the procedure defined in Section 6.5.4.3, Monitoring the Convergence, for the Total Misplaced Flow. Each assignment converged to a Relative Gap of slightly less than 0.0001; therefore, it is not useful to monitor its convergence. However, the number of assignment iterations needed to reach this Relative Gap varied substantially between the Naïve method and the other two methods.

Total Misplaced OD Flows for the three solutions are shown in Fig. 5 using the log scale to facilitate their comparison. For Naïve Feedback, TMF reached a minimum value of 34,200 persons/hour in the 7th loop, which shows a large discrepancy between the trip table in the previous feedback loop and the trip table computed based on the travel times resulting from assigning that table in the current loop. At full convergence, TMF must equal zero, which means that a link flow array has been found whose minimum route generalized travel costs produce a trip matrix equal to the trip table that yielded those links flows. A value of zero persons/hour was achieved by the Constant Weight (CW) method in the 17th loop, and a value of less than 100 was reached in the 8th loop. The MSA method reached a minimum TMF of 127 persons/hour in the 20th loop when the process was terminated. The values are acceptably small for the MSA and CW methods.

The number of assignment iterations in each loop required to achieve a Relative Gap of less than 0.0001 is shown in Fig. 6. For Naïve Feedback, the number of assignment iterations falls to 12 in loop 6, but then increases gradually to 27. For the other two methods, the number of assignment iterations falls rapidly to the minimum number of 2. In the route-based assignment algorithm applied in VISUM, the assignment is updated following the revision of the OD table. If a link-based method were used, then a new assignment solution is required in each feedback loop.

