Bar-Gera, Hillel; Boyce, David

Conference Paper

Combined travel forecasting models - formulations and algorithms

42nd Congress of the European Regional Science Association: "From Industry to Advanced Services - Perspectives of European Metropolitan Regions", August 27th - 31st, 2002, Dortmund, Germany

Provided in Cooperation with:

European Regional Science Association (ERSA)

Suggested Citation: Bar-Gera, Hillel; Boyce, David (2002) : Combined travel forecasting models - formulations and algorithms, 42nd Congress of the European Regional Science Association: "From Industry to Advanced Services - Perspectives of European Metropolitan Regions", August 27th - 31st, 2002, Dortmund, Germany, European Regional Science Association (ERSA), Louvain-la-Neuve

This Version is available at:
https://hdl.handle.net/10419/115601

Combined Travel Forecasting Models: Formulations and Algorithms

Hillel Bar-Gera[a,*], David Boyce[b]

May 30, 2002

[a] Department of Industrial Engineering and Management, Ben-Gurion University, Be'er Sheva 84105 Israel
[b] Department of Civil and Materials Engineering, University of Illinois at Chicago, IL 60607, U.S.A.

To be presented at:
42nd Congress of the European Regional Science Association
Dortmund, Germany, August 27-31, 2002

Abstract

Consistent transportation forecasting models that combine travel demand and network assignment are receiving more attention in recent years. A fixed point formulation for the general combined model is presented. Measures for solution accuracy are discussed. An origin-based algorithm for solving general (non-convex) combined models is proposed. Experimental results demonstrate the efficiency of the algorithm in comparison with prevailing alternatives.

keywords: Travel forcasting; Combined models; Origin-based assignment; Fixed point;

*Corresponding author. Tel: +972-8-6461398; fax: +972-8-6472958; e-mail: bargera@bgumail.bgu.ac.il

1 Introduction

When planning improvements to transportation systems, various alternatives are considered. Careful evaluation requires forecasts of travel patterns for each alternative. Travel patterns are the result of many choices. Traditional modeling practice considers these choices as a sequential process with unique order: activity location choice (trip generation), joint choice of origin and destination (trip distribution), mode choice and finally route choice (assignment). Despite its intuitive appeal, justification for this order is not as trivial as it may seem.

Travelers usually do not think about modes and routes until they have chosen a destination. In many cases this is simply because they have a fairly good idea about the route of choice and its properties, and even more so about the mode of choice and its properties, for most origins and destinations under consideration, prior to choosing their activities. To a certain extent this is true even for choices of work place or residential location, whether made simultaneously or sequentially in one order or another. In view of these observations it seems odd and perhaps even inappropriate to ask which choice comes first.

If all of the conditions that could affect travel choices are known in advance, the order of modeling the different choices should not matter. However, a basic assumption in most forecasting models is that travelers' choices are affected by the level of service of the transportation system. On the other hand, this level of service, and particularly travel times on the roadway network, depend upon the prevailing travel pattern and the associated congestion. The fact that the travel pattern depends on the level of service, which in turn depends on the travel pattern, is one of the main challenges of transportation modeling.

The need to consider congestion effects on route choice became apparent fairly early in the development of travel forecasting models. Early attempts included various computational procedures like quantal loading, origin by origin loading, etc. In recent years user-equilibrium models have gradually replaced previous computational procedures. In these models behavioral assumptions are translated into mathematical conditions that need to be satisfied by the model solution. These well-defined conditions allow one to evaluate approximate solutions, and to examine the convergence of various algorithms.

In contrast to the development and penetration of user-equilibrium route choice models, travel forecasts are still based by and large on sequential procedures. Sequential procedures, even if they are based on user-equilibrium route choice model, still suffer from inconsistent consideration of travel times and congestion effects in the various steps of the procedure. The inconsistent consideration of congestion is a well known and often debated flaw of traditional sequential computational procedures.

This flaw was a key issue in the San Francisco Bay Area lawsuit (Garrett and Wachs, 1996). A common remedy for this flaw is to introduce a "feedback" mechanism into the computational procedure, much like quantal loading in its different forms provides a "feedback" mechanism in computational procedures for network assignment. An alternative approach is to state the behavioral assumptions, translate them into mathematical conditions, and seek solutions that satisfy these conditions. Such models are referred to as *combined* or *integrated* models.

The authors believe that whenever possible models must be formulated mathematically. The first goal of this paper is to demonstrate that for most models used in practice a mathematical formulation requires less effort than generally believed, and that there are important benefits to mathematical formulations which are not always appreciated.

Models that combine several travel choices together are far from new. The first mathematical formulation of user-equilibrium assignment by Beckmann et al. (1956) assumed in addition that the flow of travelers between every origin-destination (O-D) pair of is a function of the level of service for that O-D pair. Their convex optimization formulation was later extended to take substitution effects into account by the introduction of origin and/or destination constraints (Evans, 1976). Evans was also the first to present an efficient convergent algorithm for solving this model. Other convex optimization models include the multi-mode model of Boyce et al. (1983); the multi-mode, multi-class model of Lam and Huang (1992); the multi-mode, multi-class model of Boyce and Bar-Gera (2001) and the origin-based algorithm for solving combined models formulated as convex optimization problems of Bar-Gera and Boyce (2002). Convex optimization formulations have the advantage of unique solutions and algorithms that are proven to converge. More general combined models were formulated as Variational Inequalities (VI) by Dafermos (1982), Florian et al. (2002) and others. Algorithms for combined models are mostly link-based, similar to Evans (1976), with the exception of the route-based algorithm of Lundgren and Patriksson (1998).

The remainder of the paper is organized as follows. Section 2 presents the general fixed point formulation for combined models. Measures for solution accuracy are discussed in section 3. Algorithms for combined models are presented in section 4. Computational results are presented in section 5. Conclusions and suggestions for future research are presented in section 6.

2 Fixed point formulations of combined models

This section presents a mathematical formulation for the general combined model. Mathematical formulations are important tools for describing the goal of a computa-

tional process. Setting the goal is a crucial step that must come prior to any consideration of computational procedures, such as the popular "feedback" mechanism. Only with a clearly stated goal can anyone judge whether a certain procedure performs well or not. Fortunately, combined models can be formulated as fixed point problems in a way that is relatively intuitive and minimal in notation.

