Public Transportation Decision-Making: A Case Analysis of the Memphis Light Rail Corridor and Route Selection with Analytic Hierarchy Process

Where Transit Use Is Growing: Surprising Results

Design of a Map-Based Transit Itinerary Planner

Valuing Urban Bus Attributes: An Experience in Kolkata


Large-Scale Transit Network Optimization by Minimizing User Cost and Transfers
Gary L. Brosch, Editor
Lisa Maitland, Assistant to the Editor

EDITORIAL BOARD

Robert B. Cervero, Ph.D.
University of California, Berkeley

William W. Millar
American Public Transportation Association

Chester E. Colby
E & J Consulting

Steven E. Polzin, Ph.D., P.E.
University of South Florida

Gordon Fielding, Ph.D.
University of California, Irvine

Sandra Rosenbloom, Ph.D.
University of Arizona

David J. Forkenbrock, Ph.D.
University of Iowa

Lawrence Schulman
LS Associates

Jose A. Gómez-Ibáñez, Ph.D.
Harvard University

George Smerk, D.B.A.
Indiana University

Naomi W. Ledé, Ph.D.
Texas Transportation Institute

Vukan R. Vuchic, Ph.D., P.E.
University of Pennsylvania

The contents of this document reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation, University Research Institute Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.
CONTENTS

Public Transportation Decision-Making: A Case Analysis of the Memphis Light Rail Corridor and Route Selection with Analytic Hierarchy Process
Reza Banai....................................................................................................................................................1

Where Transit Use Is Growing: Surprising Results
Gregory L. Thompson, Jeffrey R. Brown, Rupa Sharma, Samuel Scheib..........................25

Design of a Map-Based Transit Itinerary Planner
Christopher Cherry, Mark Hickman, Anirudh Garg................................................................45

Valuing Urban Bus Attributes: An Experience in Kolkata
C. V. Phanikumar, Bhargab Maitra..................................................................................................69

Rabi G. Mishalani, Mark M. McCord, John Wirtz........................................................................89

Large-Scale Transit Network Optimization by Minimizing User Cost and Transfers
Fang Zhao..................................................................................................................................................107

Our troubled planet can no longer afford the luxury of pursuits confined to an ivory tower. Scholarship has to prove its worth, not on its own terms, but by service to the nation and the world.
—Oscar Handlin
Public Transportation Decision-Making: A Case Analysis of the Memphis Light Rail Corridor and Route Selection with Analytic Hierarchy Process

Reza Banai, University of Memphis

Abstract

The Federal Transit Administration (FTA) New Starts process involves multiple criteria to assess funding eligibility for local public transit investments. In this article a multicriteria method—Analytic Hierarchy Process (AHP)—is used to assess light rail transit (LRT) corridor and route alternatives. Although the focus is on the current LRT corridor and route selection process in Memphis, Tennessee, the AHP-aided procedure is intended to facilitate the public transportation decision-making process generically, reflective of federal New Starts guidelines as well as local priorities and preferences. Each alternative corridor and route is assessed functionally with respect to site-specific ratings of the criteria and subcriteria in a unified framework. This framework contains the goal, participant groups, criteria, subcriteria, and alternatives as various elements of a public transportation decision process with relative influence on the outcome. The best corridor and route alignment alternative is identified by a composite score on the AHP ratio scale. Finally, with sensitivity analysis, it is
shown how a change on the importance of the criteria or participant group priority influences the trade-offs among the criteria and the outcome. The article concludes with a retrospective, reflective discussion of the planning process as a whole.

Introduction
Throughout North America, investment in rail transit by cities continues to increase while federal, state, and local funding sources decline due to budget constraints and the recent economic downturn and shifting of funding priorities to defense and homeland security. Cities and regional metropolitan planning organizations have included rail transit as an element of their federally-required, 20-year long-range transportation plans as a means to meet future air quality standards. This trend is highlighted by the increasing number of projects entering the Federal Transit Administration (FTA) New Starts pipeline. New Starts project justification and financial criteria are used by the FTA to recommend projects for funding. As projects are developed and proceed through the planning process, they are evaluated by the statutory criteria, FTA U.S.C. 5309 (e) (6), which requires that a summary rating of “highly recommended,” “recommended,” or “not recommended” be assigned to each project. The multiple criteria range from mobility to land use, environmental impact, and financial efficiency. These projects vary from a minimum capital cost of $25 million, which are exempt from the New Starts ratings, to an estimated 4.35 billion for the New York/East Side Access project. Each new extension to an existing system must go through this process.

The current light rail transit (LRT) corridor and route selection process in Memphis, Tennessee, under federal New Starts funding consideration, is the motivation for this article (Figure 1). The Analytic Hierarch Process (AHP) is used to describe the structure of the planning and decision-making process involving LRT corridor and route selection. The appeal of AHP as an ex ante method of forecasting and decision making in a wide variety of applications is the accuracy of the predictions and decision outcomes that turn out to be true when events become known later. Thus, there is a growing interest in AHP as a predictive as well as multicriteria decision analysis method used in transportation among many applications, with forecasts that are validated with later known outcomes. The AHP application in this article provides a further test of the method in the prediction of a LRT corridor selection decision outcome.
Figure 1. LRT Corridor and Route Alternatives, Memphis

Source: Adapted from MATA. 1997. Alternatives Analysis Study.
The AHP model used in this article describes a framework of the actual planning process implemented in Memphis by the public transit planning authority. The model is a case-specific mapping of the actual light rail decision making in the city. But more importantly, AHP informs as well as is informed by the decision-making and planning processes in the city. Viewed methodologically, the procedure is intended to facilitate the public transportation decision-making process generically, reflective of federal New Starts guidelines as well as local priorities and preferences of multiple participant groups. Instead of being viewed as yet another AHP application in transportation, this article is intended as a contribution toward the development of a streamlined and unified procedural framework for the purposes of federally-sponsored local public transit decision making and planning with a potential for systematic comparison of similar experiences in different cities. Thus, this article is a case analysis with a general procedural implication for public transportation decision making. A brief description of AHP follows with particular reference to applications in transportation. The article concludes with reflections on the case-specific planning process.

Multicriteria Public Transportation Decision Making: The Analytic Hierarchy Process

Planners confront complex multicriteria decisions related to alignment alternatives, different transit mode-choice, and air quality and environmental impacts. The decisions commonly involve various interest groups as well as elected officials, governmental agencies, and the general public (see also Meyer and Miller 2001). The decision criteria can be mixed with tangibles and intangibles. Commentators have observed public transportation decision making as both a technical and political process (Wachs 1985). Transportation decision making is also characterized as a process involving multiple participants or “stakeholders” (e.g., Hall 1980; Levin et al. 1999). AHP has emerged as a versatile decision support and evaluation methodology with wide-ranging applications. Transportation planning applications are equally as prolific and diverse: stakeholder preference assessment in transportation planning (Levin et al. 1999), transit market priority analysis (Khasnabis and Chaudhry 1994), transportation system improvement projects (Tabucanon and Lee 1995), and carrier selection (Bagchi 1989). Recent applications include AHP in conjunction with a geographic information system in transit-oriented development (TOD) and in freight terminal location (Banai 2000; Dantas et al. 2001; see also Saaty [1995, 1997] for a review of progress in development and
applications of AHP). AHP provides a tool to help planners structure a complex, multifaceted decision-making process.

In contrast to multicriteria or multiattribute evaluation methods, AHP is a hierarchic, systems-oriented or holistic methodology useful in defining a characteristically multilayered public transportation problem. A typical AHP hierarchy is structured by the relationship of the elements in various levels. The overarching goal is stated at the first level, followed in subsequent levels by the criteria and alternatives. When group participation is essential, the participant groups are specified explicitly as described below. A versatile ratio scale is used to compare elements pairwise for all the levels of the hierarchy—systematically comparing the elements of a level with each of the elements of the previous level, starting with each level subsequent to the goal and ending with alternatives—to compute a composite score of the alternatives. For a thorough account of the underpinning philosophy, measurement theory, and methodology of AHP, see Saaty (1996), Forman (1993), and Saaty and Vargas (2001). The transportation planning application described here uses a commercially available software for AHP, Expert Choice (2000). The criteria to assess LRT corridors and routes are varied, and thus the measurement of the intensity of the multiple criteria involves different rating, step, and utility functional types that are supported by the software and shown later in this article. A simple example of the rating methods of AHP is given to determine best LRT corridor and route alternative (Figure 2). The larger application of AHP is given below.

**Figure 2. A Simple Hierarchy for Determining Best LRT Corridor, with Linear (L) and Rating (R) Functions**

<table>
<thead>
<tr>
<th>L1: Goal</th>
<th>L2: Criteria</th>
<th>L3: Rating Intensity</th>
<th>L4: Alternatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best LRT Corridor</td>
<td></td>
<td></td>
<td>Corridor A (0.225)</td>
</tr>
<tr>
<td></td>
<td>Mobility (L) (0.540)</td>
<td>Low (472 pop/sq.mile)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TOD (R) (0.297)</td>
<td>High (1,720 pop/sq.mile)</td>
<td>Corridor B (0.540)</td>
</tr>
<tr>
<td></td>
<td>Cost (L) (0.163)</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Moderate</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>High Potential</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low ($525,862 /mile)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>High ($425,287 /mile)</td>
<td></td>
</tr>
</tbody>
</table>
In Figure 2, three criteria are used to determine the best LRT corridor alternative: mobility to job centers, TOD impact, and (operating) cost. First, the relative importance of the criteria is determined. A rating scale is then developed to evaluate alternatives. The relative importance of the criteria is determined though the paired comparison method of AHP. We use the nine-point (1–9) numerical scale of AHP, defined as: equal importance, when two activities contribute equally to the objective (); moderate importance of one over another (3); essential or strong importance (5); very strong importance (7); extreme importance (9). Intermediate values between two adjacent judgments are 2, 4, 6, 8, or finer ones using decimals, for example 1.1, 1.2, ..., 1.9.

Mobility is given near moderate importance (2) in comparison to TOD, and moderate importance (3) when compared with cost. Finally, TOD is given near-moderate importance compared to operating cost, shown in the paired comparison matrix in Table 1.

<table>
<thead>
<tr>
<th>Best LRT Corridor</th>
<th>Mobility</th>
<th>TOD</th>
<th>Cost</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobility</td>
<td>1</td>
<td>2.0</td>
<td>3.0</td>
<td>0.540</td>
</tr>
<tr>
<td>TOD</td>
<td>1/2</td>
<td>1</td>
<td>2.0</td>
<td>0.297</td>
</tr>
<tr>
<td>Cost</td>
<td>1/3</td>
<td>1/2</td>
<td>1</td>
<td>0.163</td>
</tr>
</tbody>
</table>

The reciprocal values are automatically determined, since the paired comparison method in AHP is with a reciprocal matrix. The relative importance or weight of the criteria is determined with the robust method of estimation in AHP (eigenvector, or characteristic root method; see Saaty 1996, Saaty and Hu 1998) also shown in Table 1, with mobility as the most important (0.540), followed by TOD (0.297), and cost (0.163). Furthermore, the paired comparisons of the factors are done with good consistency (Table 1). When consistency ratio CR exceeds 10 percent, the comparisons should be reconsidered so as to improve upon logical consistency (see Saaty 1996). However, inconsistency is an indicator of transitivity in judgments, which arises naturally in decision making, particularly in situations when the criteria are diverse and there is uncertainty in the environment of deci-
sion making. AHP is the only multicriteria evaluation method with which the error in judging the relative importance of factors by means of relative measurement can be detected and corrected with new observation, reflection, and discussion. Thus, the potential contribution of the AHP in transit planning is suggested as: (1) decisions involve multiple criteria or objectives; (2) the criteria are mixed tangibles and intangibles, some of which—tangibles—have no underlying scales; and (3) the relative importance or priority of the criteria represents preferences and priorities of multiple participants or “stakeholders” in the planning process through observation, reflection, communication, and negotiation.

If the criteria are “abstract” or ambiguous, they are made “concrete” in a specific context with subcriteria rating scales. The rating intensity scales indicate the desirable thresholds of the criteria that must be met in accordance with local priorities and site-specific conditions. In the example presented in Figure 2, they are shown as the subcriteria of mobility (low, high), TOD (low, moderate, high), and cost (high, low). The subcriteria are stated with a semantic scale that uses words (e.g., low, high). The intensity of the ratings is determined by paired comparisons (relative measurement). Thus, even the semantic scale is a ratio scale (not an ordinal scale), making arithmetic operations meaningful in a spreadsheet that contains mixed criteria with data values that require different functional types, ratings, utilities, or priorities entered directly to determine the total score of alternatives. The score of the alternatives is determined in a weighted linear summation method. The total score of each alternative (rows) is the weighted sum of the rating scores for the alternative across all the criteria (columns). The weighted linear summation calculation is aided in the AHP software with its spreadsheet function called “data grid.” The scores of the alternative corridors (A and B) are shown (Figure 2), with corridor B as the better choice. Since we are using a ratio scale, we can determine the rank (ordinal) as well as interval, namely that corridor B is 2.4 times better than corridor A (ratio).

In this simple example, two different types of rating intensity scale functions are used: (1) linear function and (2) rating function (denoted by L and R, respectively, in Figure 2). The choice of function types reflects the type of criteria as well as the available data on the criteria to be measured. In the measurement of mobility, for example, an increasing linear function is used, with upper and lower bounds that are determined in context. In the example above, the higher (utility or satisfaction) value of mobility is determined by the increasing value of density (pop/sq. mile), which is catered to by a LRT corridor.
Conversely, the higher utility of operating cost ($/mile) is determined by a decreasing function. The land-use criterion, TOD impact of LRT corridor alternatives, is measured by three ratings (low, moderate, high potential) with increasing utility. The existing transit planning study in the city used a similar scale for rating alternative corridors. In the larger application that follows later in this article, three types of functions—linear, step, and ratings—corresponding to different criteria and subcriteria are used (Figure 3). For example, a step function is used in the measure of percent of population in poverty. As noted above, the choice of rating functions (discrete or continuous) is determined by the type of criteria, available data, as well as empirical studies (for example, see Pushkarev and Zupan 1977; and Davis and Seskin 1996 for discussion of density function). The paired comparisons method is used even when factors or criteria require ratings. In such cases, the paired comparison method determines the ratings intensity scales, which are derived by paired comparison of the criteria on a ratio scale. Thus, the various scaling methods applied to the multiple criteria have a common ratio scale (0–100 percent) with which the total score of each alternative weighted by the importance of criteria can be determined, all on a common ratio rather than ordinal scale. AHP provides a multicriteria evaluation with a robust ratio scale method that is helpful in land-use transportation planning decisions with multiple and diverse criteria, like the LRT corridor selection problem. In addition to the flexibility of a robust ratio-scale, AHP is a multicriteria evaluation method with a structural (hierarchical) property that aids in the challenging and creative part of a complex, multifaceted land-use transportation problem: formulation.

AHP was recently introduced to the FTA as an executive decision-making tool for the resource allocation of contract funding totaling approximately $40 million for the U.S. DOT/FTA Capital Project Management Oversight (CPMO) program (Rye and Haider 2000). The result of the study indicated that the FTA did not use a structured process “or methodology to measure or quantify benefits in their decision making.” The study concluded “…such a situation can impede sound decision making for resource allocation issues” (p. 15). The FTA’s comments and feedback for not using AHP in decision making were:

...FTA decisions often involve ambiguity, conflicting goals due to the multiple objectives, trade-offs and frequently more than one decision-maker. [The director] further states that in governmental or public service domains, the objectives can be more social or political rather than financial or functional. The director indicated that when making important FTA decisions, all objectives are typically
Figure 3. A Unified Framework for Multicriteria Public Transportation Decision Making, with Linear (L), Rating (R), and Step (S) Functions

Note: The two columns of ratings for corridors and routes shown together in this figure are modeled separately in the AHP software to determine the scores of the alternative corridors and routes.
considered to have equal value of importance in order to compromise dissimilar beliefs and opinions. (Rye and Haider 2000, p. 16)

Decisions at the local level also encounter a similar context of multiplicity of objectives and of participants, as the currently ongoing Memphis corridor and route selection process suggests. The increasing popularity of AHP as a multicriteria evaluation methodology is attributed to its flexibility to deal with ambiguity of multiobjectives, with mixed tangible and intangible criteria or objectives (social, political, financial, functional), and group decision making (see, for example, Forman and Selley 2000). Thus, AHP provided a plausible methodology for our case study. AHP is used to show how complex multilayered public transit planning and decision making is unified to account for federal and local criteria, different participants, and trade-offs among multiple, diverse criteria, and choice of corridor and route alternatives.

A Unified Framework for Multicriteria Public Transportation Decision Making: The Corridor and Route Selection Process

An AHP model of a LRT corridor and route selection process is shown in Figure 3. The goal of the process is to select the best corridor and route (Figure 1). Participants include politicians, bureaucrats, community leaders, and the general public. The criteria used in the FTA New Starts rating process are incorporated. The continuum local and federal (under the criteria) connotes the notion of adaptation of New Starts general criteria in response to local context. The general criteria formed the basis for the selection of factors used in a questionnaire to solicit inputs from participant groups. The regional transit plan (1997) provided a local context. NHRCP (1999) best practice guidelines were considered also. The general criteria were ranked based on the priorities of local decision-makers. AHP is used to show how the relative importance or weights of the criteria determine corridor and route alternatives selection. The overarching purpose of the model is to help planners to structure and unify the transit planning decision-making process, to derive priorities locally in relation to federal criteria, to ensure that local decisions made are consistent with the criteria in a transparent process, and to make certain that different interest groups are equally represented. The model is intended to facilitate efficient decision making at the local level while a project is competing nationally for federal funding.
Public Transportation Decision-Making

**Participant Group**

The account of the London Motorways in the 1960s and 1970s in Peter Hall’s (1980) *Great Planning Disasters* is instructive in the present context. In retrospect—traceable to Abercrombie’s plans—Hall shows how the outcome of the transportation planning process is influenced by decisions and actions of multiple participants—experts, politicians, and community residents—and how specific transportation solutions, like highways or public transit systems, give rise to controversy and uncertainty when viewed from multiple policy objectives. Abercrombie’s famous plan, for example, involved multiple mobility, land-use, and environmental quality objectives. Hall (1980) generalized various sources of uncertainty that arises in collective decision making: the uncertainty about the planning environment, about value judgments, and about decisions of participants.

The (hierarchical) structural property of AHP frames public transportation planning generally in such a way that alternative transportation choices are determined by the relative importance given to a set of criteria (policy objectives) by participants in a collective decision-making process. Although multicriteria methods are increasing used in transportation planning, AHP provides a technique effective in decision making in the face of uncertainty, ambiguity, and limitation of information (Forman and Seeley 2000; for a review of multicriteria methods, see Nijkamp and Voogd 1983; Yoon and Hwang 1995).

The nine criteria that were ranked independently by separate members of a regional rail steering committee (MATA 2001) are incorporated in the unified AHP framework (Figure 3). Since politicians and decision makers like the flexibility to have an input on criteria weights and may not necessarily understand the complex nuances of the AHP—or any other modeling steps or procedures of transportation planning—the planner can assume the role of facilitator in the decision-making process, which is rational and structured as well as transparent and which can identify possible inconsistencies in assessing criteria, or detect if a special interest group is influencing a collective decision process in such a way that favors one group over others. Thus, a unified AHP framework of multicriteria decision making potentially guides and informs a collective decision-making process that aims toward a consensus on decisions. In addition, AHP can be used as a public educational tool to structure and facilitate complex multicriteria decision making.

The local transit authority held several public meetings and presentations in which public comments were solicited. Written comments were also encouraged to be
submitted. The relative weights of the criteria are commonly derived through paired comparison in AHP (as in the simple example above). Furthermore, to facilitate the comparisons, AHP protocol calls for the criteria to be grouped within a limit of 7+/− 2. This limit serves a practical purpose to avoid confusions in paired comparisons of criteria factors when considered simultaneously (see Saaty 1996; Simon 1981; Miller 1956). This step was skipped in this case because a questionnaire with the same upper- (10) and lower-bound values (1) to the AHP scale was already used by the local transit authority to determine criteria ranking. The ratings were converted to a scale of relative importance of the criteria expressed in percentage (0–100%), similar to the AHP ratio scale. However, the AHP has a robust ratio scale that is a natural method of ranking criteria (paired comparisons) than the weaker ordinal scale used in the questionnaire. Above all, with the AHP method the inconsistencies within the decision-making process arising in ranking of the criteria are gauged, and the means to address the incontinences are provided through observation with new information, reflection, deliberation, and communication. The AHP framework facilitates a rational planning process that is observational, reflective, deliberative, and communicative.