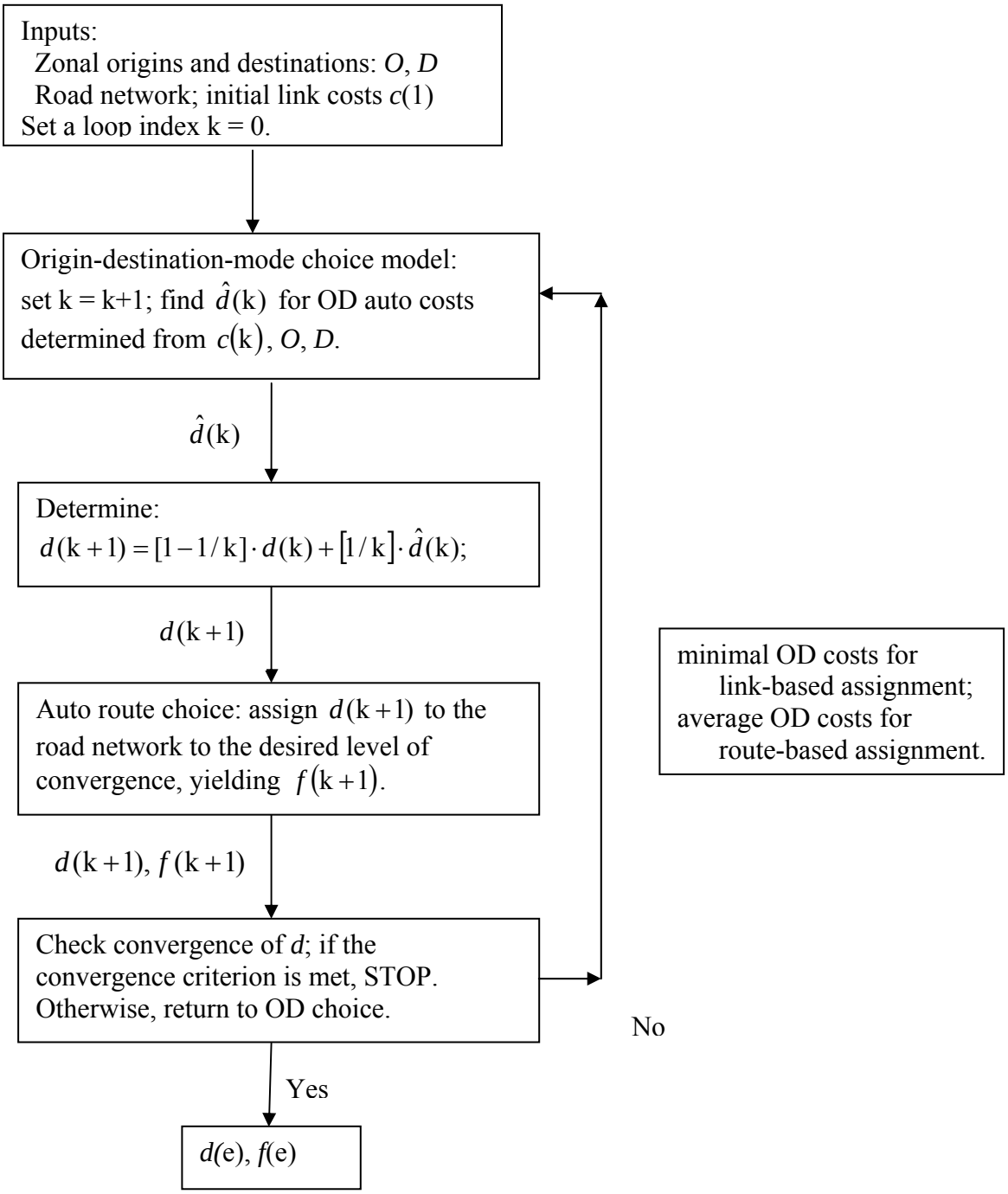


Fig. 4 Feedback Procedure with Fixed Weights

Fig. 5. Comparison of Feedback Methods - Misplaced OD Flow

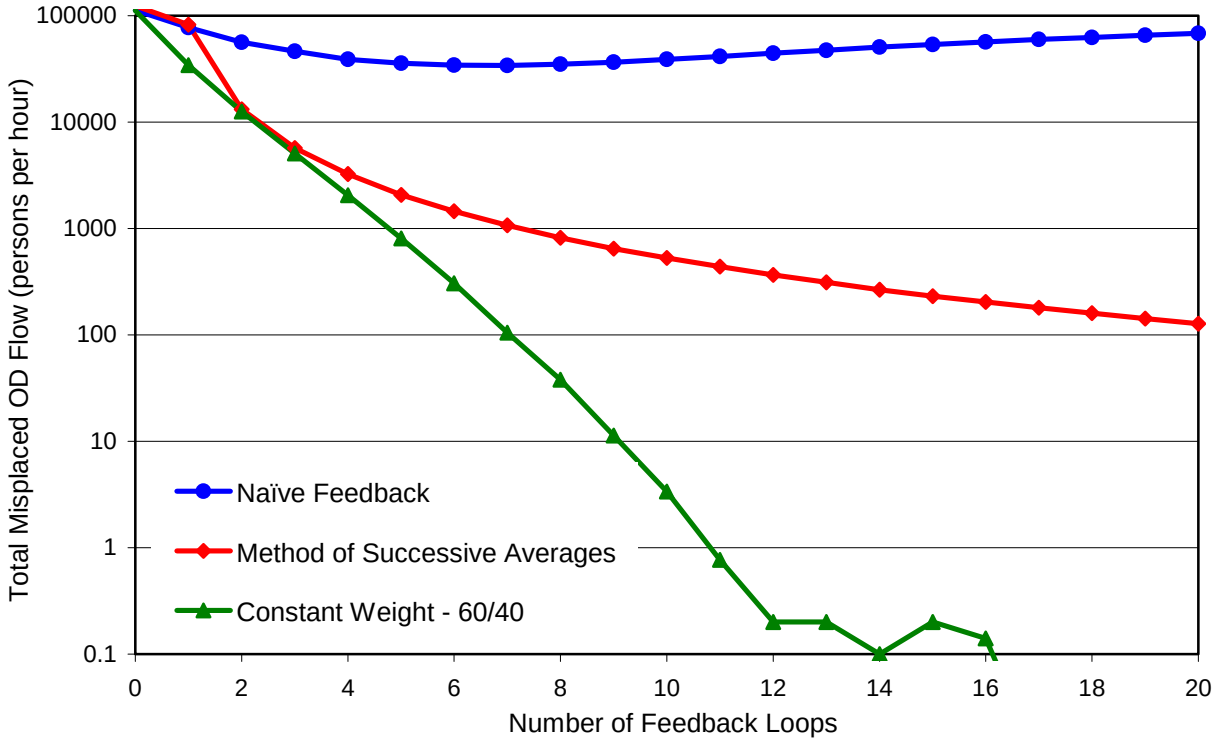
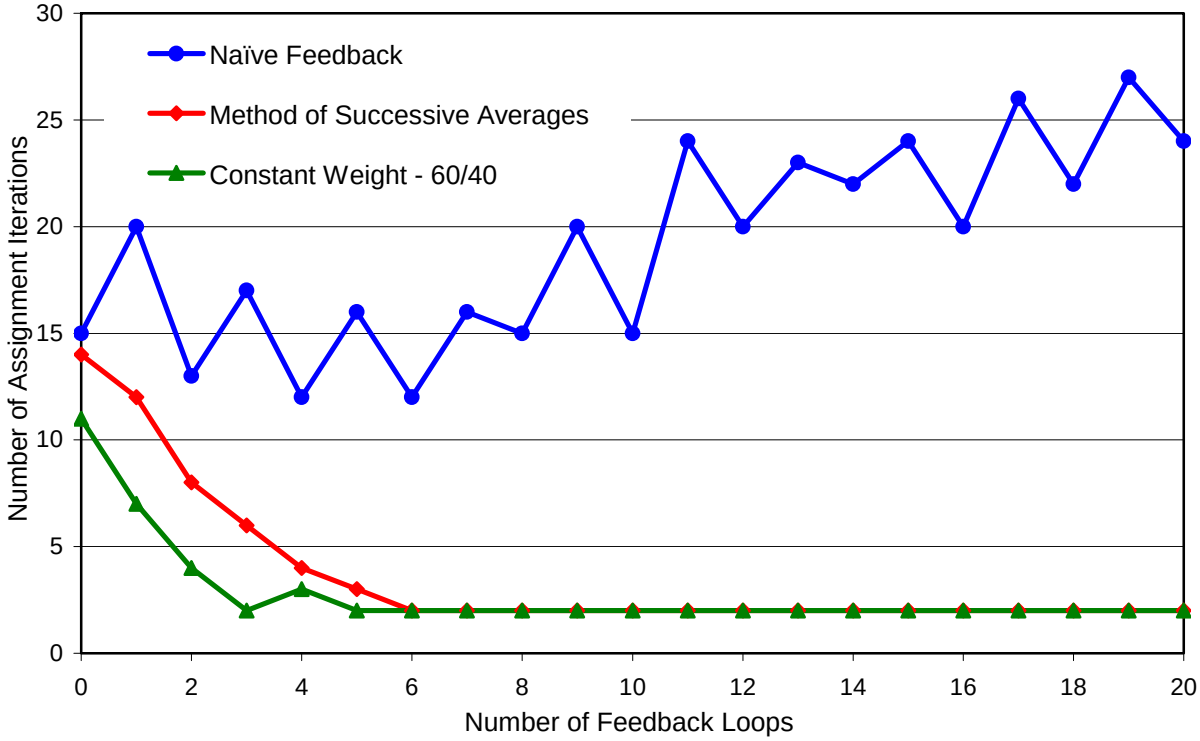