Consider a *study area* which is divided into *zones*, during a cerain time period of the day in a given year. Let Z denote the set of all zones. For every pair of origin $p \in Z$ and destination $q \in Z$ let d_{pq} denote the O-D flow (persons/hour) from p to q. $\mathbf{d}$ is the array of O-D flows. Flows are averaged over the entire modeling period (e.g., the morning peak) and over all work days during a specified year. The time period should not be too long, so that flows within it are fairly steady and reasonably represented by their average. Flows can be estimates for past years, or expected values for future years. In any case, it is important to note that as expected/average values, flows can be fractional and do not need to be integers.

The set of available routes from origin p to destination q is denoted by R_{pq}, and the set of these sets is $\mathbf{R} = \{R_{pq}\}_{p,q \in Z}$. The distribution of travelers from p to q among the routes in R_{pq} is described by a vector of non-negative route proportions (conditional probabilities) $\boldsymbol{\gamma}_{pq} = \{\gamma_{pqr}\}_{r \in R_{pq}}$. $\boldsymbol{\gamma}$ is the array of route proportion vectors. Route proportions must add up to one for each O-D pair, hence the set of all feasible route proportion arrays is

$$\Gamma(\mathbf{R}) = \left\{ \boldsymbol{\gamma} \in [0,1]^{|\mathbf{R}|} : \sum_{r \in R_{pq}} \gamma_{pqr} = 1 \quad \forall p, q \in Z \right\} \tag{1}$$

Given $\boldsymbol{\gamma}_{pq}$, the implied vector of route flows is $\mathbf{h}_{pq} = \{h_{pqr}\}_{r \in R_{pq}} = d_{pq} \cdot \boldsymbol{\gamma}_{pq}$. The array of route flow vectors is denoted by $\mathbf{h}$.

In working with these arrays of vectors it is convenient to consider two types of products. The dot product is interpreted as the sum of the product of the elements, similar to a vector dot product, that is $\mathbf{x} \cdot \mathbf{y} = \sum_{pqr} x_{pqr} \cdot y_{pqr}$. The cross product is interpreted as a dot product of array elements, one by one. That is $\mathbf{z} = \mathbf{x} \times \mathbf{y}$ means $z_{pq} = x_{pq} \cdot y_{pq}$. As the algebraic product of matrices is not used in this paper, there should not be any confusion with this notation. Using these conventions, the relationship between O-D flows, route proportions and route flows can be written in short form as $\mathbf{h}(\mathbf{d}, \boldsymbol{\gamma}) = \mathbf{d} \times \boldsymbol{\gamma}$.

According to the user-equilibrium principle of Wardrop (1952), each traveler seeks to minimize the cost associated with their chosen route; therefore, at equilibrium the cost of every used route can not be greater than the cost of any alternative route. The term cost is interpreted as a general measure of dis-utility, which incorporates

travel time. Let $\mathbf{c} = \left\{ \{c_{pqr}\}_{r \in R_{pq}} \right\}_{p,q \in Z}$ be the array of route cost vectors, which is a continuous function of the travel pattern, $\mathbf{c} = \mathbf{C}(\mathbf{h})$. The set of routes of minimum cost for a given O-D pair p, q is denoted by $R^*_{pq}(\mathbf{c}) = \operatorname{argmin}\{c_{pqr} : r \in R_{pq}\}$. The array of such sets is denoted by $\mathbf{R}^*(\mathbf{c})$. For any non-empty subset of routes $\mathbf{R}'$; $\emptyset \subsetneq R'_{pq} \subseteq R_{pq}$, define the set of feasible route proportion arrays that are limited to $\mathbf{R}'$ as

$$\Gamma(\mathbf{R}') = \left\{ \boldsymbol{\gamma} \in \Gamma(\mathbf{R}) : \gamma_{pqr} = 0 \quad \forall r \notin R'_{pq} \quad \forall p, q \in Z \right\} \tag{2}$$

In particular the set of minimum cost assignments is $\Gamma(\mathbf{R}^*(\mathbf{c}))$. It is obvious without any derivation that the travel pattern $\{\mathbf{d}, \boldsymbol{\gamma}\}$ satisfies the user-equilibrium requirements iff

$$\boldsymbol{\gamma} \in F_1(\mathbf{d}, \boldsymbol{\gamma}) = \Gamma(\mathbf{R}^*(\mathbf{C}(\mathbf{h}(\mathbf{d}, \boldsymbol{\gamma})))) \tag{3}$$

In other words, user-equilibrium route proportions must belong to the set of feasible route proportions that are limited to the set of minimum cost routes, where route costs correspond to route flows that result from the chosen route proportions.

We assume that the array of O-D flows is a continuous upper-bounded function of O-D costs, $\mathbf{d} = \Phi(\mathbf{u})$, where $\mathbf{u} = \{u_{pq}\}_{p,q \in Z}$ is the array of O-D costs. O-D costs equal average route costs, weighted by flow, $\bar{U}_{pq}(\mathbf{c}, \boldsymbol{\gamma}) = \boldsymbol{\gamma}_{pq} \cdot \mathbf{c}_{pq}$, or $\bar{\mathbf{U}}(\mathbf{c}, \boldsymbol{\gamma}) = \boldsymbol{\gamma} \times \mathbf{c}$. The fixed point formulation of the combined model is

$$\{\mathbf{d}, \boldsymbol{\gamma}\} \in F_2(\mathbf{d}, \boldsymbol{\gamma}) = \left\{ \Phi\left(\bar{\mathbf{U}}(\mathbf{C}(\mathbf{h}(\mathbf{d}, \boldsymbol{\gamma})), \boldsymbol{\gamma})\right) \right\} \times \Gamma(\mathbf{R}^*(\mathbf{C}(\mathbf{h}(\mathbf{d}, \boldsymbol{\gamma})))) \tag{4}$$

or equivalently

$$\mathbf{d} = \Phi\left(\bar{\mathbf{U}}(\mathbf{C}(\mathbf{h}(\mathbf{d}, \boldsymbol{\gamma})), \boldsymbol{\gamma})\right) \tag{5}$$

$$\boldsymbol{\gamma} \in \Gamma(\mathbf{R}^*(\mathbf{C}(\mathbf{h}(\mathbf{d}, \boldsymbol{\gamma})))) \tag{6}$$

These equations state that at equilibrium O-D flows must correspond to prevailing O-D costs, and at the same time the user-equilibrium conditions must be satisfied.