The importance of the criteria as seen by participant groups is shown in Figure 4. Criteria rankings are close among the various groups, with a consensus on corridor choice (Figure 5). The relative weights of the criteria were next used to assess the subcriteria and alternative corridors (Figure 6).

Southeast is the best corridor, with a score of 0.89 (Figure 6). The screen capture shown is the spreadsheet platform (data grid) of the AHP software (Expert Choice). The various rating functions used (increasing, step, decreasing, and the like, also graphed) are shown as the column headings. The “local” weights of the criteria (denoted by L) are shown for one group of participants: politicians. The “Total” column gives the score of each alternative (corridor) in a weighted linear summation: The rating scores of each alternative are multiplied by the weights of the criteria and the results summed across all the criteria. The “Ideal mode” of synthesis is used when we are interested in a choice of one—and only one—in a set of alternatives, and the remaining alternatives are considered as irrelevant. (For a discussion of the alternative modes of synthesis in AHP, see Saaty 1996.)
Figure 4. Criteria Rankings by Participant Group

Corridor Selection Criteria
- Politicians
- Bureaucrats
- Community Leaders
- Average

Figure 5. Scores of Alternative Corridors by Participant Group

Corridor Alternatives
- Politicians
- Bureaucrats
- Community Leaders
- Average
Figure 6. Assessing LRT Corridor Alternatives with Various Criteria Scaling Functions of AHP in a Data Grid, Ideal Mode (screen shots)

Ranking Corridor Alternatives with Data Grid in Ideal Mode

Mobility to Job Centers (Increasing Function)

Transit-Oriented Development (Rating Function)
Public Transportation Decision-Making

**Scenarios**

The local importance of the criteria is now known; however, the importance potentially given to the criteria by the FTA New Starts program is unknown (see GAO 2005). To deal with the uncertainty, a sensitivity analysis is performed to determine possible effects of weighing FTA criteria differently and the influence on outcome. Initially, all federal criteria are assumed as equally important based on a view of an administrator noted above. What if, for example, costs and land use are given strongly more importance (weight)? Knowledge of impact on outcome can help localities prepare strategically should any of the multiple criteria be given increased scrutiny or priority due to budgetary constrains or competition.

We used scenarios to examine the impact of the varying importance of the criteria on the ranking of the corridors. The relative importance of the criteria is changed in a scenario to reflect the increased weight (from 11.8% to 29.6%) given to TOD with a dynamic sensitivity analysis. This increase in the priority of TOD criterion results in decreases in the weight of mobility to job centers (from 13.2% to 10.6%), and mobility of low-income residents (12.4% to 11.5%). The results merit further investigation of the corridor alternatives to determine the proportion of jobs relative to housing in a TOD; that is, whether TODs are predominantly locations in which to live or to work. The weight of mobility of general public (from 12.1% to 20.4%) is increased, whereas, the decreased relative importance of mobility to jobs (centers) conforms to the definition of TODs as locations with both housing and jobs. The impact of the increased priority of TOD on the rest of the criteria and the rank order of alternatives are as follows:

- Mobility to jobs (13.2% to 10.6%)
- Mobility of general public (12.1% to 20.4%)
- Mobility of low-income residents (12.4% to 11.5%)
- TOD (11.8% to 29.6%)
- Operating costs (11.2% to 6.2%)
- Capital construction costs (10.8% to 6.4%)
- Use of shared rights-of-way (9.9% to 3.8%)
- Traffic congestion (9.8% to 7.6%)
- Impact on sensitive areas (8.8% to 4.0%)

The results in rank order are: southeast (47.0%), south (34.1%), and north (18.9%) corridors. Even in alternative scenarios in which the priority of low-income resi-
dents and operating cost are doubled to 23.6 percent and, 21.7 percent, respectively, or the priority of operating and capital construction cost are doubled, the southeast preserves its rank as the best corridor, followed by the south and north corridors. Other scenarios could address the impact of the state growth plan legislation, parking restrictions, local area or district plans, zoning regulations with land-use/transit joint development strategies, and cost effectiveness.

Discussion
This section provides a brief reflection on the application of the overall framework described by AHP. The purpose of this case analysis was to develop a transparent structural framework that unifies as well reflects the discrete steps of a public transportation decision-making process. Furthermore, this framework could be used repeatedly by others doing similar project planning, albeit with different priorities reflecting conditions in different contexts. Arguably, there are variations across cities reflecting differences in politics, economics, institutions, and spatial form (e.g., density). Knowledge of the differences and the effect on outcome is important in itself in a systematic comparison among the cities by using the unified, multicriteria framework of public transportation decision making to reveal different local priorities and influences notwithstanding FTA guidelines. In effect, the unified framework would reveal the differences among the cities owing to the uniqueness of the local context even when the same federal guidelines are used. In addition, the framework may be used in the city longitudinally to track changes in the priorities reflecting the environment of public decision making, political, economic, and institutional dynamics. Thus, the framework would aid in institutional learning and streamlining the planning process with each subsequent formulation and implementation. Since the basic idea of the framework was to provide a “mapping” of LRT decision making throughout the planning process, we give a brief outline of the steps in context.

Goal
The community's goal stated in the region's long-range plan is to improve the quality of life through the use of LRT. For purposes of transportation planning, the region is divided into corridors, and the goals and objectives of the community are transformed into a set of criteria. The criteria are then used to evaluate the corridors, based on a set of objectives and further refined to make specific decisions as to the best alternative routes within a corridor. The structural property of AHP is helpful here since, by definition, a hierarchy is comprised of a set of levels that
begins (atop) with the general or abstract elements (goals) and ends (at the lowest level) with the concrete or specific elements, which are the alternatives (corridors and routes in our case analysis). Commentators generally characterize this process as ill-defined, technical, and political, noting the challenges of translating the general community’s goals into specific transportation alternatives and policies. As noted above, the process involves consideration of mixed tangible and intangible factors, and requires the flexibility of a multicriteria evaluation method with a structural property helpful in problem formulation.

Participants
The planning process requires local participation in public transportation decisions. Ideally, every participant group has an equal voice in a collective decision process. But if the process is skewed in favor of one group without benefit to another, it is desirable to map the unequal weights of different groups in a transparent planning process.

During the planning process, some groups could exert an unfair influence and cause an “irrational” decision not in the public interest. In our case analysis, small business owners were opposed to the short-term construction impacts and even threatened legal action to stop an alternative (LRT route) in their “backyard.” Interestingly, the AHP model predicts Alternative 1 (Madison Ave.) as the best LRT route. The affected parties can influence local politicians and decision-makers who, in turn, can affect the outcome of the decisions. The rational framework and process of the AHP model allows for an account of the power (weights) of the influences of different parties. Group dynamics—within and between group discussions—and relationships could be similarly mapped while discerning an emerging consensus or divergence of group objectives and preferences.

Criteria
The criteria are multifaceted, derived from local and federal goals and objectives that satisfy the problems and needs of the metropolitan community and the region, ranging from transportation to land use and to the environment. The rational framework accommodates the multiplicity of goals and the intensity of their diversity. A practical consideration is the question of how to assess the criteria and how to gauge discussion of “what is important” to the decision-makers. More discussion and dialogue of what is important (relative importance of criteria) is helpful with the (questionnaire) survey of participants, elected officials, expert groups within city agencies, community leaders, public-private business representatives, and the general public. Committee membership changes, while long-term
projects are in the planning process. Periodically, the purpose and requirements of the transportation program/project should be reexamined in relationship to the criteria and the ratings (weights). The facilitator has a pivotal role in encouraging interaction and communication among the decision makers. Communication is important to address any inherent problem due to the abstract nature of the goals that are translated into criteria and which are ultimately used to evaluate performance.

**Rating Intensity**

Good, reliable data are not always readily obtained due to limited coordination with multiple participants, consultants, and other agencies involved in public transportation planning. This is especially true with complex, technically demanding data and computational requirements of travel demand forecasting and standard evaluating procedures. The AHP framework provides an alternative forecasting and evaluation methodology effective in the face of uncertainty. Specifically, the paired comparison approach with alternative rating scales surmounts the limitations encountered in data availability. We used verbal rating scales instead of step or linear function types to compensate for the vagueness of the available estimates inherent in the variable measured. One example is the measure of an alternative corridor for TOD potential, which used a rating function, in contrast to Mobility to Job Centers criterion that used a linear function with estimate of the population density (pop/sq. mile; see Figures 3 and 6). The number of households may be known for certain corridor or mode-specific alignment with precedent or baseline data but unknown for still others, contributing to the uncertainty in forecasting. Even in situations where data are available, judgment and experience play a role in the interpretation and assessment of brute data. Paired comparisons are done to determine the intensity of the (verbal) ratings on a ratio scale (0–1) comparable to fuzzy methods of deriving membership (functional) values (see Banai 1993). Thus, even verbal rating scales have numerical values that are quantified on ratio scales. Experience shows that information is not always available or may be incomplete. Alternative rating methods of AHP allow the flexibility of estimation in the face of incomplete information.

**Alternatives**

The three corridors were the result of an earlier screening process involving a larger number of corridors based on similar criteria. The AHP model described could be structured differently. Within each corridor, alternative transit modes as well as alternative routes could be examined. The criteria are then used to assess
alternative routes, shown in Figure 7, for further assessment. Similarly, the general structure can be modified to add another level, for example alternative modes, for further analytical consideration. Alternative alignments can be identified to discern which may be best based on criteria and relative weights, and scenarios can be developed to predict possible “what if” situations. New starts criteria can be set up to give an assessment of how particular projects fare when compared to one another based on similar criteria weighting assumptions.

A Task Force (1994) concluded that LRT was a feasible transportation solution for the community’s problems. However, this recommendation preempted consideration of alternatives such as monorail, bus rapid transit, or other transportation systems management (TSM) approaches. TSM alternatives should be considered early in the process. TSM is viable once the efficacy of the alternative routes is determined. Later planning studies considered alternatives required by FTA funding procedures, including monorail service on the I-40 right-of-way (see Figure 7). Public comments focused on enhanced bus service within a corridor as an incremental step toward LRT. The moral of the story: Don’t rule out competing modes of transportation, and allow for inputs from the public with the discussion of alternatives. Concomitantly, a model of this process should include alternatives.

Figure 7. Determining the Best LRT Route Alternative

<table>
<thead>
<tr>
<th>Ideal mode</th>
<th>Politicians Mobility to Job Centers (L=141)</th>
<th>Politicians Mobility of General Public (L=126)</th>
<th>Politicians Mobility of Low Income Residents (L=135)</th>
<th>Politicians Operating Costs (Annual) (L=113)</th>
<th>STEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative 1 (LRT on Madison Ave)</td>
<td>.726</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>193</td>
</tr>
<tr>
<td>Alternative 2 (LRT on Lamar Ave)</td>
<td>.489</td>
<td>Medium</td>
<td>High</td>
<td>Medium</td>
<td>208</td>
</tr>
<tr>
<td>Alternative 2B (LRT in BHSF Railroad ROW)</td>
<td>.382</td>
<td>Low</td>
<td>Low</td>
<td>Low-Medium</td>
<td>226</td>
</tr>
<tr>
<td>Alternative 3 (Monorail on I-340 ROW)</td>
<td>.474</td>
<td>Medium</td>
<td>Medium</td>
<td>Medium</td>
<td>220</td>
</tr>
</tbody>
</table>
Conclusion and Extensions

The AHP model for the selection of the top-priority corridor was completed in 2000. The model identified the southeast corridor as the best alternative. The AHP prediction later proved to be true when the southeast corridor was selected by the local transit authority board of commissioners (January 2001) based on the recommendation of the regional rail steering committee. The model was relatively easy to use even in the face of limited or incomplete information. The inductive methodology of AHP is useful in situations where deductive, predictive, observational techniques (e.g., regression analysis) encounter a limitation in the absence of precedence or with structural transformation, such as with introduction of a new public transit system. Furthermore, it provides flexibility in measurement that is helpful in situations, for example, FTA New Starts program, where multiple criteria with certain desirable thresholds of intensity must be considered strategically and adaptively responsive to local priorities and site-specific-conditions.

AHP is the only multicriteria method with a built-in procedure to account for the inconsistency of judgments of participants in the process of evaluation of a set of multiple criteria. A robust scientific framework is provided to gauge the consistency and efficacy of interpreting tangible and intangible data directly and inductively, rather than indirectly and deductively.

We then used AHP to predict the selection of the locally preferred alternative (LPA)—to use terms from FTA—best alternative alignment within the selected (southeast) corridor (Figure 7). The AHP model predicts alternative 1 (with a score of 0.726) as the best alignment selection. The LPA will be decided in the future by the local transit authority board of commissioners with concurrence from the city council and metropolitan planning organization.

The AHP-aided unified framework could be used for other public transportation planning purposes, such as in highway alignment, public transit mode-choice, and route selection decisions, which are increasing seen as a multicriteria decision-making process. By means of sensitivity analysis, alternative scenarios could be examined to determine outcomes based upon the relative importance of the criteria from local or federal agencies. Similarly, the influence of different participant groups involved in collective decision making on the selection of alternative alignments, modes, routes, and the like could be examined. The AHP prediction of such outcomes can be used as a basis for further negotiation and conflict resolution as well as for cost-benefit analysis and determination of trade-offs. Moreover, the inclusive decision-making framework allows for investigation of how different
groups’ values commensurate or conflict with different goals. The valuation (qualitative and quantitative) based upon multiple criteria and revealed preferences (values) of different participants can fill a gap created by methods known with value distortions, with a single economic efficiency criterion (e.g., benefit-cost). However, standard benefit-cost analysis plausibly supplements, if included in, the unified AHP model above in a further detailed economic efficiency analysis of LRT alternatives.

**Acknowledgments**

The author would like to thank the editor of the *Journal of Public Transportation* for the opportunity to address the comments of the reviewers in an earlier version of this paper. The author gratefully acknowledges the contributions of John Lancaster, Memphis Area Transit Authority, with data collection and analysis, and insightful discussions of the Memphis LRT planning process.

**Endnotes**

1 Research related to this article, including urban development sustainability analysis, can be viewed at http://www.people.memphis.edu/~rbanai.

2 The FY2003 annual report on New Starts project status is as follows: 25 projects have full-funding grant agreements (FFGAs); 11, in final design; 39, in preliminary engineering; and 142, additional studies and projects authorized in TEA-21 in the early planning stages or alternatives analysis. As the competition for funds increases, project sponsors have increased local matching capital funds to 50 percent, instead of the past norm of a 20 percent local match with 80 percent from the FTA. The funding split for road and highway funds is still 80 percent federal and 20 percent local, provided the 20-year long-range transportation plan meets air quality guidelines (FTA 2002).

3 The 1994 study assumed the economy of light rail with the use of the existing rail corridors. However, the Regional Rail Program (2001) raised the issue of existing right-of-way availability.
References


GAO. 2005. GAO report to congressional committees public transportation: Opportunities exist to improve the communication and transparency of changes made to the new starts program (June).


**About the Author**

Reza Banai (rbanai@memphis.edu) is professor of city and regional planning in the School of Urban Affairs and Public Policy at the University of Memphis. His recent research includes transit-supportive site suitability analysis aided with GIS and AHP as a decision-support system.
Where Transit Use is Growing: Surprising Results

Gregory L. Thompson, Jeffrey R. Brown, Rupa Sharma, and Samuel Scheib
Florida State University

Abstract

This article investigates whether transit’s fate is tied to the last vestiges of old urban forms or whether transit is finding niches in the new, largely suburban urban forms that increasingly have manifested themselves since the 1920s. The hypothesis is that most growth is in census regions with the strongest vestiges of older urban forms centered on CBDs. The hypothesis was tested by documenting how transit performance changed between 1990 and 2000 in U.S. metropolitan areas with more than 500,000 people in the year 2000. Results show that, for MSAs with fewer than 5 million people, transit use has been growing faster than very rapid population growth in the West region, but not elsewhere in the country. The conclusion is that transit growth is not tied to old urban forms. A future article will explore causality of transit use growth and service productivity change.

Introduction

This article documents changes in the magnitude of transit service, use, and productivity in U.S. metropolitan areas during the 1990s. We include transit and population variables for all metropolitan statistical areas (MSAs) and consolidated metropolitan statistical areas (CMSAs) with a year 2000 population greater than 500,000 people. Our purpose is to determine whether regional variations in service and usage exist for MSAs and CMSAs in different population size categories.
This exploration is prompted by earlier work suggesting that transit use per capita may be increasing in rapidly growing parts of the country, contrary to the commonly held belief that transit demand exists primarily in older industrial-era cities (Thompson and Matoff 2003). The earlier work was based on transit performance in only nine MSAs and prompted a desire for a more systematic documentation of transit trends, whose relationships ultimately would be analyzed statistically. The statistical analysis does not occur in this article, which describes regional and size category trends that suggest transit use and productivity are developing in ways very different from commonly held perceptions.

**Relationships among Urban Form, Travel Patterns, and Transit Performance**

The national decline in transit share is well studied. Specifically, Pisarski’s (1996) analysis of the census’s journey to work questions for 1970, 1980, and 1990 documents suburbanization of jobs as well as residents, the rise of two-worker households, the ever-increasing use of single-occupant autos, and declines in the share of travel of all other transportation modes, including public transportation. Pisarski found that the greatest absolute declines in transit work-trip mode split occurred in central cities, but greater relative declines occurred in suburban rings. The implication is that, while transit does not do well in any urban environment, it does least well in the suburbs. However, he does note several anomalies, including Houston, Los Angeles, New York, Orlando, Philadelphia, Phoenix, Rochester, San Diego, and Tampa.

Pucher and Renne’s (2003) analysis of the National Household Travel Survey (NHTS) allows for inferences about characteristics of transit demand only at the U.S. Census-defined region level, due to NHTS sample size. They note that the proportion of travelers using transit for all trip purposes (not just work trips) is larger in the East and Midwest than in other regions of the country, though there is some strength in the West. They infer that transit demand is greatest in those urban regions that experienced much of their growth when transit was the dominant mode of urban transportation, generally before the late 1920s.

Both Pisarski and Pucher and Renne conclude that, while transit use is tiny and declining, it is highest in urban structures tied to the past rather than the future. That is, it is tied to central business districts (CBDs) and dense older (streetcar) suburban areas rather than to the modern suburbs. Such thinking is encapsulated
in the often-cited work of Pushkarev and Zupan (1977, 1980), though Cervero and Seskin (1995) raise questions about the interpretation. If true, it seems likely that as the forces of decentralization further erode these old urban structures, remaining transit use will completely slip away.

We are skeptical, however, of the notion that transit’s fate is tied to the old urban forms. Tantalizing evidence suggests transit might have a niche in auto-dominated suburban areas that are rapidly developing. Pisarski’s anomalies imply this. Work by Thompson and Matoff (2003), which documents transit performance in nine metropolitan regions, also supports this theory. While the Thompson and Matoff work is not comprehensive in its choice of metropolitan regions, it shows that there are metropolitan regions with transit growth. The regions in their study are not where old urban forms remain influential but regions characterized by rapid population and employment growth, mostly in the suburbs. In some cases, they are regions with very weak CBDs where transit use has been relatively low.

**Methodology**

Transit performance in urban regions varies by population size (Downs 2004; Pucher 2004). To control for population size, we assigned each of the 81 MSAs and CMSAs with more than 500,000 persons to one of four size categories: 500,000 to 1 million people (small MSAs), 1 million to 5 million people (medium MSAs), 5 million to 10 million people (large MSAs), and greater than 10 million people (mega MSAs). To explore regional variation (our proxy for old versus new urban forms), we organized each size category by census region (Northeast, Midwest, South, and West). Table 1 shows the distribution of MSAs in each category based on population and region. We then calculated performance statistics for the aggregated transit service in each MSA. Finally, we reported the median values of each statistic in each category.

The methodology ties together two databases, the National Transit Database (NTD) and the U.S. Census (U.S. Census Bureau 2000, 2003). The NTD, which we accessed using the Florida Department of Transportation’s Florida Transit Information System (FTIS), provides information on system performance by transit agency for all agencies that received federal aid (FDOT 2004). We identified all transit agencies operating in each MSA or CMSA in the study and aggregated their performance statistics. Our approach to examining transit performance thus
differs considerably from the often-cited study by Hartgen and Kinnamon (1999), who conduct system-specific evaluations.