Fig. 6. Comparison of Feedback Methods - Assignment Iterations



These results, as well as the extensive studies of the convergence of integrated models cited above, show that forecasts of origin-destination flows and link flows depend upon solving the sequential procedure properly with respect to travel times. In congested cities with several interacting modes, representing the effects of congestion properly is bound to affect the forecast of origin-destination flows and link flows in a profound manner. In less congested cities, as found in some developed countries and in smaller and more dispersed cities, the effects of congestion may not be so important.

Results from a link-based assignment procedure are being sought to complete these comparisons.

6.5.7 Recommendations for Practice

Based on the above explanations, what can we offer as recommendations for the near term and the more distant future?

1. What has been described here as Naïve Feedback is relatively ineffective for solving the Sequential Procedure with feedback, and should not be used. Just how ineffective it is for large-scale applications remains unknown.

2. Performing any type of feedback computations is really only useful if measures of convergence are also computed. Otherwise, the practitioner has no basis for deciding how many feedback loops to perform, or how to compare solutions for different networks and models. These measures are easy to compute, as explained above.

3. For link-based assignment algorithms, feedback procedures applied in practice (other than Naïve Feedback) sometimes involve averaging of the link flow array from feedback loop to feedback loop using MSA type weights. While this procedure is likely to be preferable to Naïve Feedback, its convergence should be monitored closely to determine whether additional loops are useful.

4. The way in which OD costs are determined from link costs should be compatible with the way that OD flow is assigned to the road network:

   a) if the OD flows are assigned to minimal cost routes, as in link-based assignment, then *minimal* OD costs should be used in the feedback procedure;

   b) if route proportions are determined to allocate OD flows, as in origin-based and route-based assignment, then *average* OD costs should be used in the feedback procedure.
5. Route-based and origin-based assignment algorithms, as integrated into a larger choice model, appear to have definite advantages for finding the equilibrium solution of integrated models of OD, mode and route choice. In addition to the availability of average OD-auto costs, these algorithms allow route and link flows to be updated each time the trip table is adjusted. Therefore, in this procedure trip tables are averaged together using MSA, a constant weight, or a weight from a line search, as appropriate.