For user-equilibrium solutions $\bar{U}_{pq}(\mathbf{c}, \boldsymbol{\gamma}) = U^*_{pq}(\mathbf{c}_{pq}) \equiv \min\{c_{pqr} : r \in R_{pq}\}$. Therefore, in the above formulation, we can replace $\bar{\mathbf{U}}$ with $\mathbf{U}^*$ and obtain an equivalent formulation.

This formulation can be easily extended to multi-mode and multi-class models, by adding a mode subscript m and a class superscript l to all variables. In other words if we let $\mathbf{d} = \left\{d^l_{mpq}\right\}$; $\mathbf{R} = \left\{R^l_{mpq}\right\}$; $\boldsymbol{\gamma} = \left\{\gamma^l_{mpqr}\right\}$; $\mathbf{h} = \left\{h^l_{mpqr}\right\}$; $\mathbf{c} = \left\{c^l_{mpqr}\right\}$, and adapt the interpretation of $\mathbf{R}^*$, Γ, Φ, $\mathbf{C}$, and $\bar{\mathbf{U}}$ accordingly, then Eq. (4) is a mathematical formulation of a generic multi-mode, multi-class model.

Solution existence is demonstrated by Kakutani's extension to Brouwer's fixed point

theorem (Kakutani, 1941; Nikaido, 1968, Theorem 4.4, p. 67). Nikaido defines a set-valued mapping $f : X \to 2^Y$, where 2^Y represents the set of subsets of Y, to be closed if $x^k \to x$; $y^k \to y$; $y^k \in f(x^k)$ implies $y \in f(x)$. Every continuous function is closed; hence $\Phi, \bar{\mathbf{U}}, \mathbf{C}$ are closed. $\mathbf{R}^*$ is closed, with discrete topology on $\mathbf{R}$, and Γ is also closed. Therefore, F_2 is closed. $\Gamma(\mathbf{R}')$ is convex for every set of routes $\mathbf{R}'$, hence $F_2(\mathbf{d}, \boldsymbol{\gamma})$ is convex for every $(\mathbf{d}, \boldsymbol{\gamma})$. Due to the upper bound on O-D flows, M, the set of feasible solutions, $[0, M]^{|Z|\times|Z|} \times [0,1]^{|\mathbf{R}|}$ is non-empty, compact and convex. Under these conditions, Kakutani's extension to Brouwer's fixed point theorem guarantees that the map F_2 has a fixed point. In other words, there is at least one solution for the combined model in Eq. (4).

3 Accuracy Measures

One of the main advantages of mathematically formulated models is the ability to evaluate solutions by well defined accuracy measures, and hence to determine whether a solution is sufficiently accurate for the specific analysis under consideration.

In the case of O-D flows it is natural to compare O-D flows in the current solution $\mathbf{d}$, with the O-D flows that result from the costs of travel under current conditions. For the latter we can choose either minimum O-D costs $\mathbf{d}' = \Phi(\mathbf{U}^*(\mathbf{C}(\mathbf{h})))$ or average O-D costs $\mathbf{d}'' = \Phi(\bar{\mathbf{U}}(\mathbf{C}(\mathbf{h}), \boldsymbol{\gamma}))$. Both comparisons lead to similar results. We shall use $\mathbf{d}'$ simply because average O-D costs are not available for some algorithms. Possible aggregate measures of accuracy are the maximum positive difference, $max\{d'_{pq} - d_{pq} : d'_{pq} \geq d_{pq}\}$, the maximum negative difference, $max\{d_{pq} - d'_{pq} : d'_{pq} \leq d_{pq}\}$, and the total misplaced O-D flow, $\sum_{p,q\in Z} |d'_{pq} - d_{pq}|$. All are in units of person trips per hour.

The intuitive interpretation of these measures can be very helpful in setting conditions for sufficiently accurate solutions. For example, consider a study that examines the impact of a new commercial facility, which is expected to attract 1,000 trips per hour during the morning peak. It would be reassuring to know that the total misplaced O-D flow in the solution is less than 100 trips/hour. A total misplaced O-D flow of 1,000 trips/hour may still be acceptable, assuming that it is spread over a wide region. But, a total misplaced O-D flow of 10,000 trips/hour is probably not acceptable, as it is quite likely to have significant influence on the results of the study.

Assignment accuracy measures can be based on the distribution of *excess cost*, $ec_r = c_{pqr} - U^*_{pq}(\mathbf{c}_{pq})$, among used routes. Possible aggregate measures based on access cost include the maximum excess cost, the 95-th percentile, the portion of flow with excess cost above a certain value, say one minute, etc. The main aggregate measure used in this paper is the *average excess cost*, $AEC_a = \frac{1}{d_{\bullet\bullet}} \cdot \sum_{r\in R} h_r \cdot ec_r$, where

$d_{\bullet\bullet} = \sum_{p,q \in Z} d_{pq}$ is the total O-D flow (on the road network). In the context of fixed demand problems AEC is sometimes referred to as "normalized gap".

Setting requirements for assignment accuracy is more challenging, since typically the goal is to make sure that link flows are sufficiently close to the true equilibrium solution. A case study (Boyce and Bar-Gera, 2002) examined the impact of adding a pair of freeway ramps in the Delaware Valley Region. The goal of the study was to estimate flow differences on links in the vicinity of the proposed improvement between the Build and No-Build scenarios. It was found that solutions should have Average Excess Cost less than 0.001 vehicle-minutes, so that estimates for freeway links shall be within 3% from the true equilibrium solution, and estimates for arterials shall be within 10% from the true equilibrium solution. As additional case studies are conducted on different networks and for various levels of congestion, more definite recommendations will be available for practitioners.

A solution is considered to be sufficiently accurate only if it satisfies both conditions, that is if it has average excess cost less than, say, 0.001 vehicle-minutes, and total misplaced O-D flow less than, say, 1000 trips/hour.