The choice of the MSA or CMSA as the unit for analysis is based on the need to avoid misattribution of the population responsible for generating transit patronage. In the NTD, patronage and other transit statistics are readily available only for entire transit systems. In calculating such performance measures as passenger miles per capita or vehicle miles per capita, we found it necessary to identify the population associated with the service and ridership. Too much population is attributed to a transit system if the service area of the system does not fill up most of the area for which population is counted; too little is attributed if transit systems spill over into adjoining population areas. Figure 1 shows such situations.

### Table 1. Distribution of MSAs by Region

<table>
<thead>
<tr>
<th>Census-Defined Geographic Region</th>
<th>Mega MSAs</th>
<th>Large MSAs</th>
<th>Medium MSAs</th>
<th>Small MSAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwest</td>
<td>2</td>
<td>9</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>South</td>
<td>2</td>
<td>18</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>West</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>7</td>
<td>40</td>
<td>32</td>
</tr>
</tbody>
</table>

### Figure 1. Transit Service Provision in a Hypothetical Urban Region

![Figure 1. Transit Service Provision in a Hypothetical Urban Region](image)
Where Transit Use is Growing

Difficulties in correctly attributing population to transit statistics also occur when more than one transit system serves the same population. The propensity of the population to use transit is spread over the two or more transit systems serving it, so to understand how much ridership the population produces, we must aggregate ridership for all of the relevant transit systems. Figure 1 shows two MSAs that have grown together and that are served by several transit systems exhibiting all of these problems. A solution to overcoming the problems is to aggregate performance indicators for all transit operators in the smallest geographic area containing all of the services. In Figure 1, this is the CMSA, but if the commuter rail authority did not exist, the aggregation could be for each of the two MSAs.

Choice of Performance Measures
Our goal in comparing transit systems is to assess trends in demand, supply, and productivity, using commonly accepted measures of transit performance. Because MSAs vary in population, even within the population groups that we defined, we need to express demand and supply on a per capita basis. We use passenger miles per capita for demand and vehicle miles per capita for supply. We also determine trends in productivity of transit service between different urban regions. Fielding (1987) identified several measures of productivity: pairwise combinations of resources that transit systems consume, service they provide, and the degree to which that service is consumed. Typical measures are operating expense per vehicle mile, passenger miles per vehicle mile, and operating expense per passenger mile (Fielding 1987). We do not focus on managerial practices that affect the cost of providing vehicle hours or miles, an important topic that merits a different article, but instead focus on the productivity of each vehicle mile operated. Are there regional differences here? Is transit service in rapidly growing, auto-oriented metropolitan areas less productive than that in denser, more stable urban areas? Trends in passenger miles per vehicle mile (load factor) will tell us.

We do not show cost per passenger mile, primarily because it is a composite variable that results from dividing the cost of operating a vehicle one mile by its productivity. As we are addressing the productivity issue, the remaining insight resulting from cost per passenger mile comes from the cost of operating a vehicle one mile. Many variables affect this, such a labor contracts, congestion levels, and percent of service provided by rail to name just three, and these vary in unpredictable ways from one metropolitan area to another. We cannot interpret cost per passenger miles in terms of changing urban form. Thus, we do not include it.
Results
We examined transit performance by grouping the 81 MSAs and CMSAs with more than 500,000 people by population size and organizing these groups by census region. Results for each MSA population group are presented below.

Transit Performance in Mega MSAs
The dataset contains two mega MSAs (population greater than 10 million), one (the Los Angeles-Riverside-Orange County, California CMSA) is located in the West region; the other (the New York-Northern New Jersey-Long Island, New York New Jersey Connecticut Pennsylvania CMSA) is located in the Northeast. The dataset is compared in Table.

Transit was far more important for the New York region on almost every dimension in 1990, though the magnitude of transit change between 1990 and 2000 was more comparable. In 1990, the New York region supported almost four times as many transit service miles per capita as did the Los Angeles region while more than five times as many passenger miles per capita made use of transit in the New York region. In 1990, each mile of transit service in New York carried nearly 50 percent more passengers than that in Los Angeles.

Between 1990 and 2000, the New York region hardly expanded transit service, while the Los Angeles region expanded transit service miles by 23 percent per capita, reflecting in part the introduction of a subway line and a far-flung system of commuter trains while expanding light rail transit service that was begun the previous decade. Transit usage per capita increased in the Los Angeles region by about 11 percent, less than half the increase in service. Usage also increased by about 11 percent in the New York region, even though there was negligible service growth. Accordingly, over the decade, transit in the New York region became more productive while it became less productive in the Los Angeles region. These developments occurred despite the fact that the Los Angeles region is the densest MSA in the country. Comparison of the two mega MSAs supports the hypothesis that transit performs best in regions with traditional urban forms.

Transit Performance in Large MSAs
The dataset contains seven large MSAs (population between 5 million and 10 million). One of the four regions (West) contained one large MSA, and three regions (Midwest, Northeast, and South) each had two. Median populations in 2000 ranged from 6,003,782 to 7,306,984. Table 3 shows that service provision in 1990 varied with the Northeast and West grouped at the high end and the South and
### Table 2. Transit Performance in Mega MSAs Grouped by Region

<table>
<thead>
<tr>
<th>Census-Defined Geographic Region</th>
<th>N</th>
<th>Year 2000 MSA or CMSA Population</th>
<th>1990 Vehicle Miles per Capita</th>
<th>1990 Passenger Miles per Capita</th>
<th>1990 % Change in MSA Population since 1990</th>
<th>% Change in Vehicle Miles per Capita since 1990</th>
<th>% Change in Passenger Miles per Capita since 1990</th>
<th>% Change in Passenger Miles per Vehicle Mile (Load Factor) 1990-2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast</td>
<td>1</td>
<td>21,199,865</td>
<td>39</td>
<td>782</td>
<td>8%</td>
<td>2</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>West</td>
<td>1</td>
<td>16,373,645</td>
<td>11</td>
<td>148</td>
<td>13%</td>
<td>19</td>
<td>8</td>
<td>(9)</td>
</tr>
</tbody>
</table>

### Table 3. Transit Performance in Large MSAs Grouped by Region (Median Values)

<table>
<thead>
<tr>
<th>Census-Defined Geographic Region</th>
<th>N</th>
<th>Year 2000 MSA or CMSA Population</th>
<th>1990 Vehicle Miles per Capita</th>
<th>1990 Passenger Miles per Capita</th>
<th>1990 % Change in MSA Population since 1990</th>
<th>% Change in Vehicle Miles per Capita since 1990</th>
<th>% Change in Passenger Miles per Capita since 1990</th>
<th>% Change in Passenger Miles per Vehicle Mile (Load Factor) 1990-2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwest</td>
<td>2</td>
<td>7,306,984</td>
<td>15</td>
<td>259</td>
<td>8%</td>
<td>(6)</td>
<td>(20)</td>
<td>(16)</td>
</tr>
<tr>
<td>Northeast</td>
<td>2</td>
<td>6,003,782</td>
<td>24</td>
<td>405</td>
<td>5%</td>
<td>10</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>South</td>
<td>2</td>
<td>6,414,936</td>
<td>14</td>
<td>198</td>
<td>21%</td>
<td>7</td>
<td>5</td>
<td>(1)</td>
</tr>
<tr>
<td>West</td>
<td>1</td>
<td>7,039,362</td>
<td>24</td>
<td>364</td>
<td>13%</td>
<td>7</td>
<td>2</td>
<td>(4)</td>
</tr>
</tbody>
</table>
Midwest at the low end. This pattern also prevailed for 1990 passenger miles per capita. Productivity was about the same in the Northeast, Midwest, and West, and somewhat lower in the South.

Patterns of growth and decline in the 1990s varied more widely between the four regions. In three of the four regions, service increased faster than population between 1990 and 2000 in the median MSA. In three of the four regions, ridership also increased faster than population between 1990 and 2000 in the median MSA. The Northeast region experienced the largest gains in service per capita and ridership per capita. In the Midwest region, service and ridership failed to keep pace with modest population increases. Only in the Northeast region did productivity improve between 1990 and 2000 in the median MSA. In the other regions, ridership change failed to keep pace with service change, and thus productivity declined.

Overall, the performance of transit in large MSAs constitutes mixed evidence about the hypothesis that transit is performing best in metropolitan areas that are the more traditional in nature. The large MSA in the West region, which provided a high level of service per capita in 1990, is the San Francisco CMSA. The San Francisco CMSA does, in fact, contain a strong CBD and high population densities surrounding it, but most of the growth in transit service in the San Francisco CMSA during the 1990s was in suburban areas, and that growth has attracted usage. Much the same can be said for the Boston region, one of the two MSAs in this category in the Northeast region.

The huge productivity increase in the Northeast largely reflected growth of commuter rail service in the Boston area. Boston’s commuter rail ridership rose by 105 percent during the decade, far outstripping the 57 percent increase in commuter rail vehicle miles. Bus passenger miles rose by 21 percent in Boston, again faster than the 8 percent increase in bus vehicle miles. These changes tend to support the view that transit works well connecting distant suburban commuters with strong CBDs, which the Boston area has.

Dallas was the only other metropolitan area in the large-sized category to experience productivity improvement, though the increase occurred on top of a low base. Dallas also was the MSA area in the large-sized category whose population grew the most rapidly, and it achieved its productivity improvement on top of a very large service expansion that outpaced population growth. The Dallas experience supports the idea that transit can do well in suburban-based environments.
On the other hand, Dallas’s productivity growth was only slightly ahead of productivity change in the Chicago area, which is a much more traditionally structured metropolitan area experiencing slow growth. Chicago enjoyed the second highest degree of transit service and use in the United States in 1990. While Chicago cut back service by 8 percent during the 1990s, usage also fell by 8 percent, resulting in no change in productivity. Chicago suggests that a well-managed transit system can hold its own in a slow-growth traditional urban environment.

**Transit Performance in Medium-Sized MSAs**

The 40 MSAs in the category account for roughly one half of the MSAs in the study. Of the 40 medium-size MSAs, just under half (18 out of 40) are in the South region (see Table 1). Nine medium-size MSAs are in the Midwest, 8 are in the West, and 5 are in the Northeast. Outside of the Midwest and Northeast regions, most of the growth of the MSAs in this category took place during the auto era after World War II.

Medium-size MSAs tended to provide less service per capita than did large-size MSAs, and there was less variation in service provided across the regions of the country in 1990, as shown in Table 4. Aside from the West region, whose median MSA in this category provided 12 service miles per capita, service was uniform, ranging from 7 to 8 vehicle miles per capita. Service usage also was more uniform, ranging from 9 to 10 passenger miles per service mile. Usage per capita ranged from 49 to 08 passenger miles per capita. Usage was highest in the West where the median MSA had 108 passenger miles per capita in 1990, while at the other end of the spectrum, the South attracted only 49 passenger miles per capita.

During the 1990s, the West’s lead in passenger miles per capita grew substantially. Figure 2 shows that only in the West did patronage growth outpace population growth between 1990 and 2000. Figure 3 shows that MSAs in all four regions increased service faster than they added population. Figures 2 and 3 are boxplots that show the distribution of observations for patronage change and service change respectively. The boxes in the plot cover the distance between the upper and lower quartiles of the observations, and the line in the box denotes the median observation. Outliers are marked with an asterisk. The figures clearly show that the median MSA in all four regions increased service faster than they added passengers. The decline in productivity was much less severe in the West region than elsewhere. In the median MSA in the West, the ratio of passenger miles to vehicle miles declined by a mere 3 percent between 1990 and 2000, as compared to declines of 18 percent in the Northeast, 17 percent in the South, and 10 percent...
<table>
<thead>
<tr>
<th>Region</th>
<th>N</th>
<th>MSA Population</th>
<th>CMSA Population</th>
<th>1990 Miles per Passenger Vehicle</th>
<th>1990 Miles per Capita Vehicle</th>
<th>2000 Miles per Passenger Vehicle</th>
<th>2000 Miles per Capita Vehicle</th>
<th>% Change in MSA Miles since 1990</th>
<th>% Change in MSA Miles per Capita since 1990</th>
<th>% Change in CMSA Miles since 1990</th>
<th>% Change in CMSA Miles per Capita since 1990</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwest</td>
<td>9</td>
<td>1,776,062</td>
<td>1,886,133</td>
<td>77</td>
<td>9</td>
<td>77</td>
<td>9</td>
<td>9</td>
<td>12%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>Northeast</td>
<td>18</td>
<td>1,294,393</td>
<td>1,238,365</td>
<td>58</td>
<td>49</td>
<td>58</td>
<td>49</td>
<td>9</td>
<td>2%</td>
<td>21%</td>
<td>2%</td>
</tr>
<tr>
<td>South</td>
<td>8</td>
<td>2,423,365</td>
<td>2,483,365</td>
<td>108</td>
<td>10</td>
<td>108</td>
<td>10</td>
<td>9</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>West</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
in the Midwest. In an era of rapidly declining service productivity, the West region MSAs nearly held their ground.

Figure 2. Percent Change in Passenger Miles per Capita for Medium MSAs (Medium MSAs: 1 million to 5 million people)
These results do not support the hypothesis that transit investments are most effective only in regions of the country characterized by older urban forms. Some MSAs with industrial-era legacies, such as St. Louis, saw their transit service and usage grow substantially during the 1990s. But transit service and usage in many other such MSAs did not grow. More typically, transit service and usage grew in rapidly growing western and mountain sunbelt urban regions. One of these, Portland, is famous for coordinating development with transit, and that policy may be partly responsible for transit growth in the Portland urban region. Most other urban areas with rapidly growing transit usage in the West region do not
have such land-use controls in place, however. What accounts for the superior transit performance of the medium-sized category is not known at this time, but because of the rapid population growth of medium-sized MSAs in this region, the prevalence of urban structure dating from before the 1930s does not appear to be among the causes. In general, rapidly growing medium-sized MSAs in the South did not experience improving transit performance, so growth alone also is not an adequate explanation, but there were mid-sized cases in the South where transit’s performance did improve substantially (Norfolk-Virginia Beach, Orlando, West Palm Beach).

**Transit Performance in Small-Sized MSAs**

There are 32 small-sized MSAs spread throughout the country, as shown in Table 5. The small-sized MSAs supplied significantly less transit service per capita than their medium-sized counterparts, as a comparison of Tables 4 and 5 shows. In 1990, the median small-sized MSA in the four regions supplied between 4 and 6 service miles per capita. Small-sized urban areas in the South and West provided service close to the low end of the scale; the Northeast and Midwest provided service levels toward the top end of the range. Productivity of each service mile, which ranged from 5 to 7 passenger miles per service mile, also was much lower than in the medium-sized MSAs. Consequently, transit usage, which ranged from 23 passenger miles per capita for the South to 39 passenger miles per capita for the Northeast, was lower as well.

Between 1990 and 2000, the population of the small MSAs grew rapidly in two regions (South and West). Unlike in the medium-sized MSAs, transit service in the small MSA category generally did not grow as fast as population (see Figure 4). The exception was the West region, where transit service in the median small-sized MSA increased 6 percent more than the MSA’s 22 percent increase in population. In the Midwest region, by contrast, population change far outpaced service change.
<table>
<thead>
<tr>
<th>Census-Defined Geographic Region</th>
<th>N</th>
<th>Year 2000 MSA or CMSA Population</th>
<th>1990 Vehicle Miles per Capita</th>
<th>1990 Passenger Miles per Capita</th>
<th>1990 Passenger Miles per Vehicle</th>
<th>% Change in MSA Population since 1990</th>
<th>% Change in Vehicle Miles per Capita since 1990</th>
<th>% Change in Passenger Miles per Capita since 1990</th>
<th>% Change in Passenger Miles per Vehicle (Load Factor) 1990-2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Midwest</td>
<td>6</td>
<td>606,475</td>
<td>6</td>
<td>31</td>
<td>5</td>
<td>5%</td>
<td>(22)</td>
<td>(18)</td>
<td>(14)</td>
</tr>
<tr>
<td>Northeast</td>
<td>6</td>
<td>638,212</td>
<td>6</td>
<td>39</td>
<td>7</td>
<td>2%</td>
<td>(7)</td>
<td>(10)</td>
<td>(12)</td>
</tr>
<tr>
<td>South</td>
<td>13</td>
<td>602,894</td>
<td>4</td>
<td>23</td>
<td>7</td>
<td>15%</td>
<td>(2)</td>
<td>(30)</td>
<td>(11)</td>
</tr>
<tr>
<td>West</td>
<td>7</td>
<td>712,738</td>
<td>5</td>
<td>35</td>
<td>7</td>
<td>22%</td>
<td>6</td>
<td>23</td>
<td>13</td>
</tr>
</tbody>
</table>
Passenger miles per capita increased by 23 percent during the decade for small-sized MSAs in the West region, coming on top of a 22 percent population increase (see Figure 5). The median small-sized MSA in each of the other regions posted substantial declines in passenger miles per capita, though there was an exception in Florida (Sarasota). Productivity also improved for small-sized MSAs in the West region, but not in the other regions.
Figure 5. Percent Change in Passenger Miles per Capita for Small MSAs
(Small MSAs: 500,000 to 1 million people)

Results for small-sized MSAs generally refute the hypothesis that transit performs best in cities of old urban form. As in the case of medium-sized MSAs, the small-sized MSAs with the better-performing transit service generally were those in the rapidly growing West. In the South, service and productivity change are not pacing the growth in population. The notable exception is Sarasota, Florida.
Conclusions

Our hypothesis was that growth in transit use per capita and increasing transit productivity is associated with old urban forms. We assumed that urban areas having such characteristics would most likely be located in the Northeast and Midwest regions, where transit use historically has been large. If the hypothesis were true, we would find most urban areas with increasing transit usage per capita and productivity located in those regions.

Our analysis suggests that the truth of the hypothesis is associated with the size of urban regions. A comparison of the two megaurban regions (New York and Los Angeles) certainly suggests the truth of the hypothesis. However, analysis of the 7 urban regions with populations between 5 and 10 million people gives no clear indication of the truth of the hypothesis. For the 40 urban regions between 1 and 5 million people and the 31 urban regions between 500,000 and 1 million people, the hypothesis appears untrue. Of the 71 urban regions in those two categories, those that experienced transit growth during the 1990s were fast-growing regions located primarily in the West region. In the larger of those two categories, transit use per capita and transit productivity in the West region surpassed per capita usage and productivity in all other regions by 1990, and by 2000 the West surged ahead much farther in those performance measures. In the 500,000 to 1 million population category, there was little variation among the regions in 1990. By 2000, however, the West region far outpaced the others in service, ridership, and productivity.

The South, whose MSAs also grew rapidly in population during the 1990s, did not perform as well as the West. However, there was wide variation in experience among South region MSAs, especially in the small MSA group. This variation may be important for further analysis of why transit demand increases in some growing regions but not others. We are continuing research into this question and particularly are interested in understanding whether structural differences in urban regions of the West and South can account for the difference in transit performance. Western MSAs tend to be much denser than southern MSAs, for example. Or, are there differences in the way that transit service is supplied that might explain the differences? At this point, we do not know, but from what we present here we conclude that the prevalence of old urban forms is not a prerequisite for transit growth. This analysis is based on passenger miles as the unit of transit demand; the use of unlinked trips, which we also collected, likely would change some details of the analysis but not the final conclusion.
Acknowledgments

We are indebted to Tara Bartee and Ike Ubake in the Public Transit Office of the Florida Department of Transportation for making the FTIS available to us. We also thank Professor Albert Gan for assistance in using FTIS, and numerous transit and MPO planners for aiding our analysis. All omissions and errors remain our responsibility.

References


**About the Authors**

**Gregory Thompson** (gthompsn@garnet.acns.fsu.edu) teaches courses in transportation planning and plan evaluation and conducts research in transit policy, planning, and history in the Department of Urban and Regional Planning at Florida State University. Prior to obtaining his doctorate in social science from U.C. Irvine in 1987, Prof. Thompson worked in transit planning positions in Edmonton, Berkeley, San Diego, and Sacramento.