6. Computational experiments with appropriate measures of convergence are being performed for several test networks and related input data for all of the options discussed in this paper, so that differences in the quality of results from various procedures can be better understood.

Acknowledgement – The preliminary examples provided by Wolfgang Scherr, PTV America, using PTV's *VISUM* software system with data from the Capital District Transportation Committee of New York are gratefully acknowledged. Comments of practitioners too numerous to mention are appreciated.

6.5.8 References on Feedback

Bar-Gera, H. (2002) Origin-Based Algorithm for the Traffic Assignment Problem, *Transportation Science* 36, 398-417.

Bar-Gera, H. and D. Boyce (2003) Origin-Based Algorithms for Combined Travel Forecasting Models, *Transportation Research* 37B, 405-422.

Bar-Gera, H. and D. Boyce (2006) Solving a Nonconvex Combined Travel Forecasting Model by the Method of Successive Averages with Constant Step Sizes, *Transportation Research* 37B, 351-367.

Beckmann, M., C. B. McGuire and C. B. Winsten (1956) *Studies in the Economics of Transportation*, Yale University Press, New Haven; also published as Rand-RM-1488-PR, Rand Corporation, Santa Monica, CA, 1955.

Bothner, P. and W. Lutter (1982) Ein direktes Verfahren zur Verkehrsumlegung nach dem 1. Prinzip von Wardrop, Arbeitsbericht Nr. 1, Forschungsbereich: Verkehrssysteme, Universität Bremen, Germany.

Boyce, D. (2004) Forecasting Travel on Congested Urban Transportation Networks: Review and Prospects for Network Equilibrium Models, presented at The Fifth Triennial Symposium on Transportation Analysis, Guadeloupe.

Boyce, D. and H. Bar-Gera (2003) Validation of Multiclass Urban Travel Forecasting Models Combining Origin-Destination, Mode, and Route Choices, *Journal of Regional Science* 43, 517-540.

Boyce, D. and H. Bar-Gera (2004) Multiclass Combined Models for Urban Travel Forecasting, *Network and Spatial Economics* 4, 115-124.

De Cea, J., J. E. Fernandez, V. Dekock, A. Soto, and T. L. Friesz (2003) ESTRAUS: A Computer Package for Solving Supply-Demand Equilibrium Problems on Multimodal Urban Transportation Networks with Multiple User Classes. Presented at the Annual Meeting of the Transportation Research Board, Washington, DC.

Evans, S. P. (1976) Derivation and Analysis of Some Models for Combining Trip Distribution and Assignment, *Transportation Research* 10, 37-57.

Frank, M. and P. Wolfe (1956) An Algorithm for Quadratic Programming, *Naval Research Logistics Quarterly* 3, 95-110.

Larsson, T. and M. Patriksson (1992) Simplicial Decomposition with Disaggregated Representation for the Traffic Assignment Problem. *Transportation Science* 26, 4-17.

6.6 Some Practical Issues

In this section, we discuss some practical issues that must be addressed in the future in order to improve the solution of the four-step procedure with feedback. The discussion is based in part on our experience with transportation planning software systems.

6.6.1 Integer vs. Real Variables

There seems to be a myth in applied transportation modeling/software that the solution variables of our models are integers. That is, O-D flows and link flows are said to be properly defined as integers, and moreover that these integer values must be maintained to insure the integrity of the modeling process. Although we are unaware of the source of this myth, we assert that it is false and contrary to our goal solving travel choice models to equilibrium, that is with feedback. We offer several reasons for our viewpoint.

First, any variable defined as the flow per unit time is by definition a real-valued variable. It is a rate, such as vehicles/hour, not a count of vehicles, even though it may be based on field counts. The same is true for Census and survey data, which after all are based on expanded samples, not population counts.

Second, the treatment of these variables in the model formulation is as real-valued variables. To treat them as integers would pose a much more difficult mathematical problem in terms of its properties and its solution. We can think of only two papers in the literature that attempted to solve these models as integers, and then mainly as an intellectual challenge, not to improve their validity.

Third, the convergence of the solution algorithm depends on the variables being defined as real-valued. Fourth, and most interesting, most travel forecasting software which originated with the Urban Transportation Planning System (UTPS) uses integer-valued variables for O-D-mode flows and link flows. The reason for this choice was a completely practical one; the models could be implemented for large networks on the mainframes of the 1970s and the PCs of the 1980s only with integer-valued variables. Fortunately, this is no longer true; otherwise, we could not hope to actually solve these models with feedback. To our knowledge, older software systems in use in the U.S. still solve the models with integer-valued variables. In our early experience in solving combined models with a code implemented in UTPS, we were unable to obtain convergence beyond 5 to 10 iterations, which is not adequate for today's policy objectives.