4 Algorithms

In this paper we consider two algorithms; both are iterative. In the first algorithm, demand and route proportions are updated simultaneously in every iteration. It is similar to the algorithm proposed by Evans (1976) for a combined model formulated as a convex optimization problem. In every iteration of this algorithm, given the current solution $\left(\mathbf{d}^k, \boldsymbol{\gamma}^k\right)$; $\mathbf{h}^k = \mathbf{d}^k \times \boldsymbol{\gamma}^k$, a subproblem solution is found in the following way. First a minimum cost assignment is chosen, given the current costs, $\hat{\boldsymbol{\gamma}}^k \in \Gamma\left(\mathbf{R}^*\left(\mathbf{C}\left(\mathbf{h}^k\right)\right)\right)$; then, new O-D flows are found using the minimum costs found in the previous step, $\hat{\mathbf{d}}^k = \Phi\left(\mathbf{U}^*\left(\mathbf{C}\left(\mathbf{h}^k\right)\right)\right)$; finally, the new demand is assigned to the minimum cost routes found in the first step $\hat{\mathbf{h}}^k = \hat{\mathbf{d}}^k \times \hat{\boldsymbol{\gamma}}^k$.

Once a subproblem solution is found, a new solution is obtained by a weighted average of the current solution and the subproblem solution, $\mathbf{h}^{k+1} = (1-\lambda) \cdot \mathbf{h}^k + \lambda \cdot \hat{\mathbf{h}}^k$, where $0 \leq \lambda \leq 1$ is the *step size*, or the weight of the subproblem solution. Since total link flows are a linear function of route flows, averaging route flows and averaging total link flows lead to the same solution. Therefore, implementations of this algorithm typically store only total link flows, thus reducing memory requirements substantially. In the algorithm proposed by Evans (1976), the convex formulation was used to determine the step size. In the general case, when a convex formulation is not available, different techniques must be used to determine the step size, as discussed below.

Initialization:

Let $\mathbf{u} = \mathbf{U}^*(C(0))$
Let $\mathbf{d}^0 = \Phi(\mathbf{u})$
for p in Z do
  A_p = tree of minimum cost routes from p
  $\mathbf{f}_p$ = all or nothing assignment using A_p
end for

Main loop:

for n=1 to number of main iterations
  Update O-D flows, retain route proportions
  for p in Z do
    update restricting subnetwork A_p
    update origin-based approach proportions $\boldsymbol{\alpha}_p$
  end for
  for m=1 to number of inner iterations
    for p in Z do
      update origin-based approach proportions $\boldsymbol{\alpha}_p$
    end for
  end for
end for

Fig. 1: An origin-based algorithm for combined models

The second algorithm is similar to the origin-based algorithm proposed in Bar-Gera and Boyce (2002) for combined models with convex formulations, which is based on the origin-based assignment algorithm (Bar-Gera, 2002). The general scheme of the algorithm is presented in Fig. 1. Stopping conditions for the algorithm are based on total misplaced O-D flow and average excess cost, as discussed in section 3.

The key element in the proposed algorithm for combined models is the procedure for updating O-D flows, while retaining the route proportions of the current solution. Given a current solution, $\{\mathbf{d}^k, \boldsymbol{\gamma}^k\}$, subproblem O-D flows are determined according to average O-D costs $\hat{\mathbf{d}}^k = \Phi\left(\boldsymbol{\gamma}^k \times \mathbf{C}\left(\mathbf{d}^k \times \boldsymbol{\gamma}^k\right)\right)$. New O-D flows are obtained by a weighted average $\mathbf{d}_\lambda^{k+1} = (1-\lambda)\cdot\mathbf{d}^k + \lambda\cdot\hat{\mathbf{d}}^k$, where $0 \le \lambda \le 1$ is a chosen step size.

In models that have a convex formulation, it can be used to determine the step size. A proof of convergence for the resulting algorithm is given in Bar-Gera and Boyce (2002). As shown there, the use of average O-D costs (rather than minimum O-D costs) to determine subproblem O-D flows is critical for convergence. As with the Evans-like algorithm, when a convex formulation is not available, different techniques

must be used to determine the step size.

Once O-D flows are updated, route proportions are revised by an origin-based assignment iteration, while keeping O-D flows temporarily fixed. The main solution variables in the origin-based assignment algorithm are *origin-based approach proportions* $\boldsymbol{\alpha} = \{\alpha_{ap}\}_{a\in A; p\in Z}$; $0 \le \alpha_{ap} \le 1$; $\sum_{a:a_h=j} \alpha_{ap} = 1 \ \ \forall j \in N \ \forall p \in Z$, where N and A are the sets of nodes and links on the road network, and a_t, a_h are the tail and the head of link $a = [a_t, a_h]$. For every origin an a-cyclic restricting subnetwork is chosen, $A_p \subseteq A$; $a \notin A_p \Rightarrow \alpha_{ap} = 0$. Initial restricting subnetworks are trees of minimum cost routes. To update the restricting subnetwork, unused links are removed, ν_i - the maximum cost to node i within the restricting subnetwork is computed, and all links $[i, j]$ such that $\nu_i < \nu_j$ are added to the restricting subnetwork. Approach proportions for origin p are updated by shifting flows within the restricting subnetwork A_p according to a boundary (piece-wise linear) search in a direction determined by an approximate second order method. Given the demand $\mathbf{d}$ and the current solution $\boldsymbol{\alpha}$, the set of restricting subnetworks for the next iteration is defined by a function $\mathbf{A} = \mathcal{A}(\mathbf{d}, \boldsymbol{\alpha})$, and the next iteration solution is defined by a map $\boldsymbol{\alpha}' \in \Theta^{\boldsymbol{\alpha}}(\mathbf{d}, \boldsymbol{\alpha})$.

Route proportions are determined by $\gamma_{pqr} = \prod_{a\subseteq r} \alpha_{ap}$. It can be shown that origin-based link flows $f_{ap}(\mathbf{h}) = \sum_{q\in Z} \sum_{r\in R_{pq}; a\subseteq r} h_{pqr}$ and origin-based node flows $f_{jp}(\mathbf{h}) = \sum_{q\in Z} \sum_{r\in R_{pq}; j\in r} h_{pqr} = \sum_{a:a_h=j} f_{ap}$ maintain the relationship $f_{ap} = \alpha_{ap} \cdot g_{jp}$, demonstrating that α_{ap} is indeed the proportion of flow on approach a to node a_h for origin p. The availability of route proportions allows one to compute average O-D costs, as well as the assignment of new O-D flows by current route proportions. Due to the restriction to a-cyclic subnetworks, these computations can be done efficiently without route enumeration, in a time that is a linear function of the number of links times the number of origins. These properties are essential for the demand update procedure described above.