**Jeffrey Brown** (jbrown2@fsu.edu) is an assistant professor of urban and regional planning at Florida State University. His research interests include urban transportation history, transportation finance, and public transit policy.
Abstract

Geographic Information Systems (GIS) have provided a platform to present information over the Internet to potential users of public transportation. The advantage of using a GIS is that it allows the user to select an origin and destination on a map, easing the task of inputting information to the itinerary-planning process. In addition, the mapping features of GIS can provide a user-specific map showing the route(s) used in the itinerary, as well as local access, egress, and bus stop information. In this article, the design issues associated with the use of GIS in itinerary generation are discussed. Specific design principles are articulated, based on existing knowledge of requirements for the human-computer interface (HCI). In application of these principles, this article describes the implementation of an ArcIMS GIS-based itinerary planner for the Sun Tran bus network in Tucson, Arizona. This system provides users the option of selecting their origin or destination on the map, manually entering an address, or selecting a landmark from a pull-down menu. The routing algorithm then finds the optimum path, and the output is presented to the user both in text and on the map. This is unique from other itinerary planners because it provides an interactive point-and-click map feature that can be implemented using commercially available GIS software.
Introduction

Transit agencies have always struggled to attract riders in a highly competitive transportation market. Potential riders have a large number of options available to them that would encourage use of other modes of transportation. One of the major problems associated with transit ridership is the presentation of information. Abdel-Aty (2001) performed a survey of the effect of advanced transit information and stated, “About 38% of non-transit users indicated that they might consider transit use if appropriate transit information was available to them” (p. 276).

Historically, transit agencies offered transit information including trip planning through call centers. One of the limitations of this system is that graphical information was difficult to relay through a telephone call. Recently, transit agencies have made their service information available on the Internet, using maps, schedules, and on-line automated trip planners. The Internet provides certain benefits when presenting information but cannot replace call centers for more complex trip-planning requests or for people without Internet access. Hence, rather than replacing call centers, the Internet can serve some passengers’ routine trip-planning needs, relieving call centers of some of this traditional workload. Radin et al. (2002) provides an excellent report on the current state-of-the-practice of on-line trip planners. There have been several approaches to create on-line trip planners. One of the more advanced applications is the introduction of interactive maps using Geographic Information Systems (GIS) software. This software provides functionality that allows a map to be created and accessed that is personalized to the user’s preferences and route choices.

Several researchers have described itinerary-planning systems and have recommended functions and features that should be provided. Trépanier et al. (2002) indicate that the itinerary calculation process includes:

- origin, destination, and circumstance specifications,
- access and egress calculations,
- path calculation, and
- schedule integration.

When developing a trip planner, the developer must consider the user’s needs and provide tools that allow the user to plan his/her trip using the same decision-making factors that he/she would use without the trip planner. Several authors have identified some options that mimic this decision-making process. Huang
and Peng (2001) note that important trip-planning options should include the shortest travel time, the minimum amount of transfers, the minimum fare, and/or the least amount of walk time. Donovan (1998) recommends allowing the user to choose which mode he/she prefers and allow the user to choose his/her maximum walking time.

Trip planners can also provide a large amount of information to developers and planners related to the origins, destinations, and timing of trips, and on the use of the trip planner software. Trépanier et al. (2005) show that people using map-based trip planners employ a mixture of methods when determining their origins and destinations, including clicking on the map, entering a landmark or intersection, or any combination of those for both the origin and destination. The authors also demonstrate that users often plan more than one trip during their visit to the transit website. These are important considerations when developing an on-line trip planner.

**Text-Based Trip Planners**

Many times the input and output of the trip planner is a text-based interface. The user enters his/her address at the origin and destination and time requirements including day of travel in text fields. These time inputs can include the time at the origin or the time at the destination (Huang and Peng 2002). The information is sent to the web server, which sends it to a routing algorithm. The completed itinerary is returned to the web server and back to the browser. This has been the state-of-the-practice of on-line transit itinerary planners since their evolution from call centers (Radin et al. 2002).

Peng and Huang (2000) recognize some problems associated with text-based trip planners. One major problem is that, many times, the user either does not know the exact address of both the origin and the destination or the user does not enter the address correctly. Another problem is that sketch maps often do not provide enough detail or scale to be useful when attempting to plan a trip. Peng and Huang state that a solution to this problem is to allow users to pick their origins and destinations from pull-down menus and provide interactive maps during the itinerary-planning process. This is reinforced by a large literature on human-computer interface design (Brinck et al. 2002; Shneiderman 1998).

**Map-Based Trip Planners**

To incorporate maps into this process, Smith (2000) identifies the two major functions of a map-based trip planner. The first function is that the planner must make
spatial decisions. This function requires that a planner find all the nearby transit stops within walking distance of a desired origin or destination. The trip planner must also be able to make temporal decisions; that is, the planner must be able to link schedule times to the origin and destination as well as to determine total trip time and maximum time allowed for the trip. Smith (2000) explained that the goal of a map-based trip planner is to minimize total trip time subject to spatial, temporal, and system constraints.

Peng and Huang (2000) indicate that on-line, map-based trip planners have a three-tier architecture. The first tier of the architecture contains a user interface on the web browser, which is the client-server tier. The second tier is a web server, which is the server tier. The third tier is an application server that contains a GIS application server and/or a database server, which is the application tier. As an example of this third tier, Karimi et al. (2004) developed an address geocoding and routing application that finds the route with the minimum number of transfers between an origin and destination and displays them on a GIS-based map.

To date, there are very few map-based trip planners in which the user can interact with the map while choosing origins and destinations (Caliper Corporation 2003; STM 2003), particularly those that use commercial off-the-shelf GIS tools. Lee et al. (1999) discuss the challenges of developing a map-based trip planner using commercial off-the-shelf GIS tools. The authors outline data needs and difficulties associated with identifying accurate locations on a map using GIS software.

The objective of this research is to develop some general design principles for a GIS-based itinerary-planning tool. To illustrate these principles, the research also implemented a prototype GIS map-based itinerary-planning website for a transit agency, using readily available software. This paper documents the design principles and subsequent development of the prototype website. The second section of the article provides more detail about the design principles for a map-based trip planner. The process of data acquisition and development for such a trip planner are discussed in the third section. The fourth section presents the development of a prototype map-based itinerary planner using ArcIMS and transit schedule information. Finally, the fifth section offers conclusions and recommendations to improve the functionality and performance of such itinerary-planning tools.
System Design Principles
The goal of a map-based itinerary-planning system is to provide the transit rider with an interactive, easy-to-use, trip-planning application on the Internet. This system must be GIS map-based to provide interactivity. Its tools must be intuitive and easy to use, yet it must include powerful GIS functions to perform the analysis.

User Interface Properties
The user interface serves two different purposes: (1) it provides a common interface by which information is collected from the user as inputs to the itinerary planner; and (2) it provides the interface by which the specific itinerary is communicated back to the user. Both of these purposes require specific considerations on the user interface.

When developing the trip planner, one important element in the design is to determine what input from the GIS is needed for the routing algorithm and then decide how to provide the data to the routing algorithm. Most commonly, the primary inputs provided by the user are the origin and the destination of the trip as well as the circumstances of the trip, implying the day of week, the time at which the trip begins, the time when the trip should end, or both. More specifically, routing algorithms require that the user enter in unique locations, in the form of land parcels, addresses, or landmarks, as origins and destinations. The GIS analysis must then deliver nearby bus stop identifiers to the routing algorithm. Finally, more sophisticated itinerary-planning tools ask the user for specific information such as how far they are willing to walk or drive to get to and from a transit stop, what preferences they may have for transfers, whether they prefer bus or rail modes, and whether they would like to minimize cost (fares) or travel time.

While there has been substantial research on the design of interactive websites, of relevance here is the use of map information in this input process. Specifically, when designing the display of the system, consideration must be made to determine what cartographic information should be presented that is relevant to trip planning in specifying a trip origin and a trip destination. When deciding how users will determine the origin and destination bus stops, the developer must provide the options that the rider normally uses when planning a trip. Many people desiring to use transit either know the address of their origin or destination, or they know the name of a large landmark where they would like to leave from or go to. However, this does not constitute all of the ways that a transit rider would plan a trip. Many riders plan trips by looking at maps and identifying their approximate
location based on street intersections or other orientation landmarks. From that point, they decide which bus route to take.

In attempting to provide this same level of information, many types of GIS-based cartographic information are available, including specific land parcels or locations (e.g., organized by street address), landmark data, and intermodal (transportation network) data. Landmark and intermodal data contain map features that would aid in the orientation of the users, as landmarks and street networks can be major inputs to the cognitive process of navigation. Landmarks show the user points of interest or large trip generators served by the transit network. Landmarks and the local street network can also aid the user in orienting themselves on the map. In addition, through the intermodal data, the user can also determine what modal options are available within or outside of the transit network.

Overall, all features of the map must have some purpose that will aid the rider when determining the trip plan. GIS interfaces have the advantage of displaying different levels of detail and information depending on the scale or zoom level of the map. The primary challenge in designing the interface for user input is to determine how much information to display on the map. More cartographic information (e.g., additional landmarks, a background aerial photograph, a more detailed street network, etc.) may be very helpful to the user to assist in orientation and to find specific locations that serve as a trip origin and/or a trip destination. Conversely, more information can provide greater visual clutter, possibly confusing the user. In addition, more information in the display means that there can also be substantial latency in generating and manipulating the maps through a GIS-based system. As a result, a trade-off must be made in determining the type of information provided in the interface.

Once the system has generated an itinerary, this information must then be presented to the user. Again from a cartographic perspective, the same map features for user input may also be useful for the map output. The display should also provide a clear indication of the recommended itinerary. Specifically, transit route(s) to be used can be accentuated by highlighting specific bus stop locations and paths of access from origin to bus stop and from bus stop to destination.

**Key Functions**

A GIS-based interface provides all of this functionality for identifying origin and destination bus stops: an address search function, a landmark search function, and a function that allows the user to click (or select) any point on the map, allowing
the user to select specific landmarks, specific locations, and/or individual land parcels (specific addresses).

Figure 1 displays the core series of events that occurs when a user runs a map-based itinerary planner, specifically over the Internet. With the GIS functionality, the user can have numerous options to select the trip origin and destination. If the user knows the address or landmark, they can enter that directly as a text field. The GIS system can then conduct a query for that location (feature) in the database. Once the feature is found, it can be highlighted on the map and a buffering tool in the GIS can be used to find all of the bus stops within a given (maximum) walking or driving distance. Those bus stops are then sent to the routing algorithm.

**Figure 1. Flowchart of Events Required for the Trip Planner**
If the user prefers, he/she can utilize the point-and-click capability of the GIS. With this function, the user can simply click a point on the map and the selected location is highlighted (and can even be displayed in text form). With this location, the system then identifies the same buffer around that parcel and the bus stops are sent to the routing algorithm. Essentially, manually inputing an address or landmark performs the same task as clicking a point on the map. All of the query tools access the same database for the location and generate a buffer around that location.

The final critical input required by the routing algorithm is the desired time of departure from the origin, or the desired time of arrival at the destination. This information can be input by using a simple text form or pull-down menu. All of this information is sent to the routing algorithm, and the routing algorithm returns an itinerary to the viewer in the form of text directions and a highlighted route map.

The main purpose of these functions is to provide an interface that can be accessed through a standard web browser, allowing everyone who has Internet access to use the system. It is also designed to provide all of the functionality of the system through a single, easy-to-use interface. As a design principle, all of this functionality can be provided by modifying the tools of existing GIS software to create a custom application. The adaptation of existing GIS tools is important in that many of the existing GIS resources that a transit agency may have can be applied directly to the system without the additional cost of accommodating a propriety system.

Software and Data Requirements
In a GIS-based itinerary system, spatial and temporal decisions are required and likewise, spatial and temporal data are required to make those decisions. The spatial data are usually stored as georeferenced shapefiles, covering the usual sets of shapes (points, lines, polylines, polygons, etc.) with associated database files. The temporal data required are schedule data for the bus stops. With these data, the GIS software, a routing algorithm, and itinerary planner can be developed.

Software
Based on the functions described in the previous section, the primary functions that might be considered internal to the GIS software tools include the ability to:
• generate maps for the World Wide Web (i.e., of reasonable size, electronically);
• pan, zoom, refresh, and otherwise manipulate the map;
• change display features based on the map scale;
• select using point-and-click functionality on the map;
• query for landmarks, addresses, and other points of interest;
• generate buffers for determining passenger access and egress distances to and from bus stops; and
• select and highlight portions of a route (bus route, street network, etc.) for display with the itinerary.

In addition, the ability to generate a shortest path (itinerary) through the network can also be seen as a function that may be included directly in the GIS. Some commercial GIS software have this capability today.

**Spatial Data**

The spatial data used in the GIS can be categorized into three groups: those required for routing functionality, those complimenting an intermodal transportation network, and those landmarks that might aid in the use of the program. The basic routing functionality within the itinerary-planning problem requires bus route information, bus stop information, as well as a parcel layer that provides complete coverage of the transit network. The bus route information is used within the GIS to display available routes to the user when requesting input, and to display the recommended itinerary after it has been generated. The bus stop information is used to locate bus stops in the vicinity of the origin and destination (a set of parcels), typically using some sort of buffering function. Most GIS systems have a simple Euclidean distance buffer tool. More sophisticated buffering tools calculate distance along the street network, which is more appropriate for walking trips, but also more difficult to implement because the physical representation of access (existence of sidewalks, bicycle lanes, physical barriers, turning movements, etc.) requires a significant amount of additional data and computation.

The intermodal data can be valuable to the user in determining potential access and egress routes, as well as providing a general spatial orientation to the map. Data in this category include the street network (at perhaps various levels of detail, such as major or minor streets, local roads, etc.), public parking lots or park-and-ride lots, transit centers, bike routes, walking paths, public trails, and other
such transportation facilities. Similarly, landmark information can include any number of possible forms to orient users to the map and to provide salient locations that serve as origins and destinations: major employment centers, shopping centers, schools, hospitals, parks, post offices, libraries, bodies of water, or other topographic features. The list of potential landmarks can be extensive. However, one is reminded of the potential trade-off between additional information for the user, and the creation of visual clutter, confusion, and longer computer processing times that might also be caused by the higher information display.

**Temporal Data**

The other major component of the trip-planning data is the temporal data, in the form of transit schedule data. Schedule information gives the departure time at each time point (and preferably every bus stop) along each route for each run (bus trip from terminal to terminal). Schedule data must be included for all operating schedules, such as weekday, Saturday, Sunday, and holiday schedules. In some cases, some routes or route segments are not operated on some days (e.g., express bus service is not operated on weekends); in these cases, separate data for routes and bus stops must be maintained, conditioned on the day of the week.

In some instances, schedule data for specific bus stops is not available directly from the transit agency. In these cases, it is necessary to interpolate to determine the estimated bus departure times from these intermediate stops. This in and of itself can be done directly in the GIS, if a linear reference model has been created that gives the locations of bus stops and time points along a route. A less precise method simply uses the preceding time point when determining a passenger’s departure time from the origin, and the subsequent time point is used when determining a passenger’s arrival time at the destination.

**Prototype Design and Implementation**

Under these design principles, a prototype system was developed with commercial off-the-shelf GIS software and existing spatial and temporal data in Tucson, Arizona. This prototype system serves as a simple test of the functionality for Sun Tran, the fixed-route bus transit operator in Tucson.

**Developing a Custom Website**

The software used for this project was ArcIMS (Arc Internet Map Server) distributed by the Environmental Systems Research Institute (ESRI 2003). ArcIMS provides the functionality to present GIS information on the Internet or an intranet,
allowing users who do not have GIS software or who are not familiar with GIS operations to use the system. In addition, ArcIMS has all of the features described earlier in this article in terms of built-in functionality for producing a map-enabled itinerary-planning website.

ArcIMS provides a simple interface with potential for customization that will allow both users and developers to easily use the system. To develop a website using ArcIMS, the developer must go through three stages of site development. The steps within ArcIMS are:

1. ArcIMS Author—develop an initial GIS map.
2. ArcIMS Administrator—create a map service using the map created in Author.
3. ArcIMS Designer—create a web browser interface in which to display the map.

Authoring a Map. The first step in developing a website is to create a map using Author. This step allows the most customization of the map display. Once all of the shapefile data are collected, the desired layers are added to the map, and displayed on the map and in a list. The layers are essentially stacked on top of each other; because of this the developer must take care to order the layers to avoid obscuring an important layer.

This project contains 15 shapefiles, provided and maintained by the City of Tucson and Pima County. The shapefiles, which are ordered according to desired visibility, are:

- parcels,
- bus routes,
- bike routes,
- major streets,
- minor streets,
- bus stops,
- park-and-ride lots,
- public trails,
- schools,
- post offices,
• parks,
• malls and colleges,
• libraries,
• lakes, and
• hospitals.

From this list, one may see that the parcels are the top layer; this was done to enhance user selection of origins and destinations. While the bus routes and stops can be considered critical to understanding the itinerary, the parcel layer was placed at the top to provide user point-and-click functionality for origin and destination locations. ArcIMS’s default setting requires that any spatial queries be conducted on the top layer in the map. The only data required to operate the algorithm are bus stops, bus routes, and parcel coverage layers. All other layers are included at the discretion of the developer and primarily provide spatial information to orient the user and have no application to the algorithm generating the itinerary.

Major transportation intermodal data form the next several layers, covering bus routes and stops, streets, bike routes, park-and-ride lots, and public trails. Curiously, bike routes were near the top of the layers, as Tucson is considered a very bicycle-friendly community and there are a large percentage of bus users who access the bus using a bicycle.

Finally, the lower layers include major landmarks, such as schools, post offices, parks, malls, colleges, libraries, lakes, and hospitals. In this particular layer, many other potential landmarks were omitted to limit the total information provided in the map interface. While a user may not see their desired landmark in this list, the constraints on computational time had a significant effect on the design of the prototype. Further experimentation on the selection of landmarks would be useful, but was not in the scope of this research.

Once these layers are added, the map is made into a map service through Administrator. For this application, an additional overview map was created displaying only the bus routes layer; this was added to simplify presentation of the bus network. In this way, users can have a very simple spatial representation of the bus service provided in the community. This can be seen in Figure 2.
Administering a Service. The next step for the creation of this website is to create map services. These services actually generate the relevant map(s), upon request from a browser for a specific map. Creation of map services is done through the Administrator component of ArcIMS. In the case of this website, two map services must be created: one for the main map and one for the overview map. After these map services are administered, they can be accessed by the web server and displayed on a web browser.

Designing a Website. Developing a browser interface that displays the map services is done through the Designer component of ArcIMS. See Figure 2 for a default display of the main map and the overview map. One of the goals of this project is to make a powerful yet simple interface in which a user, who is assumed to have no GIS experience and does not understand the use of GIS tools, can easily navigate through the system without confusion. Because of this, we have to customize the interface to eliminate or combine some of the advanced GIS tools that are inherent in ArcIMS so that the user only needs to perform the least number of steps.

Customization of the Interface

The GIS interface provided by ArcIMS is very powerful but is also too complicated for the average Internet user. This project therefore requires that the interface be simplified in such a way as to provide all of the functionality required for the trip-planning application, yet not so complicated as to discourage use of the system. Simplification must occur in the interface of the browser as well as in the functionality of several of the tools provided by the ArcIMS software.

Interface Customization. The interface of the software provides some features that could be confusing to a casual user and might discourage use of the system. As shown in Figure 2, the original viewer contains three columns: toolbar frame, map frame, and table of contents frame. The viewer also contains three rows: title frame, map frame, and bottom frame. The text frame is contained within the map frame and is a fixed height. The map fills in the space above the text frame.

To make the website more user friendly, several of the frames were moved to aid in the simplicity of the interface. The intent was to provide basic functionality with the map (visualization, pan and zoom, point-and-click, etc.), and input of other text, but without providing visual clutter or otherwise overstimulating the user. Several of the frames had little meaning to a user with no GIS experience, so these were moved or modified to maximize the size of the map. A color-coded legend was included to ease the use of the map. Tools that are not relevant to the func-
tions of this website were also taken out of the toolbar. Toolbar icons were also changed to text to be more intuitive to a casual user. Finally, a simple HTML frame was included that contained all of the text input fields such as origin address, destination address, and travel time for users of the system who choose not to use the map-based system.

**Tool Customization.** The main function required from the GIS software is to determine the location of the bus stops surrounding the user’s origin or destination on the map. In this case, since the user is looking for addresses, landmarks, and points on the map, all of the analysis tools are used on the parcel layer. To provide the point-and-click functionality that is desired, the Buffer tool was used to find bus stops. Because of its complexity, this tool was modified in such a way as to automatically execute the buffer when the Select by Rectangle tool was used.