6.6.2 Parameter Estimation/Calibration

A significant challenge facing practitioners in solving models with feedback is the question of parameter estimation. In the past, each model was estimated separately, and then the sequence was solved. In the process, it is likely that the travel costs used in the model estimation process were incorrect or invalid, since they were not the equilibrium costs.
A proper model estimation procedure requires that the entire four-step procedure or combined model be solved when parameters are estimated. That is, parameter values should be chosen, and the entire model sequence solved with feedback to evaluate that parameter value. Then the value should be adjusted, the sequence resolved, etc. Adjusting the parameter values during each feedback iteration may result in unintended biases and misleading results. This requirement is another reason that efficient codes are needed to solve the four-step procedure. For additional information on parameter estimation, see Boyce and Bar-Gera (2003).

6.6.3 Travel Forecasting Software

Some comments about the status of transportation planning software are also in order. Among software systems with which we are generally familiar, there are two types:

1. PC derivatives of UTPS and other main frame software, updated in various ways: TRANPLAN, MINUTP and TRIPS, now being marketed as a more comprehensive software system called CUBE http://www.citilabs.com.

2. software systems developed independently, generally from university-based research: EMME/2 http://www.inro.ca/, TRANSCAD http://www.caliper.com/, VISUM http://www.ptv.de/ and most recently, ESTRAUS http://www.mctsoft.com/

3. A more specialized system, SATURN, is a suite of flexible network analysis programs http://www.its.leeds.ac.uk/software/saturn/index.html

Until recently, none of the existing software systems was able to solve the four-step procedure with feedback as a feature of the system; now the PTV software system VISUM has this capability, and others may also. Moreover, ESTRAUS solves a combined model. In 1996 and subsequently, we developed an EMME/2 macro which solved our Chicago combined model. Although it worked properly, it was relatively slow, and in our experience rather tedious to program. Such options need to be incorporated into the main features of the software.

Directly, or indirectly, we have observed that there are many deficiencies of existing software systems, including the question of integer variables discussed above. This statement is particularly true with regard to assignment methods, which are evidently not well understood by some developers. Software developers and also consultants are sometimes required to implement improper methods, simply because clients are misinformed or out of date in their understanding of the problem. Regrettably, university researchers often do not agree on what is proper practice. This situation leaves much to be desired as a scientific basis for transportation planning practice.

References for Further Study (see also Section 6.5.8)

Boyce, D. (2002) Is the Sequential Travel Forecasting Procedure Counterproductive? **ASCE** *Journal of Urban Planning and Development* **128**, 169-183.

Boyce, D. and H. Bar-Gera (2001) Network Equilibrium Models of Travel Choices with Multiple Classes, Chapter 6 in *Regional Science Perspectives in Economic Analysis*, M. L. Lahr and R. E. Miller (eds.), Elsevier Science, Amsterdam, 85-98.

Boyce, D. E. and M. S. Daskin (1997) Urban Transportation, Chapter 7 in *Design and Operation of Civil and Environmental Engineering Systems*, C. ReVelle and A. McGarity (eds.), John Wiley and Sons, New York, 277-341.

Boyce, D. E., L. J. LeBlanc and K. S. Chon (1988) Network Equilibrium Models of Urban Location and Travel Choices, *Journal of Regional Science* **28**, 159-183.

Boyce, D., B. Ralevic-Dekic and H. Bar-Gera (2004) Convergence of Traffic Assignments: How Much Is Enough? *ASCE Journal of Transportation Engineering* **130,** 49-55.

Dafermos, S. C. and F. T. Sparrow (1969) The Traffic Assignment Problem for a General Network, *Journal of Research of the National Bureau of Standards* **73B**, 91-118.

Erlander, S. and N. F. Stewart (1990) *The Gravity Model in Transportation Analysis*, VSP, Utrecht.

Lundgren, J. and M. Patriksson (1998) An algorithm for the combined distribution and assignment model, *Transportation Networks: Recent Methodological Advances*, M. G. H. Bell (ed.), Elsevier, Oxford, 239-253.

Patriksson, M. (1994) *The Traffic Assignment Problem: Models and Methods*, VSP, Utrecht. (The entire book can be downloaded from http://www.math.chalmers.se/~mipat/traffic.html.)