In both the Evans-like algorithm and the origin-based algortihm the main obstacle towards a general implementation for non-convex models is the determination of the step size. The most well known technique to determine step sizes in general problems is the Method of Successive Averages (MSA), introduced in the seminal work of Robbins and Monroe (1951). In this technique, step sizes are predetermined as $\lambda_k = 1/k$. (Where k is the iteration index.) Polyak (1990) argues that in the context of stochastic approximation techniques, under certain conditions, it is better to use either a constant step size, or step sizes of $\lambda_k = k^{-\beta}$ where $0 < \beta < 1$.

Basic intuition for the behavior of different algorithms with various choices of step sizes can be developed by considering a very simple example of a single dimensional problem with a given feasible range $[0, M]$, and an unknown optimal solution x^*. Consider an averaging algorithm, $x^{k+1} = (1 - \lambda_k) \cdot x^k + \lambda_k \cdot y^k$ that is based on subproblem

solution $y^k = f_a(x^k)$, where

$$f_a(x) = \begin{cases} M & x \le b_1 \\ x^* - a(x - x^*) & b_1 \le x \le b_2 \\ 0 & b_2 \le x \end{cases} \tag{7}$$

$$b_1 = x^* - \frac{M - x^*}{a}; \quad b_2 = x^* + \frac{x^*}{a} \tag{8}$$

The value of a controls the accuracy of the subproblem. $a = 0$ indicates perfectly accurate subproblem solutions, since $f_0(x) = x^* \ \ \forall x$. On the other hand, as $a \to \infty$ subproblem solutions are less accurate, and at the limit a semi-continuous point to set function is obtained

$$f_\infty(x) = \begin{cases} M & x < x^* \\ [0, M] & x = x^* \\ 0 & x > x^* \end{cases} \tag{9}$$

When subproblem solutions are accurate, i.e. $a \to 0$, larger step sizes lead to faster convergence. MSA does not take advantage of accurate subproblem solutions, however, since even if $a = 0$, the convergence of MSA is given by $x^k - x^* = \left(x^0 - x^*\right)/k$, which is quite slow. The progress under constant step size, λ, while in the linear range $[b_1, b_2]$, is given by:

$$x^{k+1} = (1 - \lambda) \cdot x^k + \lambda \cdot (x^* - a \cdot (x^k - x^*)) = x^* + (1 - \lambda - \lambda \cdot a) \cdot (x^k - x^*) \tag{10}$$

$$x^{k+1} - x^* = \left(x^k - x^*\right) \cdot (1 - \lambda \cdot (1 + a)) = \left(x^0 - x^*\right) \cdot (1 - \lambda \cdot (1 + a))^k \tag{11}$$

The sequence x^k converges to x^* for every step size $0 < \lambda < 2/(1 + a)$. Oscillations are avoided by any step size $0 < \lambda < 1/(1 + a)$. Further reducing step sizes only slows convergence. In the case of f_∞, any constant step size leads to oscillations around x^*. Smaller step size leads to oscillations closer to x^*, but also to slower progress towards x^*.

In monitoring the progress of the algorithm, typically, the value of x^* is not known, hence $x^k - x^*$ can not be evaluated. Instead, we can monitor the value of $gap^k = x^k - y^k$. This is similar to the accuracy measure for O-D flows proposed in section 3. For an algorithm that is based on f_a, when the slope a is finite, the relative reduction in the gap within the linear range $[b_1, b_2]$ is given by

$$\frac{gap^k - gap^{k+1}}{gap^k} = 1 - \frac{x^{k+1} - y^{k+1}}{x^k - y^k} = \lambda \cdot (1 + a) \tag{12}$$

With perfectly accurate subproblem solutions ($a = 0$) the relative gap reduction is

equal to the step size λ. Notice that in the first iterations of an algorithm based on f_∞, that is until oscillations start, the relative gap reduction is also equal to the step size. It is therefore interesting to monitor the relative gap reduction in computational experiments.

It is clear from the above discussion, that the optimal step size strategy depends on the algorithm. If subproblem solutions are continuous as a function of the current solution, as is the case with f_a when a is finite as well as for the origin-based algorithm described above, a constant step size, if sufficiently small, should not cause oscillations. Our goal in that case is to use the largest constant step size that does not lead to oscillations. On the other hand, if subproblem solutions are not continuous, as is the case with f_∞ and with the Evans-like algorithm, any constant step size will eventually lead to oscillations; therefore, we must use a decreasing sequence of step sizes.

5 Experimental Results

This section presents computation results comparing the convergence of the proposed origin-based algorithm and the Evans-like algorithm for different step size strategies. All experiments were conducted on the same Compaq Alpha Unix Server model DS20E, with CPU speed of 666 MHz, and 256MB RAM.

The algorithms were applied to a multimodal model, which is similar to the model presented in Boyce and Daskin (1997). The network of the model, referred to as the "Chicago Sketch Network," consists of 387 zones, 933 nodes, 2,950 road links, and total O-D flow of about 1.25 million person trips per hour. The main inputs to this model are: the flow of person trips per hour from each origin $\bar{d}_{p\bullet}$; the flow of person trips per hour to each destination $\bar{d}_{\bullet q}$; free flow travel times tt_a^0, capacities k_a, and lengths l_a for each link on the roadway network; parking costs pc_z and walking times to or from the parking place wt_z for each zone; in vehicle travel times c_{tpq}^{ivtt}, fares c_{tpq}^{fare}, and out of vehicle times c_{tpq}^{ovt} when traveling by transit from origin p to destination q (these are fixed regardless of flows); and truck flows d_{pq}^{truck} by O-D in passenger cars equivalents.