Since the only feature that the routing algorithm needs is the bus stops, the bus stops layer is chosen as the highlighted feature for buffering. The user is given the option of choosing an acceptable walking distance, and the bus stops within that Euclidean distance are selected. The display shows the highlighted parcel surrounded by a circle with a radius of whatever walking distance the user has specified. Within this circle, all of the bus stops are highlighted. Other metrics, such as a right-angle metric for a grid street network, could also be used to generate the buffer. Such an approach would approximate the actual walking distance more closely than the Euclidian distance, although it still may not adequately capture the walking times in the actual street network. As mentioned previously, a full representation of the street network in this step generates additional data and computational requirements to the itinerary planner.

A similar technique is used for the user’s option of entering an address or selecting a landmark. Instead of using the Select by Rectangle tool, the user can take advantage of the Query tool. With the modified interface, the format of the query is automatically generated so that the user does not have to develop a query on their own. In the case of entering an address or landmark, the user enters his/her exact address in the field or selects a landmark from a pull-down list. The system then queries the parcel database for a match. At this point, the Buffer tool is immediately run and the surrounding bus stops are selected. Again, those bus stops are used in the routing algorithm as potential origins and destinations.
Routing Interface
Transit routing algorithms are very different than standard street network algorithms. These algorithms are more complex because they generally operate on fixed routes rather than standard street networks. The intricacies of the schedule’s timings, especially in the case of transfers, adds to the complexity. As a result, rather than using existing network tools in the ESRI suite of software, the ArcIMS map-based system was designed to connect with a back-end, separate routing algorithm. This interface sends the selected bus stops to the routing algorithm, and, after the routing is done, presents the results of that algorithm to the user back through ArcIMS. Therefore, the main focus of this application from ArcIMS is to provide the existing routing algorithm with the information that it needs to find the shortest path, and, once the path is found, to present the path to the user.

Routing Algorithm. The itinerary-planning algorithm of Hickman (2002) was used for this application. This is a forward-searching algorithm (from the origin to the destination) and can also be implemented from the destination back to the origin. It finds the shortest path between a given origin-destination pair of bus stops, starting at the origin at a certain time and working toward the destination. An itinerary can be generated for every origin-destination bus stop pair, using any one of the buffered bus stops at the origin and any buffered bus stop at the destination. With each bus stop pair, the algorithm traces a minimum possible time path from node to node along the transit network; these nodes are actually the time points in the Sun Tran schedule. If the system cannot find a path to the destination bus stop on the same route as the origin, it then looks for a transfer node. The transfer nodes are also schedule time points serving several routes. Once the algorithm finds a transfer node, it begins performing a similar search along the new route toward the destination. The algorithm repeats this process until it finds a path to the destination bus stop. All possible routes are considered and the one with the shortest travel time is selected. This algorithm avoids the computationally difficult task of finding all possible origin-destination paths by pruning paths with excessive travel times, as the paths are generated (see Hickman 2002).

This particular routing algorithm gives the user the option of choosing whether to determine the optimum path from the existing schedule or from historic bus arrival time data. Historic bus arrival and departure data are taken from an archive of vehicle location data from an automatic vehicle location (AVL) system. These data are used to find the optimum path based on the total trip time as well as
the probability of the passenger arriving at the destination by a particular time. In this way, reliability of service can be considered. For instance, a bus that has high variance in arrival times might produce an itinerary with a low probability of making scheduled transfers and of arriving at the destination by a certain time. Alternatively, the user may choose to find a shortest path using only the static schedule. The user then can choose the deterministic schedule-based shortest path or the nondeterministic, historic shortest path.

**Data Input.** For the routing algorithm, the data input required is the time point that the rider would like to board, the time point that the rider would like to alight, and the time of day the user would like to leave (or arrive) on a given day (weekday, Saturday, Sunday, or holiday). In our case, the schedule data do not correspond to actual bus stops; they only correspond to time points. For the purposes of the prototype, rather than interpolating the schedules for all bus stops, we used time points in the following way. In the selection of origin and destination, the user clicks the approximate locations on the map, inputs an address in the text field, or chooses a landmark from a pull-down list for both the origin and destination. The database is queried and a parcel corresponding to that click location, address, or landmark is identified; then, the bus stops within the buffer distance of the origin or destination parcel are found. Many of the bus stops found are not actually time points for the routing algorithm. In the case of the origin bus stop, the nearest upstream time point is found. In the case of the destination bus stop, the nearest downstream time point is found. The program then runs the routing algorithm. By using these time points, the algorithm delivers a conservative itinerary. However, since the time taken for a bus to traverse between time points is often rather small, this overestimation should not make a large impact on the overall estimated travel time.

**Output**

The output of the routing algorithm must be presented to the user in a manner that is clear and easy to understand. This is done through text directions and a map display. Once the routing algorithm has determined an optimum path between time points, it sends the results back in the form of text directions to an HTML frame that lists the arrival time at the originating bus stop, and the probabilities of the bus arriving at that time if the user chooses the historic arrival times. Specifically, the user can be given the probability of arriving at the destination by a specific time; they can also see, historically, what the earliest and latest times of arrival at the destination have been for that specific itinerary. The HTML frame
also lists the transfer nodes that are required and the bus stop where the rider will alight. The name of the intersection where the bus stop exists is also sent to the display and presented to the user. This is done for the origin and destination bus stops as well as the transfer points in the itinerary. Because of the additional data requirements to generate local walk directions, our prototype does not include text walk directions from the origin and destination to the corresponding bus stops; however, the user can use the map to navigate to or from their destination or origin. The time that the bus arrives at each of those stops is also displayed. Presenting this information gives the user complete and specific directions.

Map Display. The map display is one of the unique features of this application; it includes a display of a personalized map of the highlighted transit routes that will be taken. This highlighting process is done using the existing functionality provided by ArcIMS. The Query tool automatically highlights all of the features that have been queried; in the prototype, the query is run on the bus routes and they are highlighted on the map. Figure 3 shows the final output screen, displaying the map and the text directions to the user. The user can then use GIS tools to navigate the map.

Another improvement that could be included is a more detailed map output. Currently, the map simply has the routes highlighted that will be used. A more desirable output would be to highlight the route from the actual origin to the actual destination, graphically displaying relevant bus stops and transfer nodes as well as relevant times along the route.

Overall, this application performs the appropriate function of finding those bus stops within a predetermined distance from a parcel and converting those unique stop identification numbers into a format that could be used by the routing algorithm. The routing algorithm then finds a set of optimum routes based on the total travel time, and, if desired, the reliability of the travel time. The route is delivered to the user in text format and the routes are highlighted on the map. This map gives the user the same functionality as the original interactive map; the only difference is that the routes are highlighted, giving the user the option to identify landmarks near the origin, destination, or along the route.
Figure 3. Final Output Including Text Directions and Highlighted Map
Conclusions and Recommendations
One of the major obstacles associated with transit is the complexity of trip planning. The goal of this research was to explore different design issues associated with map-based itinerary-planning tools. This will become more important as such map features are added to transit websites. In this way, a prototype was demonstrated to use and modify existing GIS tools and software to create an on-line map-based trip planner for Sun Tran, the transit agency of Tucson, Arizona. This trip planner contains much of the functionality that is currently available in state-of-the-practice websites, but it also contains an interactive map. The interactive map provides the functionality to point and click on a location for the origin and/or the destination, thus eliminating the need for the rider to know exact location information.

System Development
One of the main objectives of this research was to explore the design capabilities and tools that are provided within the readily available GIS software. This would aid future developers of similar projects by using a customized application of existing and widely available GIS software tools. The article has specifically documented system design considerations for a map-based itinerary planner in terms of: (1) interface requirements to allow users to interact with system maps, and associated trade-offs in the level of detail; (2) overall system function and information flows; (3) requisite GIS software functionality; and (4) spatial and temporal data requirements.

ArcIMS 4.0 was used for the prototype. This application was developed beginning with the standard steps through which to create a GIS website in ArcIMS, but the interface had to be customized to make it more intuitive to the user. Some of the tools that allow the user to change settings were removed and many of the tools were automated. For trip planning, a new input frame was added to allow the user to select origin, destination, and desired departure time and send those data to the routing algorithm. The routing algorithm finds the optimum path and sends the output to the browser in the form of text-based directions and a highlighted route map.

Potential Improvements
This project was developed as a prototype of a trip planner that could be implemented by transit agencies. There are several issues related to performance and user features. The main problem with the existing system is that the processing
time to perform one trip calculation is rather slow. Most of the processing occurs on the server, so increasing the processor speed and efficiency of the server would greatly reduce the processing times of each of the steps.

Related to system design, much of the processing speed issues revolve around the very large parcel file that must be visible for most of the functions to work, particularly the point-and-click function. This results in the slow generation of thousands of polylines every time the map display is redrawn. Also, the processing time increases as the zoom level increases for many of the tools. Finally, because of the large number of parcels, much time is spent querying the parcel database for an address or landmark. Directly buffering the bus stops around a point on the map without using the parcel layer would likely improve the processing time.

Some improvements could be made to enhance the user features in any web-based itinerary system. One improvement would be to provide an alternative text-based trip-planning site that is accessible to those with visual impairments who utilize screen-reading software. This could be an alteration of the existing site, prompting the user to use the text-based fields only, or it could be a completely alternate website that executes the routing algorithm with only a text-based interface. As the need for accommodating these patrons continues to grow, provision of these text-only alternate sites may also become more critical.

The system design and prototype described here has some limitations, mostly with processing times and with the detailed features provided in the map display. Nonetheless, it is a successful prototype and provides exceptional map-based functionality that is rarely found with other on-line trip planners. This trip planner provides current state-of-the-practice functionality with its text-based search capability, but it also includes a point-and-click functionality that is very rare with other trip planners.

**Acknowledgments**

The authors would like to express their appreciation to several people for their contribution to this project: Ron Platt from the City of Tucson for supplying updated map information; Steve Whitby from Pima County for providing much of the spatial data used in this project; and Mike Beel and Aimee Ramsey from Sun Tran transit agency for supplying the appropriate schedule data.
This material is based upon work supported by the National Science Foundation under Grant No. 9984906. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References


### About the Authors

**Christopher R. Cherry** (cherry@berkeley.edu) is a Ph.D. student and graduate student researcher in the Department of Civil and Environmental Engineering at the University of California, Berkeley. He holds a master’s degree in civil engineering from the University of Arizona. His research interests include public transit operations and planning, transit security, transportation policy, and transportation needs of developing countries.

**Mark Hickman** (mhickman@engr.arizona.edu) is an associate professor in transportation engineering in the Department of Civil Engineering and Engineering Mechanics at the University of Arizona. At the university, he has taught courses and performed research in transportation and traffic engineering, traffic flow and capacity analysis, and public transit planning and operations. His areas of research interest and expertise include public transit planning and operations, the
application of new information technologies in transportation, and transportation systems analysis and quantitative modeling.

**Anirudh Garg** (*anirudh@coreweb.com*) is a senior software engineer at Core-Web Inc. in Woburn, Massachusetts. He holds a master’s degree in computer science from the University of Arizona. His interests include GIS, web-based mapping software, distributed computing, and web programming.
Valuing Urban Bus Attributes:
An Experience in Kolkata

C. V. Phanikumar and Bhargab Maitra
Indian Institute of Technology, Kharagpur

Abstract

The article presents the marginal willingness-to-pay (WTP) estimates for various qualitative and quantitative attributes of travel with reference to the bus transportation system in Kolkata City, India. A stated choice experiment is designed to capture the responses for estimating marginal WTP values for various attributes. WTP values are estimated separately for commuting and noncommuting trips. The effects of model specification and socioeconomic parameters on WTP values are also studied. Estimates from standard multinomial logit (MNL) and different random parameter logit (RPL) models indicate that WTP values vary with model specification. In the process of developing RPL models, successful application of sparsely used constrained triangular distribution is also demonstrated.

Introduction

Travel needs in developing countries are largely served by public transportation systems, especially bus transportation systems. In the recent years, bus fares have risen at regular intervals due to frequent increases in the price of petroleum fuels all over the world. However, poor quality of travel in bus transportation systems continues, with a resulting declining trend in bus patronage. Policy-makers and practitioners are constantly in search of solutions for improving bus patronage, especially in urban areas of developing countries.
It is essential to improve the bus patronage in urban areas, not only for minimizing the usage of private vehicles and resulting road congestion, but also to safeguard the environment for greater community benefit. Any improvement is expected to bring some benefits to the users, and estimation of user benefits is an essential input for evaluating any improvement plan. The literature shows evidence of estimation of user benefits in the form of willingness-to-pay (WTP) values in the context of transport and nontransport improvements (Hensher 1994; Jose Holguin-Veras 2002; Hensher and Greene 2001; Hensher 2001; Hensher and Sullivan 2003; Adamowicz, Louviere, and Williams 1994; Carlsson, Frykblom and Liljenstolpe 2003; Onyango, Govindasamy, and Nayga 2004). However, most of the studies are carried out in developed countries with limited information available about WTP values in developing countries like India.

In the present study, an attempt is made to estimate WTP values associated with various qualitative and quantitative attributes of bus transportation system with reference to Kolkata City in India. The travel demand in Kolkata City is largely served by more than 5,000 buses. Longer travel times, poor levels of comfort inside buses (based on crowding), poor appearance of buses (both internal and external), and high noise levels are common features of the bus transportation system in Kolkata. Therefore, all these attributes along with travel cost are considered for the estimation of WTP values.

WTP estimates vary depending on the approach and/or model specification adopted. Users’ willingness to pay also depends on the trip purpose. In addition, WTP values may also be influenced by one or more sociodemographic parameters such as age, income, household size, etc. The objective of this research is to estimate users’ WTP values associated with various attributes of bus transport, and study the variation with different model specifications and trip purposes. Trips are classified as commuting (work and business) and noncommuting (recreation and social), and separate models are developed for estimating WTP values. The effect of socioeconomic characteristics on the mean of random parameter (“mean heterogeneity”) is also investigated.

**Methodology**

**Approach**

To estimate WTP values, it is necessary to develop utility models on the basis of user preferences collected in the form of either Revealed Preference (RP) or
Stated Preference (SP) data. Both RP and SP data have been used in diverse fields for estimating WTP values (Adamowicz, Louviere, and Williams 1994; Bates 1982; Kroes and Sheldon 1988; Louviere 1988; Hensher 1994; Jose Holguin-Veras 2002). At times, RP data may be inappropriate as they cannot accommodate nonexisting attributes or variability of attributes, which in turn does not permit the establishment of their influences. On the other hand, SP data facilitate inclusion of hypothetical attributes and variability of attributes. Due to the overall poor service quality of buses in Kolkata, the RP data does not include the variability of attributes. Although it is not advised to use stand-alone SP models for predictions, they are rich and effective in estimating marginal WTP values (Hensher and Sullivan 2003). Therefore, SP data is used for the development of the utility model to estimate WTP values.

SP data may be collected in the form of rating, ranking, and choice. The Stated Choice (SC) method has strong theoretical foundations based on economic theory and is an established approach for understanding and predicting consumer trade-offs and choices in marketing research. SC experiments provide a framework where one can study the relative marginal disutility of variations in attributes and their potential correlations (Louviere, Hensher, and Swait 2000). SC methods are extensively used to model the behavior of individuals (Hensher and Greene 2001; Hensher 2001; Hensher and Sullivan 2003; Carlsson, Frykblom and Liljenstolpe 2003; Onyango, Govindasamy, and Nayga 2004). In the present study, the SC method, where profiles generated using various attributes and their levels are presented to the respondent in the form of choice set, is adopted for observing preferences. Responses in the form of “choice” among the presented choice alternatives are utilized to develop utility models and estimate the WTP values.

Generally, SP and/or RP data are analyzed using Multinomial Logit (MNL) models due to simplicity in estimation. However, MNL models impose restrictions such as independence of irrelevant alternatives (IIA). Modifications to the MNL models to reduce the influence of restrictions lead to Random Parameter Logit (RPL) or Mixed Logit (ML). In this article, travelers’ marginal WTP values are estimated using SC data and RPL models with constrained triangular distributions over random parameters. RPL models are developed considering uncorrelated choice sets and correlated choice sets across each individual traveler.

**Econometric Model**

MNL models are essentially econometric models developed on the basis of Random Utility Theory (Thurstone 1927; McFadden 1974), where the utility of each
element has an observed (deterministic) component denoted by \( V \) and a random (disturbance) component denoted by \( \varepsilon \):

\[
U = V + \varepsilon
\]

If the deterministic part \( V \) is again a function of the observed attributes \( (z) \) of the choice as faced by the individual, the observed socioeconomic attributes of the individual \( (S) \) and a vector of parameters \( (\beta) \), then

\[
V = V (z, S, \beta)
\]

A probabilistic statement can be made (due to presence of the random component) as, when an individual “\( n \)” is facing a choice set, \( C_n \), consisting of \( J_n \) choices, the choice probability of alternative i is equal to the probability that the utility of alternative “i,” \( U_{i_n} \), is greater than or equal to the utilities of all other alternatives in the choice set. For example,

\[
P_n (i) = \Pr (U_{i_n} \geq U_{j_n}, \text{for all } j \in C_n)
\]

\[
P_n (i) = \Pr (V_{i_n} + \varepsilon_{i_n} \geq V_{j_n} + \varepsilon_{j_n}, \text{for all } j \in C_n, j \neq i)
\]

Assuming IID (Gumbel distribution) for \( \varepsilon \), the probability that an individual chooses i can be given by the MNL model (McFadden 1974; Ben-Akiva and Lerman 1985)

\[
P_n = \frac{e^{V_{i_n}}}{\sum_{j \in j_n} e^{V_{j_n}}}
\]

(1)

This model can be estimated by Maximum Likelihood techniques, and is a useful first cut at modeling choice behavior. However, several well-known limitations apply. The most severe is the IIA property, which states that a change in the attributes of one alternative changes the probabilities of the other alternatives in proportion. This substitution pattern may not be realistic in all settings. Secondly, the coefficients of all attributes are assumed to be the same for all respondents in a choice experiment, whereas in reality there may be substantial variability in how people respond to attributes. To overcome these limitations, a generalized form of MNL (i.e., a random parameters logit model) is used to account for unobserved heterogeneity. Let us assume the utility function of alternative i for individual n is

\[
U_{i_n} = \beta x_{i_n} + \varepsilon_{i_n} = \beta' x_{i_n} + \hat{\beta}_n x_{i_n} + \varepsilon_{i_n}
\]

(2)
Thus, each individual’s coefficient vector $\beta$ is the sum of the population mean $\beta^I$ and individual deviation $\beta_n \cdot x_n^i$, are error components that induce heteroskedasticity and correlation over alternatives in the unobserved portion of the utility. This means that an important implication of the ML specification is that we do not have to assume that the IIA property holds. Let tastes, $\beta$, vary in the population with a distribution with density $f(\beta | \theta)$, where $\theta$ is a vector of the true parameters of the taste distribution. If the error terms ($\varepsilon_{in}$) are IID type I extreme value, it is a random parameter logit model (Train 1998). The conditional probability of observing a sequence of choices is the product of the conditional probabilities

$$S_n(\beta_n) = \prod P(k(n,t) | \beta_n)$$

(3)

where:

$k(n,t)$ denotes the sequence of choices from choice sets that person $n$ chooses in situation $t$

In the choice experiment, the sequence of choices is the number of hypothetical choices each respondent makes in the survey. The unconditional probability for a sequence of choices for individual $n$ is then expressed as the integral of the conditional probability in (3) over all values of $\beta$:

$$P_n(\theta) = \int S_n(\beta) f(\beta | \theta) d\beta$$

(4)

In general, the integral cannot be evaluated analytically, and one has to rely on a simulation method for the probabilities. Here a simulated maximum likelihood estimator, using Halton draws, to estimate the models is used (Train 1999). This type of random parameter model is less restrictive than standard conditional logit models. However, these less restrictive models should be applied cautiously. Apart from being more difficult to estimate, the literature shows that the results can be rather sensitive to the distributional assumptions and the number of draws applied in the simulation (Hensher and Greene 2001). Therefore, the gain in terms of precision of the estimates of WTP is unclear.
**Distributions**

For the development of RPL models, it is necessary to assume suitable distributions for random parameters. Different distributions for random parameters like normal, lognormal, uniform, and triangular have been attempted by researchers while developing RPL models (Algers et al. 1998; Hensher 2001; Revelt and Train 1997; Hensher and Greene 2001; Train 2001).