Link travel time functions are of the BPR form

$$tt_a(f_a) = tt_a^0 \cdot (1 + 0.15 \cdot (f_a/k_a)^4) \tag{13}$$

Auto operating costs, including gasoline consumption, are a linear function of link length and travel time,

$$oc_a = \eta_1 \cdot tt_a(f_a) + \eta_2 \cdot l_a \tag{14}$$

Link generalized costs are

$$t_a(f_a) = \beta^{au}_{time} \cdot tt_a(f_a) + \beta^{au}_{cost} \cdot oc_a(f_a) \tag{15}$$

where the β's are calibration parameters. Parking costs and walking times are components of the additional auto costs, defined as

$$ac_{apq} = \beta^{au}_{park} \cdot \frac{pc_p + pc_q}{2} + \beta^{au}_{walk} \cdot (wt_p + wt_q) \tag{16}$$

Route generalized costs by auto are $c_{apqr} = ac_{apq} + \sum_{a \in r} t_a$. O-D generalized costs by auto are $\bar{U}_{apq}(\mathbf{c}, \boldsymbol{\gamma}) = \boldsymbol{\gamma}_{apq} \cdot \mathbf{c}_{apq}$, as before. O-D generalized costs by transit are

$$u_{tpq} = \beta^{tr}_{bias} + \beta^{tr}_{ivtt} \cdot c^{ivtt}_{tpq} + \beta^{tr}_{fare} \cdot c^{fare}_{tpq} + \beta^{tr}_{ovt} \cdot c^{ovt}_{tpq} \tag{17}$$

O-D flows are of the compound LOGIT form

$$d_{apq} = A_p \cdot B_q \cdot \exp(-\mu \cdot u_{apq}) \cdot u^{-\rho}_{apq} \tag{18}$$

$$d_{tpq} = A_p \cdot B_q \cdot \exp(-\mu \cdot u_{tpq}) \cdot u^{-\rho}_{tpq} \tag{19}$$

Flows of person trips by auto are converted to vehicle flows by a predetermined auto occupancy factor, aof. The same route proportions are used for auto flows and for truck flows, hence

$$\mathbf{h}_{pq}(\mathbf{d}, \boldsymbol{\gamma}) = \left(d_{apq}/aof + d^{truck}_{pq} \right) \cdot \boldsymbol{\gamma}_{pq} \tag{20}$$

The fixed point formulation in Eq. (4) applies to this model almost directly, except that the definition of $\mathbf{h}_{pq}(\mathbf{d}, \boldsymbol{\gamma})$ mentioned above is slightly different than before. The only difference between this model and the model considered in Bar-Gera and Boyce (2002) is the power term $u^{-\rho}$ in (18) and (19). For $\rho = 0$, we get the original model, that can be formulated as a convex optimization problem. In that case, the performance of the algorithms based on step sizes determined using the convex objective function can be used as a reference for the performance of any other step size strategy.

Practitioners have reported that in some cases observed data is better explained by models with $\rho \approx 1$. Figures 2-7 show results of the two algorithms for $\rho = 0, 1, 2$. In the case of $\rho = 0$, the results of the algorithm described in Bar-Gera and Boyce (2002), which is based on a line search over the convex objective function, are included as a reference.

Equilibrium solutions are quite different for different values of ρ. For example total intra-zonal flow values are about 105,900, 88,708 and 63,203 person trips per hour for $\rho = 0, 1, 2$ respectively. On the other hand, the behavior of the different algorithms is