The lognormal distribution is suitable if the mean of random parameter needs to be of specific (nonnegative) sign. However, a long upper tail of lognormal distribution results in extremely high WTP values. A uniform distribution with a (0, 1) bound is suitable for dummy variables. The triangular distribution, where the density function looks like a tent with a peak in the center and dropping off linearly on both sides of the center, is advantageous over normal or lognormal distributions due to its bounded nature. However, like normal or uniform distribution, triangular distribution also has the disadvantage of producing the wrong sign to some shares due to spread or standard deviation. It is possible to overcome the disadvantage of triangular distribution by imposing a constraint on the spread. In constrained triangular distribution, mean and spread are made equal to minimize the effect of spread on WTP estimates, yet producing WTP estimates with proper signs (Hensher and Greene, 2001). The advantages of constrained triangular distribution over other distributions are:

1. The bounded nature of the triangular distribution helps in early convergence of the model.
2. It keeps the sign of the estimate the same for all respondents (i.e., there is no reversal of sign throughout the respondents) unlike normal or triangular distributions.
3. It provides simplicity in WTP estimations.

When mean=spread, the impact of the spread is negligible and the ratio of mean to any random parameter (with constrained triangular distribution) over mean of cost will give the WTP value directly. It is not so with normal or triangular distributions, where standard deviation/spread is also to be considered while calculating the WTP value of any random parameter. When mean and spread are made equal, the constrained distribution has a peak in the density function with two endpoints of the distribution fixed at 0 and 2*mean, so that there is no free variance (scaling) parameter. Although constrained triangular distribution has several advantages over the other distributions, its application in WTP estimate has not
been explored adequately. In the present work, the application of constrained triangular distribution is explored while developing RPL models.

**Correlations Among Choice Sets**
In stated choice studies, respondents are often asked a series of hypothetical choice questions. For each experiment, a set of alternatives with different attributes/levels is described, and the respondent is asked to state which alternative he/she would choose. A series of such questions is asked, with the attributes of the products varying so as to determine how the respondent’s choice changes when the attributes change. In addition, this process also allows researchers to make sure that each respondent gets an opportunity to evaluate several SP alternatives. This process improves the richness of the data but may lead to correlated responses across observations, which is in violation of independence of observations assumption in classical choice model estimation (Hensher and Greene 2001). RPL models are developed in the present work taking into account the correlations among responses across each individual.

**Survey Instrument**
Survey instruments are designed for collecting respondent’s trip characteristics, socioeconomic characteristics, and stated preference “choice” from the choice set. Six attributes are considered for the design of choice sets. During the preliminary investigation it is observed that the journey speed for buses is considerably low, comfort is less, appearance of buses is poor, and noise level is high. Therefore, the primary attributes of travel speed and travel cost and the secondary attributes of discomfort, waiting time, appearance of bus, and noise level are considered for the preparation of choice sets. Each attribute is further described by suitable levels. Levels are decided following discussions with experts and trip-makers. The attributes and corresponding levels as used in the study are:

- travel speed (km/h): 20, 15, 12.5, 10
- waiting time (minutes): 4, 8, 12, 16
- travel (dis)comfort: comfortable seating, congested seating, get seat during journey
- comfortable standing, standing in crowd
- noise level: very low, low, high, very high
• appearance: good, average, poor
• Travel cost (paise/km): 50, 75, 100, 125

The full factorial technique with above-mentioned attributes and levels would produce $4^4 \times 3 \times 5 (3840)$ combinations/alternatives/profiles. However, it is practically not possible to include all these profiles in the SP experiment. Fractional factorial orthogonal main effects design (by SPSS 7.5) produced 32 alternatives by eliminating the dominating and dominated alternatives. These alternatives are randomly grouped into eight blocks, each containing four SC alternatives and each respondent is asked to choose an alternative from four blocks. Two sets of questionnaires are prepared each having four SP blocks. While presenting the alternatives, travel time and the corresponding travel speed both are presented to respondents for their convenience. For collection of SP responses, enumerators were trained in multiple sessions to improve the quality of the work as these are personal, paper- and-pencil interviews. A sample of the SP choice set is presented in Figure 1.

**Figure 1. Sample Choice Set**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Alternative 1</th>
<th>Alternative 2</th>
<th>Alternative 3</th>
<th>Alternative 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Travel (dis)comfort</td>
<td>Standing in crowd</td>
<td>Get seat during journey</td>
<td>Congested seating</td>
<td>Standing comfortably</td>
</tr>
<tr>
<td>Travel speed (kph)</td>
<td>12.5</td>
<td>20</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>Travel time (min/km)</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Waiting time (min)</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>External appearance</td>
<td>Average</td>
<td>Poor</td>
<td>Average</td>
<td>Good</td>
</tr>
<tr>
<td>Noise level</td>
<td>Low</td>
<td>High</td>
<td>Very low</td>
<td>Very high</td>
</tr>
<tr>
<td>Travel cost (paise/km)</td>
<td>50</td>
<td>125</td>
<td>100</td>
<td>75</td>
</tr>
<tr>
<td>Choice</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Database
Data were collected from Kolkata bus users in October 2004. The study sample was intercepted while they were at shopping centers, recreational places, and at offices spread over the city. Respondents’ recent trip characteristics, socioeconomic characteristics, and responses to the choice sets were collected. Every respondent was asked to make a choice among four alternatives presented in each choice set. This was repeated four times so that each respondent gets an opportunity to evaluate 16 SP alternatives. During the study, 1,700 respondents were approached and 1,200 (73.5%) gave their consent. Of the 1,200 respondents, only 1,021 respondents’ data were found useful for the development of the models. The remaining data were eliminated due to nonresponses to various items. Information collected included respondent’s socioeconomic characteristics (age, occupation, personal income, family members, household income) along with trip characteristics (trip length, cost of recent trip, trip purpose, etc.), and SP choice. Summary statistics of the information about trip purposes and socioeconomic details like gender, age, household income, and car ownership forming the database are given in Table 1.

Model Development
A total of 3,261 observations were used for the development of models. Models for commuting trips were developed using 853 observations, while 2,048 observations are used for developing models for noncommuting trips. LIMDEP 8.0 (2005) was used for the analysis of SC data using MNL and RPL models. Three RPL models were attempted for each trip purpose—one (RPL 1) with independent choice sets, another (RPL 2a) assuming correlations among choice sets across each individual, and the final one (RPL 2b) considering heterogeneity (i.e., effect of sociodemographic parameters such as age, gender, income, etc.) around the mean of random parameter. In the process of developing models for commuting trips, heterogeneity around mean could not be observed. In RPL models, all the parameters except travel cost are considered random parameters. Travel cost was considered a fixed parameter because (1) it simplifies the estimation of marginal WTP for other parameters (i.e., simple division of coefficient of attribute by coefficient of cost); (2) the distribution of the marginal WTP for an attribute becomes the distribution of that attribute’s coefficient; and (3) it ensures the price variable to be nonpositive for all individuals.
Table 1. Summary Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total sample</td>
<td>1,021</td>
</tr>
<tr>
<td>Net observations</td>
<td>3,261</td>
</tr>
<tr>
<td>Gender (male)</td>
<td>2,964 (91%)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
</tr>
<tr>
<td>&lt;15</td>
<td>8 (0.2%)</td>
</tr>
<tr>
<td>15–25</td>
<td>526 (16.4%)</td>
</tr>
<tr>
<td>26–40</td>
<td>2,008 (61.6%)</td>
</tr>
<tr>
<td>40–60</td>
<td>635 (19.4%)</td>
</tr>
<tr>
<td>&gt;60</td>
<td>84 (2.4%)</td>
</tr>
<tr>
<td>Household income (INR)</td>
<td></td>
</tr>
<tr>
<td>≤ 5000</td>
<td>286 (8.7%)</td>
</tr>
<tr>
<td>5,001–10,000</td>
<td>819 (25.1%)</td>
</tr>
<tr>
<td>10,001–20,000</td>
<td>1,012 (31.1%)</td>
</tr>
<tr>
<td>20,001–30,000</td>
<td>996 (30.6%)</td>
</tr>
<tr>
<td>&gt;30,000</td>
<td>148 (4.5%)</td>
</tr>
<tr>
<td>Trip purpose</td>
<td></td>
</tr>
<tr>
<td>Work</td>
<td>1,759 (54.0%)</td>
</tr>
<tr>
<td>Business</td>
<td>649 (20.0%)</td>
</tr>
<tr>
<td>Recreation</td>
<td>853 (26.0%)</td>
</tr>
<tr>
<td>Car ownership</td>
<td></td>
</tr>
<tr>
<td>0 car</td>
<td>2,855 (87.5%)</td>
</tr>
<tr>
<td>1 car</td>
<td>386 (11.8%)</td>
</tr>
<tr>
<td>2 cars</td>
<td>20 (0.6%)</td>
</tr>
</tbody>
</table>

In the process of developing models, the attributes were coded according to their levels. Quantitative attributes for travel time, waiting time, and travel cost were entered in cardinal linear form (i.e., continuous scale), while qualitative attribute levels were effects coded (-1, 0, 1). In RPL models, all random parameters are assumed to follow constrained triangular distribution. RPL models are estimated with simulated maximum likelihood using intelligent Halton draws with 200 replications (Train 1999). Initially all attributes and effects-coded variables are considered for model estimation. However, in the initial runs one level (i.e., “congested seating”) of the attribute travel comfort and one level (i.e., “average appearance”) of the attribute appearance were found to be insignificant. Therefore, models were reestimated excluding the insignificant levels of attributes (Hensher, Rose
Model estimates for commuting and noncommuting trips are presented in Tables 2 and Table 3, respectively.

### Table 2. Coefficient Estimates from MNL and RPL Models for Commuting Trips

<table>
<thead>
<tr>
<th>Attribute</th>
<th>MNL Coefficient</th>
<th>RPL (Constrained Triangular)</th>
<th>RPL 1 Coefficient</th>
<th>RPL 2a Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-vehicle travel time</td>
<td>-0.243(7.38)</td>
<td>-0.263(7.52)</td>
<td>-0.277(5.59)</td>
<td></td>
</tr>
<tr>
<td>Waiting time</td>
<td>-0.101(15.23)</td>
<td>-0.107(14.53)</td>
<td>-0.110(14.56)</td>
<td></td>
</tr>
<tr>
<td>Comfortable seating</td>
<td>0.234(4.22)</td>
<td>0.256(4.48)</td>
<td>0.247(4.27)</td>
<td></td>
</tr>
<tr>
<td>Get seat enroute</td>
<td>0.175(2.87)</td>
<td>0.201(3.19)</td>
<td>0.203(3.34)</td>
<td></td>
</tr>
<tr>
<td>Comfortable standing</td>
<td>-0.126(1.85)</td>
<td>-0.151(2.16)</td>
<td>-0.144(2.08)</td>
<td></td>
</tr>
<tr>
<td>Very low noise</td>
<td>0.428(7.77)</td>
<td>0.424(7.61)</td>
<td>0.417(7.45)</td>
<td></td>
</tr>
<tr>
<td>Low noise</td>
<td>0.378(5.44)</td>
<td>0.355(5.13)</td>
<td>0.348(5.14)</td>
<td></td>
</tr>
<tr>
<td>High noise</td>
<td>-0.364(4.51)</td>
<td>-0.318(4.06)</td>
<td>-0.307(4.06)</td>
<td></td>
</tr>
<tr>
<td>Good appearance</td>
<td>0.148(4.12)</td>
<td>0.168(4.73)</td>
<td>0.166(4.75)</td>
<td></td>
</tr>
</tbody>
</table>

**Nonrandom parameters**
- Travel cost: -0.033(23.47) 0.034(23.32) -0.034(23.34)

**Random parameter spread**
- In-vehicle travel time: - 0.263(7.52) 0.277(5.59)
- Waiting time: - 0.107(14.53) 0.110(14.56)
- Comfortable seating: - 0.256(4.48) 0.247(4.27)
- Get seat enroute: - 0.201(3.19) 0.203(3.34)
- Comfortable standing: - 0.151(2.16) 0.144(2.08)
- Very low noise: - 0.424(7.61) 0.417(7.45)
- Low noise: - 0.355(5.13) 0.348(5.14)
- High noise: - 0.318(4.06) 0.307(4.06)
- Good appearance: - 0.168(4.73) 0.166(4.75)

**Heterogeneity in mean**
- In-vehicle travel time:
  - Household income:
    - Observations: 2408 2408 2408
    - Log likelihood: -2450.82 -2452.25 -2447.59
    - $R^2$: 0.232 0.231 0.233
Table 3. Coefficient Estimates from MNL and RPL Models for Noncommuting Trips

<table>
<thead>
<tr>
<th>Attribute</th>
<th>MNL</th>
<th>RPL (Constrained Triangular)</th>
<th>RPL 1</th>
<th>RPL 2a</th>
<th>RPL2b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>Coefficient</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random parameter means</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In-vehicle travel time</td>
<td>-0.256(4.69)</td>
<td>-0.277(4.76)</td>
<td>-0.288(4.78)</td>
<td>-0.437(4.27)</td>
<td></td>
</tr>
<tr>
<td>Waiting time</td>
<td>-0.096(8.60)</td>
<td>-0.102(8.26)</td>
<td>-0.105(8.29)</td>
<td>-0.101(8.13)</td>
<td></td>
</tr>
<tr>
<td>Comfortable seating</td>
<td>0.268(2.89)</td>
<td>0.289(3.02)</td>
<td>0.282(2.94)</td>
<td>0.290(3.01)</td>
<td></td>
</tr>
<tr>
<td>Get seat en-route</td>
<td>0.183(1.78)</td>
<td>0.210(1.98)</td>
<td>0.210(2.06)</td>
<td>0.208(1.95)</td>
<td></td>
</tr>
<tr>
<td>Comfortable standing</td>
<td>-0.199(1.68)</td>
<td>-0.228(1.86)</td>
<td>-0.224(1.83)</td>
<td>-0.223(1.82)</td>
<td></td>
</tr>
<tr>
<td>Very low noise</td>
<td>0.468(5.07)</td>
<td>0.465(4.97)</td>
<td>0.455(4.88)</td>
<td>0.462(4.96)</td>
<td></td>
</tr>
<tr>
<td>Low noise</td>
<td>0.369(3.15)</td>
<td>0.352(3.02)</td>
<td>0.342(3.01)</td>
<td>0.350(2.99)</td>
<td></td>
</tr>
<tr>
<td>High noise</td>
<td>-0.361(2.66)</td>
<td>-0.320(2.41)</td>
<td>-0.303(2.42)</td>
<td>-0.322(2.43)</td>
<td></td>
</tr>
<tr>
<td>Good appearance</td>
<td>0.163(2.67)</td>
<td>0.180(2.99)</td>
<td>0.177(3.05)</td>
<td>0.184(3.04)</td>
<td></td>
</tr>
<tr>
<td>Nonrandom parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Travel cost</td>
<td>-0.032 (13.94)</td>
<td>-0.033(13.85)</td>
<td>-0.034(13.93)</td>
<td>-0.033 (13.81)</td>
<td></td>
</tr>
<tr>
<td>Random parameter spread</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In-vehicle travel time</td>
<td>-</td>
<td>0.277(4.76)</td>
<td>0.288(4.78)</td>
<td>0.437(4.27)</td>
<td></td>
</tr>
<tr>
<td>Waiting time</td>
<td>-</td>
<td>0.102(8.26)</td>
<td>0.105(8.29)</td>
<td>0.101(8.13)</td>
<td></td>
</tr>
<tr>
<td>Comfortable seating</td>
<td>-</td>
<td>0.289(3.02)</td>
<td>0.282(2.94)</td>
<td>0.290(3.01)</td>
<td></td>
</tr>
<tr>
<td>Get seat enroute</td>
<td>-</td>
<td>0.210(1.98)</td>
<td>0.210(2.06)</td>
<td>0.208(1.95)</td>
<td></td>
</tr>
<tr>
<td>Comfortable standing</td>
<td>-</td>
<td>0.228(1.86)</td>
<td>0.224(1.83)</td>
<td>0.223(1.82)</td>
<td></td>
</tr>
<tr>
<td>Very low noise</td>
<td>-</td>
<td>0.465(4.97)</td>
<td>0.455(4.88)</td>
<td>0.462(4.96)</td>
<td></td>
</tr>
<tr>
<td>Low noise</td>
<td>-</td>
<td>0.352(3.02)</td>
<td>0.342(3.01)</td>
<td>0.350(2.99)</td>
<td></td>
</tr>
<tr>
<td>High noise</td>
<td>-</td>
<td>0.320(2.41)</td>
<td>0.303(2.42)</td>
<td>0.322(2.43)</td>
<td></td>
</tr>
<tr>
<td>Good appearance</td>
<td>-</td>
<td>0.180(2.99)</td>
<td>0.177(3.05)</td>
<td>0.184(3.04)</td>
<td></td>
</tr>
<tr>
<td>Heterogeneity in mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In-vehicle travel time:</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.212 (2.00)</td>
<td></td>
</tr>
<tr>
<td>Household income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>853</td>
<td>853</td>
<td>853</td>
<td>853</td>
<td></td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-866.94</td>
<td>-867.35</td>
<td>-866.15</td>
<td>-865.11</td>
<td></td>
</tr>
<tr>
<td>$\rho^2$</td>
<td>0.232</td>
<td>0.232</td>
<td>0.233</td>
<td>0.234</td>
<td></td>
</tr>
</tbody>
</table>

Results and Discussions

In Tables 2 and Table 3, the signs of the parameter estimates are as expected and in agreement with the actual condition of the study route. It is evident from the t-ratios that the parameter estimates are statistically significantly different from 0. The overall goodness of fit is considered using Pseudo $R^2$ ($\rho^2$). Value of the $\rho^2$
between 0.2 and 0.4 indicates acceptable model fit (Louviere, Hensher, and Swait 2000). The $\rho^2$ values indicate that these models are good in fit.

The $\rho^2$ value is also improved from MNL to RPL with uncorrelated choice sets models and to RPL with correlated choice sets models indicating superior model fit. Parameter estimates from Tables 2 and 3 clearly indicate that in addition to in-vehicle travel time and waiting time, the discomfort level, appearance, and noise level also have significant effect on use of the service. Often these attributes are ignored while formulating improvement proposals. The negative signs associated with quantitative attributes indicate that use of the service decreases with an increase in the value of in-vehicle travel time and waiting time.

The interpretation of model coefficients is not straightforward except for significance. Therefore, the marginal rates of substitution between attributes and cost are calculated. These substitution rates (ratios between coefficient of attribute/level and coefficient of cost) can be interpreted as marginal WTP for a unit change for continuous attributes. In the case of effects-coded qualitative attributes, estimation of WTP values is based on rescaled coefficients of the levels where the last level is considered the reference level (made equal to 0) and the estimated values are with reference to the last level (i.e., for a shift from last level to the level under consideration). While standing in a crowded vehicle is taken as a reference level for the attribute “travel discomfort,” very high noise is taken as reference level for the attribute “noise.” For the attribute “appearance,” poor appearance is taken as the reference level. The marginal WTP estimates for various attributes/levels as obtained from MNL and RPL models are shown in Tables 4 and 5 for commuting and noncommuting trips, respectively.

Table 4 shows the WTP values for commuting trips from different model specifications. The WTP value of in-vehicle travel time ranges from 7.35 paise/min or INR 4.4 (≈ 0.10 USD) per hour to 8.13 paise/min or INR 4.87 (≈ 0.11 USD) per hour. The WTP value for the waiting time ranges from 3.08 paise/min to 3.3 paise/min.