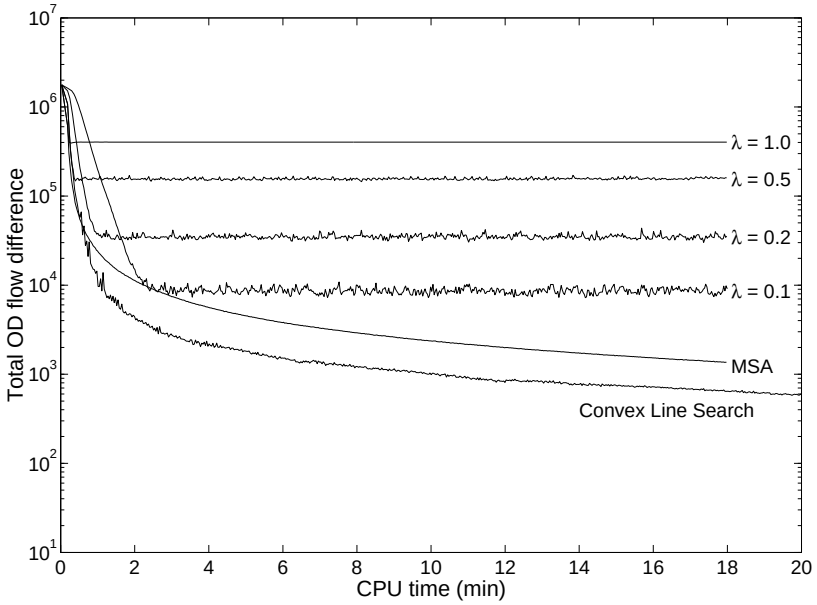

Fig. 2: Convergence of Evans-like algorithms, $\rho = 0$

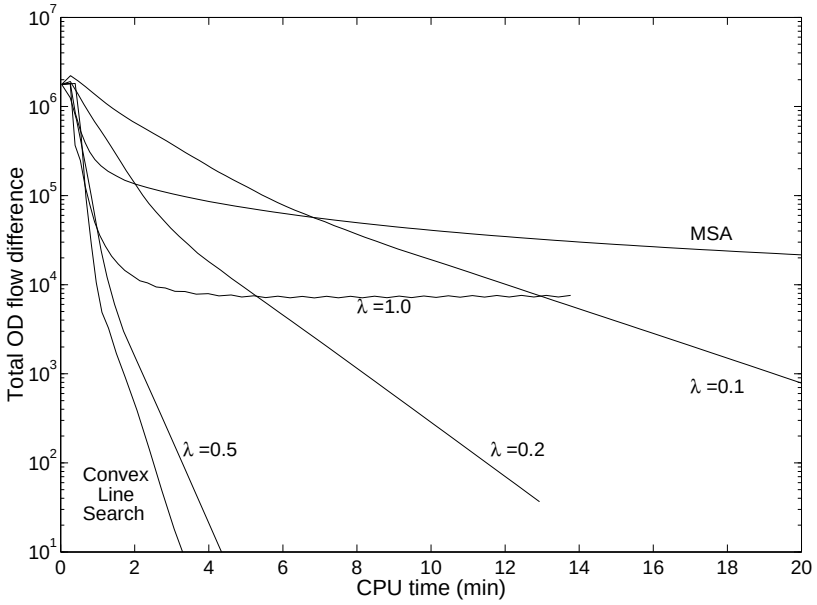

Fig. 3: Convergence of origin-based algorithms, $\rho = 0$

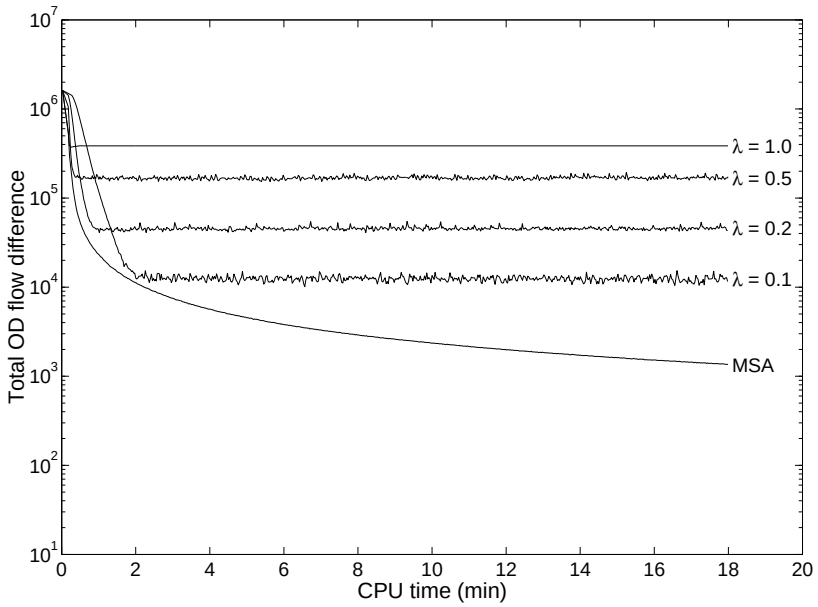

Fig. 4: Convergence of Evans-like algorithms, $\rho = 1$

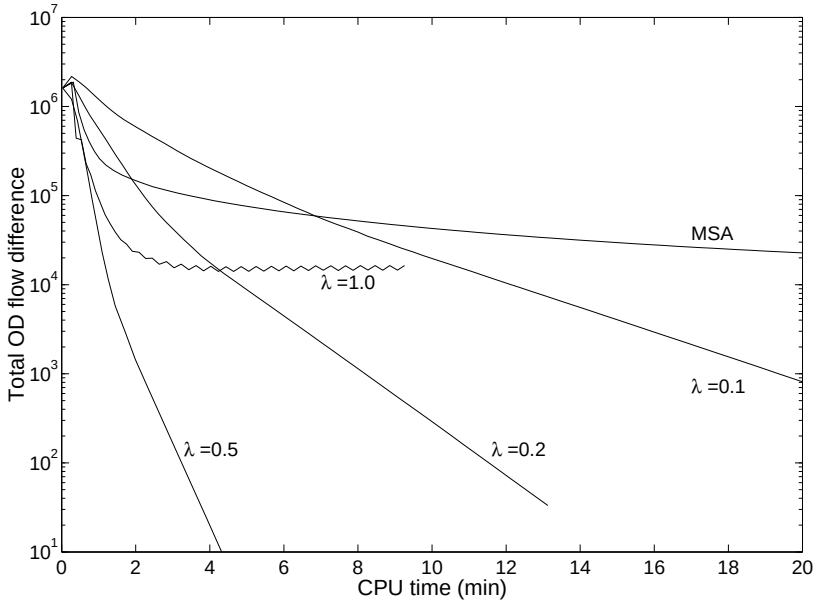

Fig. 5: Convergence of origin-based algorithms, $\rho = 1$

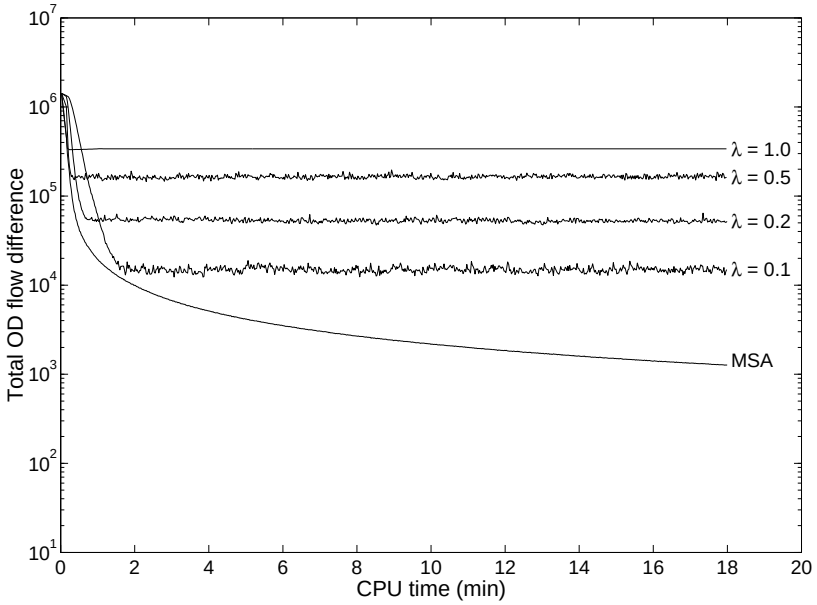

Fig. 6: Convergence of Evans-like algorithms, $\rho = 2$

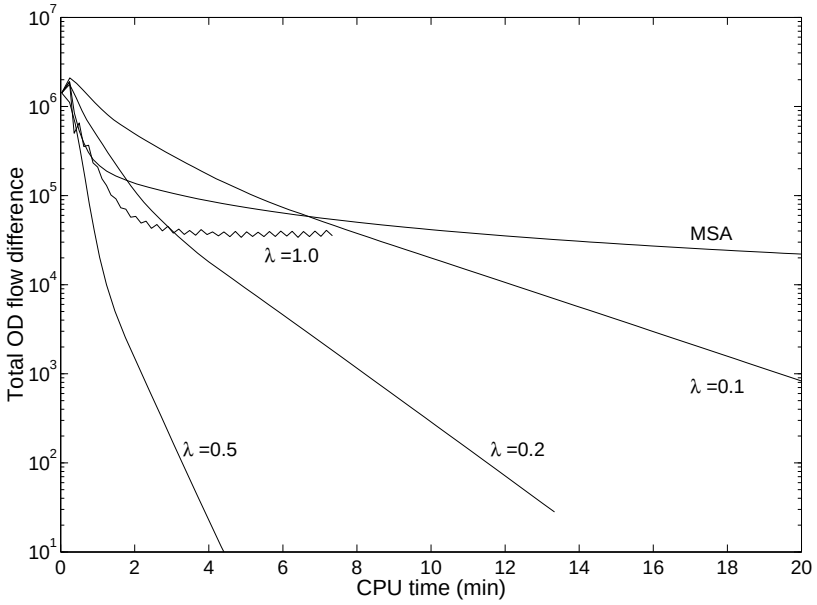

Fig. 7: Convergence of origin-based algorithms, $\rho = 2$

almost identical in all three cases.

In the case of Evans-like algorithms, we find that any constant step size at some point leads to oscillations, as expected. In this case MSA performs reasonably well, although it is possible that a more sophisticated strategy will provide better results. On the other hand, in the case of origin-based algorithms, any step size less than 0.5 leads to convergence much faster than with MSA step sizes.

Comparing the convergence of the two algorithms with MSA step sizes with respect to CPU time, shows substantial advantage for the Evans-like algorithm. This is mainly because each origin-based iteration takes more CPU time. For example 50 origin-based iterations take about 15 minutes of CPU time, and lead to total O-D flow difference of about 30,000 person trips/hour (with MSA step sizes); 50 Evans-like iterations take about 1.3 minutes of CPU time, and lead to total O-D flow difference of about 15,000 person trips/hour (with MSA step sizes); in 15 minutes of CPU time about 600 Evans-like iterations are performed, leading to total O-D flow difference of about 1500 person trips/hour (with MSA step sizes).

Assignment convergence of solutions obtained by origin-based algorithms is substantially superior to that obtained by Evans-like algorithms. Solutions after 50 origin-based iterations with $\lambda = 0.5$, obtained in 4-5 minutes of CPU time, have average excess cost of less than 1E-10 vehicle minute equivalents. Solutions after 1000 Evans-like iterations with MSA step sizes, obtained in 20-30 minutes of CPU time, have average excess cost of 0.005 to 0.01 vehicle minute equivalents.

It is interesting to point out that in most cases, in all algorithms while not in oscillation, the relative reduction in gap is fairly close to the step size. As discussed in the previous section, this is expected to happen when the subproblem solution does not change much as a function of the current solution, at least for most dimensions. Given the enormous number of dimensions in this problem, there may be other reasons for this observation as well. In any case, it may be possible to use this observation to develop an algorithm that adjusts the step size during the run. This remains a subject for future research.

6 Conclusions and Future Research

Traditional travel forecasting methods were based on sequential computational procedures. The need for integrated or combined models is becoming more and more convincing, in view of court decisions and legislative mandates in the U.S. in the last decade. The fixed point formulation presented in this paper seems to be a natural tool to formulate general combined models mathematically, including most models used in practice. The need for intuitive accuracy measures leads to separate consideration of

assignment accuracy and the accuracy of O-D flows.

General combined models can be solved with either an Evans-like algorithm, or with an origin-based algorithm. In the first case, the sequence of step sizes used for averaging should be decreasing, as in the MSA. As a result, convergence is relatively slow. The proposed origin-based algorithm can provide much faster convergence, when a constant step size is used, as long as the step size is not too large.

The results presented in this paper should be examined and validated in other models, and particularly for different levels of congestion.

Acknowledgments

The authors wish to thank the Chicago Area Transportation Study, Chicago, IL. for providing the Chicago network data.

References

Bar-Gera, H., 1999. Origin-based Algorithms for Transportation Network Modeling. Ph.D. Thesis, University of Illinois at Chicago, U.S.A.

Bar-Gera, H., 2002. Origin-based Algorithm for the Traffic Assignment Problem. Accepted for publication in Transportation Science.

Bar-Gera, H., and Boyce, D., 2002. Origin-based Algorithms for Combined Travel Forecasting Models. Accepted for publication in Transportation Research B.

Beckmann, M., McGuire, C. B. and Winston, C. B., 1956. Studies in the Economics of Transportation, Yale University Press, New Haven, CT.

Boyce, D. and Bar-Gera, H., 2001. Network Equilibrium Models of Travel Choices with Multiple Classes. In: Lahr, M.L. and Miller R.E. (Eds.). Regional Science in Economic Analysis, Elsevier Science, Oxford, UK. Chapter 6, pp. 85–98.

Boyce, D., and Bar-Gera, H., 2002. Convergence of Traffic Assignments: How Much Is Enough? The Delaware Valley Region Case Study. Working paper, Department of Civil and Materials Engineering, University of Illinois at Chicago, Chicago, IL.

Boyce, D., Chon, K.S., Lee, Y.J., Lin, K.T. and LeBlanc, L.J., 1983. Implementation and evaluation of combined models of location, destination, mode and route choice. Environment and Planning A 15, pp. 1219–1230.

Boyce, D. and Daskin, M.S., 1997. Urban Transportation. In: ReVelle, C. and McGarity, A.E. (Eds.). Design and Operation of Civil and Environmental Engineering Systems, John Wiley and Sons, New-York. Chapter 7, pp. 277-341.

Dafermos, S., 1982. The general multimodal network equilibrium problem with elastic demand. Networks 12, pp. 57–72.

Evans, S. P., 1976. Derivation and Analysis of Some Models for Combining Trip Distribution and Assignment. Transportation Research 10, pp. 37–57.

Florian, M., Wu, J.-H., and He, S., 2002. A multi-class multi-mode variable demand network equilibrium with hierarchical logit structures, In: M. Gendreau and P. Marcotte (eds.). Transportation and Network Analysis: Current Trends, Kluwer Academic Publishers, Dordrecht, Netherlands. Chapter 8, pp.119-133.

Garrett, M. and Wachs, M., 1996. Transportation Planning on Trial : The Clean Air Act and Travel Forecasting, Sage Publications, Thousand Oaks, CA.

Kakutani, S., 1941. A Generalization of Brouwer's Fixed Point Theorem. Duke Math. J. 8, No. 3.

Lam, W.H.K. and Huang, H.J., 1992. A combined trip distribution and assignment model for multiple user classes. Transportation Research 26B, pp. 275–287.

Lundgren, J.T. and Patriksson, M., 1998. An Algorithm for the Combined Distribution and Assignment Model. Transportation Networks: Recent Methodological Advances, M. G. H. Bell (ed.), Oxford: Elsevier, 239-253.

Nikaido, H., 1968. Convex Structures and Economic Theory, Academic Press, New York.

Polyak, B.T., 1990. New method of stochastic approximation type. Automation and Remote Control, 51(7), pp. 937-946.

Robbins, H., and Monro, S., 1951. A stochastic approximation method. Mathematical statistics, Vol. 22, pp. 400-407.

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