Table 5 shows the WTP values for noncommuting trips from different model specifications. The value of in-vehicle travel time ranges from 6.62 paise/min or INR 4 (≈ 0.175 USD) per hour for the income group with household monthly income less than INR 20000 (≈ 455 USD) to 12.9 paise/min or INR 7.74 (≈ 0.175 USD) per hour for the income group with household monthly income more than INR 20000 (≈ 455 USD). The WTP value for the waiting time ranges from 2.96 paise/min to 3.11 paise/min. WTP values for in-vehicle travel time obtained from the present study are comparable with those reported by other studies in developing countries. The
Table 4. WTP Values for Commuting Trips

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Unit</th>
<th>WTP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MNL</td>
</tr>
<tr>
<td>In-vehicle travel time</td>
<td>Paise/min</td>
<td>7.35</td>
</tr>
<tr>
<td>Waiting time</td>
<td>Paise /min</td>
<td>3.07</td>
</tr>
<tr>
<td>Comfortable seating</td>
<td>Paise /km</td>
<td>15.66</td>
</tr>
<tr>
<td>Get seat en-route</td>
<td>Paise /km</td>
<td>13.89</td>
</tr>
<tr>
<td>Comfortable standing</td>
<td>Paise /km</td>
<td>4.76</td>
</tr>
<tr>
<td>Very low noise</td>
<td>Paise /km</td>
<td>26.34</td>
</tr>
<tr>
<td>Low noise</td>
<td>Paise /km</td>
<td>24.84</td>
</tr>
<tr>
<td>High noise</td>
<td>Paise /km</td>
<td>2.35</td>
</tr>
<tr>
<td>Good appearance</td>
<td>Paise /km</td>
<td>8.99</td>
</tr>
</tbody>
</table>

*100paise = 1 Indian Rupee (INR) and 44 INR=1 US$

Table 5. WTP Values for Noncommuting Trips

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Unit</th>
<th>WTP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MNL</td>
</tr>
<tr>
<td>In-vehicle travel time</td>
<td>Paise/min</td>
<td>7.90</td>
</tr>
<tr>
<td>Waiting time</td>
<td>Paise /min</td>
<td>2.96</td>
</tr>
<tr>
<td>Comfortable seating</td>
<td>Paise /km</td>
<td>16.04</td>
</tr>
<tr>
<td>Get seat en-route</td>
<td>Paise /km</td>
<td>13.42</td>
</tr>
<tr>
<td>Comfortable standing</td>
<td>Paise /km</td>
<td>1.65</td>
</tr>
<tr>
<td>Very low noise</td>
<td>Paise /km</td>
<td>29.08</td>
</tr>
<tr>
<td>Low noise</td>
<td>Paise /km</td>
<td>26.03</td>
</tr>
<tr>
<td>High noise</td>
<td>Paise /km</td>
<td>3.51</td>
</tr>
<tr>
<td>Good appearance</td>
<td>Paise /km</td>
<td>10.04</td>
</tr>
</tbody>
</table>

*100paise = 1 Indian Rupee (INR) and 44 INR=1 US$. **For household income more than INR 20000
WTP values for in-vehicle travel time as reported for Bangladesh, Tanzania, and Ghana are 0.06 USD/hr, 0.18 USD/hr and 0.18 USD/hr, respectively (I.T. Transport 2005). The WTP for in-vehicle travel time as reported for Mumbai (India) bus users is 0.28 USD/hr (MMPG 1997).

Very low noise (WTP in the range of 25.7 paise/km to 26.3 paise/km) is valued almost 3 times as much as good appearance (WTP in the range of 8.99 paise/km to 9.77 paise/km), and comfortable seating (WTP in the range of 15.66 paise/km to 16.64 paise/km) is valued nearly 1.5 times than good appearance by commuting trip-makers. Similarly, noncommuting trip-makers also valued very low noise (WTP in the range of 27.93 paise/km to 29.17 paise/km) about 3 times as much as good appearance (WTP in the range of 10.04 paise/km to 10.94 paise/km) and comfortable seating (WTP in the range of 16.04 paise/km to 17.02 paise/km) nearly 1.6 times than good appearance. Tables 4 and Table 5 show that there is a big leap in WTP values between high noise to low noise, and between comfortable standing to get seat en-route. Not surprisingly, levels of the attribute “noise” carry high WTP values across qualitative attributes, which closely map the current noise levels in buses. Marginal WTP for in-vehicle travel time is nearly 2.5 times than that for waiting time. High WTP values for qualitative attributes not only indicate the importance of these attributes but also reflect the poor quality of services being offered now. Across MNL and RPL models, these is a little gain in WTP values for all the attributes and levels except for levels very low noise, low noise, and comfortable standing from RPL models. Similar observations with gains in some attributes and loss in others are reported by Train (1998) for recreational demand, Revelt and Train (1999) in household appliance study, Bhat (1998) in mode choice modeling, Algers et al. (1998) while estimating value of travel time, Carlsson (1999) while estimating value of travel time for business class, and Alpizar and Carlsson (2001) in mode choice modeling.

A comparison of WTP estimates between commuting and noncommuting trips indicates that WTP values are sensitive to trip purpose. While commuting trip-makers have higher WTP values for some attributes/levels, noncommuting trip-makers generally have higher WTP values for qualitative attributes. In addition, heterogeneity around the mean of the in-vehicle travel time indicates that the WTP for in-vehicle travel time is more for noncommuting trips made by the high-income group.

WTP estimates indicate that MNL models in this case marginally underestimated the benefits that can be derived from travel time and waiting time. In general,
a comparison of estimates from a standard logit and a random specification depends on the data and the assumed distribution for random parameters.

An interesting observation from the estimates is that in all the three estimates the travel time is valued higher than waiting time, unlike in developed countries. This high value of travel time may be attributed to the poor comfort conditions inside the vehicles and longer journey periods. A similar observation is reported by Mumbai Metro Study (1997) for bus users in Mumbai (India). The findings outlined in this article may be helpful while estimating the marginal WTP values for other cities of developing countries.

Conclusions
Users willingness to pay for various quantitative and qualitative attributes of bus transportation system is estimated with reference to a case study in Kolkata, India. It is found that apart from quantitative attributes, the urban bus user’s choice is also influenced by qualitative attributes. This emphasizes the need for considering qualitative attributes while formulating improvement proposals and estimating user benefits in developing countries. The effect of model specification on WTP estimates is studied. For some attributes, the WTP estimates obtained from standard MNL are found lower than those from RPL models. However, for other attributes, RPL models produced lower WTP estimates than MNL models. WTP values are also estimated separately for commuting and noncommuting trips. The noncommuting trip-maker’s WTP values are generally high for qualitative attributes. While studying the effect of socioeconomic attributes on WTP values, WTP is shown to be higher for noncommuting trips made by the high-income group.

The article also demonstrates the development of acceptable RPL models with constrained triangular distribution for random parameters. Though constrained triangular distribution has several advantages over other distributions, its application is not explored widely by researchers in RPL model development. The present application is expected to encourage the use of constrained triangular distribution as an alternative to other commonly used distributions while developing RPL models. Though contextual, the findings of the article may be used by planners and policy-makers to formulate strategies for improvement of urban bus transportation system in developing countries.
Acknowledgments

The work presented in this article is based on the research project sponsored by the Department of Science and Technology (DST), Ministry of Science and Technology, Government of India. Suggestions from John Rose (ITLS, Sydney) are greatly acknowledged.

References


**About the Authors**

**C.V. Phanikumar** (phani@civil.iitkgp.ernet.in) is a research scholar in transportation engineering in the Department of Civil Engineering at Indian Institute of Technology, Kharagpur, India. His research interests include transport economics, transport planning, and framing-related policy issues using behavioral analysis. He has experience in geometric design of highways and traffic engineering.

**Bhargab Maitra** (bhargab@civil.iitkgp.ernet.in) obtained his Ph. D from IIT, Bombay, India, and is presently an assistant professor in transportation engineering in the Department of Civil Engineering at Indian Institute of Technology, Kharagpur. His research interests include traffic management, operation and maintenance, congestion modeling, and improving public transport systems using behavioral analysis. He has published numerous papers in both national and international journals.

Rabi G. Mishalani and Mark M. McCord, Ohio State University

Abstract

This study quantifies the relationship between perceived and actual waiting times experienced by passengers awaiting the arrival of a bus at a bus stop. Understanding such a relationship would be useful in quantifying the value of providing real-time information to passengers on the time until the next bus is expected to arrive at a bus stop. Data on perceived and actual passenger waiting times, along with socioeconomic characteristics, were collected at bus stops where no real-time bus arrival information is provided, and relationships between perceived and actual waiting times are estimated. The results indicate that passengers do perceive time to be greater than the actual amount of time waited. However, the hypothesis that the rate of change of perceived time does not vary with respect to the actual waiting time could not be rejected (over a range of 3 to 15 minutes). Assuming that a passenger’s perceived waiting time is equal to the actual time when presented with accurate real-time bus arrival information, the value of the eliminated additional time is assessed in the form of reduced vehicle hours per day resulting from a longer headway that produces the same mean passenger waiting time. The eliminated additional time is also assessed in the form of uncertainty in the headway resulting in the same extra waiting time. Naturally, such benefits of passenger information can
only be confirmed when the actual effect of information on the perception of waiting time is quantified.

Motivation and Hypothesis
The background motivating this study is first discussed, and the objective is then presented. Real-time bus arrival information—for example, delivered to prospective passengers waiting at bus stops via variable messages signs (VMSs)—can be useful to transit passengers for a multitude of reasons. Passengers can use their waiting time more productively, select which route they would want to take, or choose to select an alternative mode of transportation. Whatever the prospective passengers’ choices are, providing them with real-time information reduces the uncertainty inherent to transit systems. Empirical evidence shows that the time travelers spend outside the transportation vehicle of choice (e.g., waiting at a stop) is more onerous than the time they spend inside the vehicle in motion to their destination (Ben-Akiva and Lerman 1985). This is partly due to the higher degree of uncertainty associated with waiting for a transit vehicle. This phenomenon is well characterized by Duffy (2002): “People don’t mind waiting for a bus if they know how long it’s going to be. Even if they have to waste the time, at least they know it’s going to be 15 minutes. Otherwise they’re sitting there thinking the bus will be along in about two minutes, and when it doesn’t show, then they start getting frustrated.” In general, reducing waiting time uncertainty is expected to improve passenger satisfaction, and ultimately increase bus ridership.

Mishalani et al. (2000) studied the value of information to passengers in terms of using the waiting time more effectively, while Hickman and Wilson (1995) studied the value in terms of improved route choice. This research focuses on passengers’ perceptions of their waiting time at stops (outside the vehicle) and, as a result, the possible reduction in such times when real-time passenger information is provided. To study the perceptions of waiting time, a survey of prospective passengers at bus stops was conducted. The collected data were then analyzed whereby the relationship between perceived and actual waiting times was investigated.

The main hypothesis of this study is that without real-time bus arrival information, passengers are likely to perceive waiting time to be greater than it actually is. When accurate bus arrival information is provided, it is assumed that passengers will perceive their waiting time to be equal to the actual waiting time. In this case, a passenger will arrive at a stop and look at the VMS, which will display the min-
utes until the next bus is expected to arrive on the route of interest. The VMS will continue to update the expected time while the passenger waits.

Without bus arrival information provided to passengers, the relationship between perceived and actual waiting times is expected to follow a function where the perceived time is greater than the actual time. The form of the function might depend on the magnitude of the waiting time. For example, the additional time due to exaggerated perceptions may be further magnified under long waits in comparison to short waits. One can also imagine the opposite situation, where longer waits may be perceived more accurately due to the more conscious recognition of time under such conditions.

The main objective of this study is to model and quantify the difference between perceived and actual passenger waiting times at bus stops in the absence of accurate real-time bus arrival information and to investigate the effects of duration of the actual waiting time and socioeconomic variables on this difference. This objective is achieved in the context of a pilot study by estimating models that describe waiting time as perceived by passengers waiting for buses at stops. To do so, appropriate field data were collected. Once the difference between perceived and actual waiting times were quantified, possible factors that might affect the magnitude of the difference were explored. Moreover, an analysis of the potential benefits of providing accurate bus arrival time information at bus stops was carried out.

**Data Collection**

Data were collected by surveying passengers waiting at bus stops for Campus Area Bus Service (CABS) buses, which are operated by the Transportation and Parking Office of Ohio State University in Columbus. CABS serves the campus community, which includes close to 50,000 students, resulting in an annual ridership of approximately 4 million. The operation consists of 15 to 20 40-foot buses running simultaneously on several routes of lengths ranging from 2 km to 8 km on and in the areas surrounding the campus.

The transit service used in this pilot study is small enough to be manageable, yet large enough to reflect situations pertinent to more extensive transit services in urban areas. Nevertheless, it would be important to build upon this pilot study in future research by examining larger transit systems. Extensions to larger systems would render the findings more applicable to a wider set of conditions, most notably, longer routes and more heterogeneous traveling populations.
Three students surveyed 83 passengers over a period of approximately one year, from spring 2001 to spring 2002. A surveyor went to a bus stop, noted the arrival time of a passenger, and later asked him or her a series of questions. A response rate close to 100 percent was achieved.

**Bus Stops**

A set of appropriate bus stops was first determined for the purpose of conducting the surveys on the basis of four criteria. The first criterion was to choose a stop that does not serve many routes. In fact, a stop that serves only one bus route is ideal. Fewer routes will help the interviewer know, or possibly guess, which route a passenger is going to choose before he or she gets on the bus. A surveyor sitting at a stop has a general idea of when the next bus will arrive based on both the published headway and observations of the buses over a period of time. If a bus stop serves many different routes, it becomes more difficult, if not impossible, to know the route a random passenger plans to use. The passenger might then board a bus before the surveyor has a chance to conduct the interview, thus missing a data collection opportunity.

The second criterion for selecting a good bus stop for surveying purposes is to use a stop that serves routes with longer headways. Longer headways are attractive because it is desirable to have the option to survey a passenger after a notable wait. While data with short waiting times are needed for a complete data set, not all the data should be collected after a 3- to 4-minute wait. A longer headway allows the interviewer more options on when to survey passengers and observe longer waits.

One issue with longer headways, though, is whether a schedule is published. This issue leads to the third criterion. It is helpful to select bus stops where passengers arrive totally randomly. Such arrivals typically occur when only the headway on a route is published, rather than the scheduled time of bus arrivals. When schedules are published and headways are long, most passengers will likely arrive shortly before published arrival times, thus reducing the opportunity of observing relatively longer waiting times.

Finally, it is productive to collect data at stops with relatively high demand. Passengers must arrive frequently at a bus stop to ensure that interviewers do not experience a large amount of idle time. Otherwise, longer survey times will be needed to produce the same number of observations. Based on the above criteria, a total of five stops on three routes were selected. Two of the routes have published head-
ways of 10 minutes and one a published headway of 15 minutes. None, however, has a published schedule.

**Interviews and Observations**
During the survey process, when a passenger arrived at the bus stop, the arrival time was noted. The first one or two passengers that arrived after the previous bus departure were selected. This selection strategy ensured that passengers with a wider range of wait times were interviewed and that the interviewer was able to survey them without the bus coming before the interview was complete. The interviewer decided when to survey the passenger. This is selected largely on when the next bus was thought to be arriving. The interviewer typically began the interview with a passenger at least a minute before the expected bus arrival to ensure enough time to complete the interview without the passenger feeling anxious about catching the bus. As already discussed, the interviewers had a good idea as to when the next bus would arrive because of their knowledge of the previous bus arrival time and service operations.

In addition to the surveyor’s name, date, and weather conditions, the following is a list of variables observed for each interviewed waiting passenger:

1. origin of the passenger (bus stop where the survey is conducted);
2. passenger’s arrival time to the stop;
3. time the passenger was surveyed;
4. passenger’s gender;
5. passenger’s race;
6. passenger’s perceived waiting time;
7. whether the passenger was wearing a watch;
8. destination of the passenger;
9. maximum time the passenger would be willing to wait were real-time bus arrival information provided;
10. approximate walking time to the destination from the bus stop at which the passenger was waiting;
11. whether the passenger had a time constraint (such as making it to a class at a certain time);
12. familiarity of the passenger with the transit service, as measured by frequency of use in number of trips per day;
13. passenger’s car ownership status; and
14. passenger’s affiliation with the university (undergraduate student, graduate student, staff, faculty, or visitor).

Items 1 through 7 were always recorded by the surveyor, while items 8 through 14 were generally collected, time permitting. The first question, asked after a brief introduction, was “In minutes, how long do you think you have been waiting?” It was important to specify the desired level of accuracy to the passengers to avoid the possibility of their rounding off to the nearest 5 minutes.

**Perceived Waiting Time Models**

The model development and estimation aimed at quantifying the difference between perceived and actual waiting times, along with the exploration of factors influencing that difference, are discussed in this section. The modeling consists of two parts. First, a simple ordinary least squares (OLS) regression model of perceived versus actual waiting times is estimated. Hypothesis tests are then applied to determine whether a significant difference between the two variables exists and to assess the nature of the relationship. Second, the impact of socioeconomic variables on the relationship between perceived and actual waiting times is investigated.

**Perceived versus Actual Waiting Times**

A simple linear regression model of the following form is estimated:

\[ p = \beta_0 + \beta_1 a + \epsilon \]  

(1)

where:

- \( p \) is the perceived waiting time
- \( a \) represents the actual waiting time
- \( \beta_0 \) is the parameter representing the intercept of the regression line
- \( \beta_1 \) denotes the parameter representing the slope of the regression line
- \( \epsilon \) is a random variable with mean of 0
The estimation results, shown in Table 1, indicate that the perceived waiting time is greater than the actual within the range of the data set whereby the actual waiting time varies between 3 and 15 minutes. However, it is important to apply statistical hypothesis testing to assess the significance of this finding.

**Table 1. Estimation Results of Perceived vs. Actual Waiting Times**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Est. Parameter</th>
<th>Standard Error</th>
<th>t-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.33</td>
<td>0.48</td>
<td>2.77</td>
</tr>
<tr>
<td>Actual wait time</td>
<td>0.92</td>
<td>0.076</td>
<td>11.96</td>
</tr>
</tbody>
</table>

No. of observations = 83, $R^2 = 0.634$

In addition to the fairly high corrected goodness-of-fit measure $R^2$, the parameter estimate $\hat{\beta}_1$ with a t-statistic of 11.96 is significantly different from 0 at the 0.01 level, clearly indicating that the perceived waiting time varies positively with the actual waiting time. More interestingly, though, is the null hypothesis that the parameter estimate $\hat{\beta}_1$ is equal to 1. For the general case when testing whether a parameter estimate is equal to some non-0 value, the t statistic is given by Pindyck and Rubinfeld (1998). In the case of the null hypothesis of interest, namely that $\hat{\beta}_1 = 1$, the t-statistic takes the value of -1.11 (see Wirtz [2002] for the details behind this calculation), resulting in the failure to reject the hypothesis at the 0.10 level. That is, the slope of the regression line is not statistically different from 1 and, hence, the data do not support the notion that passengers perceive time any more or less differently after waiting for longer periods.

In addition to the actual time ranging between 3 and 15 minutes in the data set, 94 percent of the actual waiting time observations range between 3 and 9 minutes. Therefore, it is quite possible that, under longer waiting times, the perceptions may increase at a different rate than unity with respect to the actual time. Moreover, given that the rejection of a hypothesis does not imply its acceptance, it is also worth pointing out that the less-than-1 value of the estimated slope means that (on the basis of the parameter estimates of Table 1), as the wait gets longer, the exaggerated perceptions become more closely aligned with reality for actual waiting times as long as 15 minutes. This observation is consistent with the notion that longer waits are associated with more accurate perceptions of time.
In any event, the failure to reject the hypothesis that $\hat{\beta}_1$ is equal to 1 renders the testing of the additional null hypothesis that $\hat{\beta}_0$ is equal to 0 more critical. More specifically, only a parameter estimate $\hat{\beta}_0$ that is statistically greater than 0 would confirm that the perceived waiting time is greater than the actual waiting time. Indeed, with a t-statistic of 2.77, the estimate of 1.33 is significantly different from 0 at the 0.01 level.

Thus far, the null hypotheses $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$ were tested separately. It is also useful to test the two hypotheses jointly using an F-test (Pindyck and Rubinfeld 1998) where the joint null hypothesis is $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$, which reflects the function $p = a + \varepsilon$. In other words, on average, the perceived waiting time equals the actual waiting time. If this test fails to reject the null hypothesis, there is no statistical difference between what passengers perceive and the actual waiting time. The F-statistic for the hypothesis of interest is 0.46 (see Wirtz [2002] for the details behind the calculation). This value implies the rejection of the null hypothesis at the 0.01 level (the critical value corresponding to this level is 4.91), thus confirming that passengers do perceive waiting times to be greater than actual times, at least in the range between 3 and 15 minutes.

Finally, it is worth pointing out that the finding that perceived waiting times are statistically greater than actual waiting times within the range of the data identifies one possible contributor to the well-established conclusion that out of vehicle travel time (of which waiting time at a bus stop is a possible component) is more onerous to travelers than in-vehicle travel time (Ben Akiva and Lerman 1985).

**Socioeconomic Variables**

Introducing socioeconomic variables to the original specification of equation (1) is also considered. Investigating the socioeconomic variables listed above, none are found to have parameter estimates significantly different from 0 at the 0.05 level. This result could partly be attributed to the fact that data on some of these variables were not collected for each interviewed passenger because of time constraints during the survey process. For some variables, the number of observations is as low as 46. Nevertheless, it is worthwhile to point out two variables that exhibit some noteworthy effect.

In particular, the walking time to the destination from the bus stop at which the passenger is waiting and the presence of a time constraint do exhibit some explanatory value. Data on both variables were provided by passengers in response to interview questions. The former is measured in minutes, and the latter is a dummy
variable that takes a value of 1 if a time constraint is present and 0 otherwise. These two variables are included in the model as follows:

\[ p = \beta_0 + \beta_1 a + \beta_2 w + \beta_3 c + \varepsilon \]  

(2)

where:

- \( w \) is the walking time from the bus stop the passenger is waiting at to the destination
- \( c \) represents the time constraint dummy variable
- \( \beta_2 \) and \( \beta_3 \) are parameters

The details behind the identification of these variables are discussed in Wirtz (2002). The estimation results are given in Table 2. The t-statistics of the estimates of the intercept and the parameters of the two additional variables reflect values significantly different from 0 at the 0.10 level. Furthermore, the null hypothesis that \( \beta_2 = \beta_3 = 0 \) is rejected at the 0.10 level by applying the F-test, indicating that this model adds significant explanatory value (at the 0.10 level) with respect to the original model.

### Table 2. Estimation Results with Socioeconomic Variables Included

<table>
<thead>
<tr>
<th>Variable</th>
<th>Est. Parameter</th>
<th>Standard Error</th>
<th>t-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.68</td>
<td>0.94</td>
<td>1.80</td>
</tr>
<tr>
<td>Actual wait time</td>
<td>0.77</td>
<td>0.18</td>
<td>4.36</td>
</tr>
<tr>
<td>Walking time</td>
<td>0.047</td>
<td>0.028</td>
<td>1.71</td>
</tr>
<tr>
<td>Time constraint</td>
<td>-0.91</td>
<td>0.62</td>
<td>-1.47</td>
</tr>
</tbody>
</table>

No. of observations = 46, \( R^2 = 0.529 \)

One possible explanation of the positive parameter value of the walking time to destination variable is that a passenger going farther is more dependent on the bus service (i.e., practically, the passenger does not have an alternative transportation option for the desired trip on the university campus) and is therefore likely to be
more frustrated from repeatedly having to take the bus than a passenger taking the bus only occasionally. Such frustration would result in a further exaggeration of perceived waiting time. Moreover, passengers with shorter walk times who choose to take the bus rather than walk to their destination are likely to be less sensitive to the time of their wait—given the choice that they have made—than those with longer walking times.

As for the presence of the time constraint variable, one possible explanation for the negative sign of its corresponding parameter estimate is that a passenger with a time constraint is more conscious of how much time is actually elapsing. Being conscious of the elapsed time would tend to lessen exaggerations in misperceptions compared to the case when a time constraint is not present. Alternatively, one might expect the opposite effect, whereby the presence of a time constraint introduces anxiety that leads to an increase in the perceived waiting time. The negative value of the parameter estimate corresponding to the time constraint dummy variable supports the dominance of the former effect. Future studies employing more comprehensive data sets would shed more light on this interesting finding. Finally, similar hypothesis tests to those applied to the original model (i.e., Table 1) are also applied to the model including the walking time and time constraint variables (i.e., Table 2). As in the case of the original model, the t-test fails to reject the hypothesis that the parameter associated with the actual waiting time is equal to 0 at the 0.05 level. Moreover, the estimate of the intercept is significantly different from 0 at approximately the 0.07 level.

Potential Value of Providing Passenger Information

Empirically, the mean perceived waiting time of the passengers surveyed is 6.61 minutes, while their mean actual waiting time is 5.77 minutes. In light of the quantified difference of 0.84 minutes, the assessment of the value of providing passenger information on the expected time until the next bus arrival consists of two parts. First, the additional 0.84 minutes of perceived waiting time are converted to equivalent vehicle-hours per day saved if real-time information is provided at stops while maintaining the same mean perceived passenger waiting time, under the assumption of deterministic headways. Second, the additional 0.84 minutes of perceived waiting time are converted into an equivalent headway uncertainty. In the event that the effects of actual waiting time and socioeconomic variables on the difference between the perceived and actual waiting time are found to be statistically significant in the context of a more extensive study, the expected dif-
ference given the conditions of interest—instead of the unconditional difference of 0.84 minutes—would be used in the forthcoming analyses.

In both cases, it is assumed that real-time passenger information will eliminate exaggerated perceptions of waiting time. Thus, the two measures provide an assessment of the value of introducing such information in terms of known performance measures. As illustrated below, the introduction of real-time information systems could result in a reduced operating cost or an increased passenger satisfaction. Such benefits could ultimately lead to increased ridership for public transit, depending on the policies adopted by the transit agency in conjunction with the introduction of such information systems.

Naturally, the subsequent discussion of the value of introducing real-time passenger information depends on the validity of the assumption that such information will produce accurate perceived waiting times. While this assumption requires further research to be validated, it is not unreasonable to adopt this assumption in the absence of evidence suggesting otherwise.

**Equivalent Vehicle-Hours**

If the exaggeration in the perceptions of waiting time is eliminated by providing real-time passenger information, the quantified difference of 0.84 minutes will be reduced to 0. Under such a condition, the transit agency can achieve the same level of service, as measured by the mean passenger waiting time, by employing fewer vehicle-hours per day. Decreasing vehicle-hours would translate to reduced labor, energy, and vehicle maintenance costs. Assuming that the headway is deterministic (i.e., the headway variance is 0) and passengers arrive in a totally random manner (i.e., arrivals follow the Poisson process), the mean actual passenger waiting time is equal to half the deterministic headway. Furthermore, under deterministic headways and running times, the vehicle-hours per day required for service provision is the ratio of the running time to the headway multiplied by the total hours of service provided per day. Naturally, for a given running time, fewer vehicle-hours per day are required when the headway is longer.

In the event that real-time information is provided and the additional mean of 0.84 minutes of waiting time as perceived by passengers is eliminated, operating at a headway of \( h + 2 \times 0.84 \) minutes—where \( h \) is the headway when real-time passenger information is not provided—will result in the same mean perceived passenger waiting time as that when information is not provided and waiting times are exag-
generated due to passenger perceptions. As a result, the percent reduction $Z$ in the vehicle hours per day required for service provision is given by the following:

$$Z = 1.68(h+1.68)^1$$  \hspace{1cm} (3)

This function is plotted in Figure 1 for values of mean headway ranging from 5 to 30 minutes where the percent reduction in vehicles-hours per day varies between 25.1 and 5.3 percent.

A typical CABS route has a mean running time of 30 minutes, a mean headway of 10 minutes, and operates over a period of 18 hours during a weekday. Under these operating conditions, providing real-time information while maintaining the same mean passenger waiting time results in a reduction of 7.77 vehicle-hours per day from the original total of 54, amounting to a 14.4 percent saving (as seen in Figure 1).

**Figure 1. Percent Reduction in Vehicle-Hours per Day vs. Headway**
**Equivalent Coefficient of Variation**

In this exercise, the difference of 0.84 minutes is converted to an equivalent coefficient of variation in the headway of a route that would increase mean passenger waiting times by 0.84 minutes with respect to the mean wait time under a deterministic headway. In this case, the value of real-time information is assessed in terms of eliminating a certain level of uncertainty in the headway. Assuming that passengers arrive in a totally random manner, the expected waiting time is given by the following (Larson and Odoni 1981):

\[ E[w] = \frac{E[h]}{2} \left( 1 + \frac{\text{Var}[h]}{E[h]^2} \right) = \frac{E[h]}{2} \left( 1 + \text{CV}[h]^2 \right) \]  

(4)

where:

- \( E[w] \) is the expected waiting time for a randomly arriving passenger
- \( E[h] \) represents the expected headway
- \( \text{Var}[h] \) denotes the variance of the headway
- \( \text{CV}[h] \) is the coefficient of variation of headway (the ratio of the standard deviation to the mean)

To obtain the coefficient of variation of headway that results in an increase in mean waiting time equivalent to the difference between the perceived and actual waiting times, \( \text{CV}[h] \) is solved for using equation (4) for given values of \( E[w] \) and \( E[h] \), where the difference \( E[w] - E[h]/2 \) remains equal to 0.84 minutes. (Note: \( E[h]/2 \) is the mean waiting time when the headway is deterministic.) Naturally, the result depends on the value of \( E[h] \), and is given by the following:

\[ \text{CV}[h] = 1.296 E[h]^{1/2} \]  

(5)

This function is plotted in Figure 2 for values of mean headway ranging from 5 to 30 minutes. The equivalent coefficient of variation is seen to vary between 0.580 and 0.237.
The equivalent coefficient of variation of headway is a measure of the bus service quality as perceived by passengers. Elimination of increased perceived waiting times on a bus route that does not exhibit any variability in its headway could conceptually be traded off against increased variability in headways. For example, an increased passenger wait time perception of 0.84 minutes on a route with a deterministic headway of 10 minutes has an equivalent headway coefficient of variation of 0.410, as seen in Figure 2, on a route with random headways and wait time perceptions equal to actual wait times. This value of CV[h] corresponds to a headway standard deviation of 4.10 minutes. As an illustration, such a standard deviation could be produced by a total of two buses bunching, with a shorter headway of 7.1 minutes and a longer headway of 12.9 minutes. Thus, for the data set used in this study and a 10-minute headway, eliminating exaggerated perceived waiting times by providing accurate real-time information on the time until the next bus is expected to arrive can be thought of as equivalent to eliminating bus bunching producing a headway standard deviation of 4.10 minutes.

Figure 2. Equivalent Coefficient of Variation vs. Expected Headway

Summary and Future Research
Through the observation of 83 prospective passengers waiting at bus stops, the relationship between perceived and actual waiting times was investigated. The
results of estimating a linear relationship between the two variables indicate that, while the intercept is significantly greater than 0, the slope’s equality to 1 cannot be rejected. This finding implies that the data do not reveal evidence that the additional perceived waiting time varies with the actual time within the range of 3 to 15 minutes reflected in the data set. Moreover, some socioeconomic variables are found to have explanatory value. In particular, a passenger’s walking time to the destination from the bus stop at which he or she is waiting and the presence of a time constraint reflect an impact on the perceived waiting time. A longer walking time produces a greater exaggeration in the perceived waiting time, while the presence of a time constraint brings the perceived waiting time closer to the actual time.

The mean difference between the perceived waiting time by passengers and the actual time is 0.84 minutes. This additional time is related to both an equivalent savings in vehicle-hours per day and an equivalent coefficient of variation (the ratio of the standard deviation to the mean) in the headway of a route. These relationships are developed in an effort to assess the benefits of eliminating the exaggeration in passengers’ perceptions of waiting time by providing accurate real-time passenger information on the time until the next bus arrival. If the additional amount of time—eliminated by the provision of real-time information—is used to increase the headway to a value such that the mean passenger perceived waiting time remains unchanged, the reduction in the vehicles-hours per day required to provide service is derived assuming a deterministic headway and totally random passenger arrivals. For a running time of 30 minutes, a headway of 10 minutes, and a duration of service of 18 hours per day, a reduction of 7.77 vehicle-hours per day is achieved, amounting to a 14.4 percent saving.

If the additional 0.84 minutes of perceived time is added to the expected waiting time of a passenger waiting for a bus when there is no variation in the headway and assuming totally random passenger arrivals, the equivalent headway coefficient of variation CV\(h\) producing such an additional waiting time is derived. For a mean headway of 10 minutes, CV\(h\) is 0.410 reflecting a standard deviation of 4.10. Thus, providing real-time information at bus stops—for example, via VMSs—could reduce passengers’ mean perceived waiting time by an amount equivalent to that achieved through eliminating the corresponding standard deviation in the headway.

Because passengers perceive waiting times to be greater than actual waiting times at a bus stop, real-time passenger information systems could potentially reduce
the perceived waiting time for buses when providing accurate information. The reduction in perceived waiting times could potentially be translated into reduced operating costs or increased passenger satisfaction and, ultimately, into increased ridership for public transit, depending on the policies adopted by the transit agency in conjunction with the introduction of real-time passenger information systems.

This study demonstrated the feasibility of examining the questions of interest and points to several directions regarding future research. Based on this pilot study, there is good evidence that a difference between perceived and actual waiting times does exist to motivate a more comprehensive data collection and modeling effort. It would be valuable to observe passengers traveling on larger transit systems than the university campus-based system of this study. Larger systems present additional complexity in service that might affect waiting time perceptions (e.g., while waiting during transfers between routes, and when traveling on longer routes) and reflect a more heterogeneous traveling population. A wider range of actual waiting times is also important to allow for testing various specifications characterizing the relationship between perceived and actual waiting times. Moreover, a larger data set with more complete observations of socioeconomic variables is necessary, given the indication that such variables could have an important impact on the perceptions of waiting time. Also, additional socioeconomic variables might influence the perceptions of waiting time and, hence, would be worth observing. Such variables include time of day, whether the passenger has access to time-telling devices other than a watch (e.g., mobile phone, other portable electronics, or public clocks visible from the bus stop), whether the passenger is traveling with a group, and whether a bus operating on a route that does not provide service to the passenger’s desired destination arrives while the passenger is still waiting for a service on another route. In addition, in conducting a larger study on a more extensive transit service, which will inevitably present more degrees of freedom, it is important to pay attention to the design of the sampling strategy and to monitor the response rate. Finally, the assumption that passengers perceive waiting time to be equal to the actual time when accurate real-time information is provided should be investigated. In particular, a study whereby passengers are surveyed both before and after the introduction of real time passenger information systems would be particularly valuable.
Acknowledgments

This study was funded by the Transportation and Parking Office at the Ohio State University. The support of Sarah Blouch and Steven Basinger of the same office is appreciated. Valuable comments by the anonymous referees are also gratefully acknowledged.

References


About the Authors

**Rabi G. Mishalani** (mishalani@osu.edu) is an associate professor in the Department of Civil and Environmental Engineering and Geodetic Science at Ohio State University. He received his doctor of philosophy degree in civil and environmental engineering from Massachusetts Institute of Technology in 1993, his master of science degree in transportation systems from MIT in 1989, and his bachelor of engineering degree in civil engineering from American University of Beirut in 1987. Prior to joining the faculty at OSU in 1997, Dr. Mishalan was a research scientist at MIT’s Center for Transportation and Logistics. He won the National Science Foundation CAREER award in 2001. His main areas of expertise include the application
of probability modeling and statistical methods to the planning and managing of transportation systems. His current research focuses on public transit operations in the presence of vehicle location information, highway operations control and traveler guidance and the role of remotely sensed traffic data, and infrastructure maintenance decision making in the presence of automated condition data collection technologies.

Mark R. McCord (mccord.2@osu.edu) is a professor in the Department of Civil and Environmental Engineering and Geodetic Science at Ohio State University. He received his doctor of philosophy degree in civil engineering from Massachusetts Institute of Technology in 1983, his master of science degree in industrial engineering from Stanford University in 1978, and his bachelor of science degree in civil engineering from Purdue University in 1977. While a doctoral student at MIT, Dr. McCord was a Chateubriand Fellow at Ecole Centrale des Arts et Manufactures in Paris during the 1981–82 academic year. Shortly after joining the faculty at OSU in 1983, he won the National Science Foundation Presidential Fellowship award. His main areas of expertise include the application of decision analysis and subjective probability theory to the planning and managing of transportation systems. His current research focuses on the use of vehicle location data in public transit operations, the assessment of remote sensing technologies in transportation planning, and regional travel demand measurement and forecasting.

John Wirtz (wirtz.13@osu.edu) is an engineer at Edwards and Kelcey, Inc. in Chicago. He received his master of science degree in civil and environmental engineering from Northwestern University in 2004 and his bachelor of science degree in civil and environmental engineering from Ohio State University in 2002. During both his undergraduate and graduate studies, his research focused on transportation systems analysis and modeling.
Large-Scale Transit Network Optimization by Minimizing User Cost and Transfers

Fang Zhao, Florida International University

Abstract

This article proposes a methodology for developing optimal transit networks (route structures and headways) that minimizes transit transfers and total user cost while maximizing service coverage, given information on transit demand, transit fleet size, and the street network of the transit service area. The research provides an effective mathematical computational tool with minimal reliance on heuristics. The methodology includes representation of transit route networks and solution search spaces, objective functions representing total user cost and unwillingness of users to make transfers, and a global search scheme based on simulated annealing. The methodology has been tested with published solutions to benchmark problems and has been applied to a large-scale realistic network optimization problem in Miami-Dade County, Florida.

Introduction

As congestion in large urban areas continues to worsen and gas prices began to rise in the recent years, the attractiveness of public transit as an alternative to private cars has also been growing. However, for a public transit system to help meet the growing travel demand and alleviate the congestion problem, it must be able to provide reasonable travel time and convenience relative to private vehicles. Travel time and convenience are affected directly by the configuration of a transit net-
work (TN) and service frequency, although other service and traffic characteristics and pedestrian environment will also have an impact on the willingness of the public to use transit. The quality of a TN may be evaluated in terms of a number of parameters including route directness, service coverage, operator cost, transit user cost (including waiting, in-vehicle, and transfer times), and the average number of transfers required to accomplish a trip. Route directness may be measured by the additional travel time incurred to a transit user when a bus does not follow the most direct route between the user’s origin and destination. Service coverage refers to the percentage of total estimated demand (i.e., transit trips) that may be potentially satisfied by the transit services provided, based on a given transit route network. Operator cost is the cost to a transit property to provide transit services within a given network. Transfers are a result of not being able to provide direct services between all pairs of origins and destinations.

Transfers are known to discourage transit use. According to a survey conducted by Stern (1996) of various transit agencies in the United States, about 58 percent of the respondents believed that transit riders were willing to transfer only once per trip. Reducing transfers, therefore, has great potential in increasing the attractiveness of public transit and ridership. Transfers may be reduced by optimizing transit network configuration, or optimally laying out transit routes such that the services are as direct as possible and transfers are minimized. Improvements of TN configuration may also lead to lower transit operating cost and more services provided, which, in turn, help increase transit use.

In TN optimization, route network layouts and route headways are sought that minimize the overall cost of providing transit services, which is generally considered to have two components: user cost and operator cost. Unfortunately, TN design optimization processes that attempt to find global optimal solutions from a search space with reasonable completeness suffer from combinatorial intractability. Newell (1979) observes the difficulty in developing efficient TN optimization methods with traditional mathematical programming techniques and points out that TN design optimization “is generally a nonconvex (even concave) optimization problem for which no simple procedure exists short of direct comparisons of the various local minima.” Furthermore, the resultant system for a TN problem is usually a NP-hard, mixed combinatorial optimization problem that is unlikely to be solved with traditional mathematical optimization techniques. The NP-hard problem (the hard problem in nondeterministic polynomial problem/algorithm class) refers to a problem for which the number of elementary numeri-
cal operations is not likely to be expressed or bounded by a function of polynomial form where the variable(s) of the function reflect(s) the size of the problem. The NP-hard intractability is due to the need to search for optimal solutions from a large search space made up by all possible solutions. A mixed problem refers to a problem that involves both continuous and discrete variables; a combinatorial problem usually refers to an integer optimization problem where the unknown variable set (called combinatorial set) consists of all feasible integer subsets of a larger base integer set. In TN optimization, the base set is the set of all street nodes that are suitable to serve as transit stops, and the combinatorial set consists of all street paths (subsets or integer vectors of the base street node set) in the street network that are suitable for transit vehicle operations. Even for a small street node set, the corresponding combinatorial set (i.e., the set that includes all possible paths) may be huge. Baaj and Mahmassani (1991) observe that large-scale TN optimization problems tend to suffer from several forms of difficulties with traditional mathematical approaches, such as nonlinearity, nonconvexity, multiobjectives, and combinatorial intractability due to the discrete nature of the problems. Similar observations are also made by Ceder and Wilson (1986), Charkroborty and Dwivedi (2002), and Zhao and Ubaka (2004), among others. These seem to be why the solutions to most TN optimization problems in practice are either relying on certain heuristic assumptions or are limited to relatively small or idealized networks. To date, the solutions to large-scale transit network problems that include both route network and headway as design components have been mostly limited to the use of various heuristic approaches where the solution search schemes are based on a collection of design guidelines, criteria established from past experiences, and cost and feasibility constraints.

In recent years, genetic algorithm (GA) has been applied to various TN optimization problems. GA is a stochastic algorithm based on natural evolution principle (i.e., genetic inheritance and the Darwinian strife for survival process). Mathematically, GAs may be categorized as weak solution search schemes that make few assumptions about problem domains and function properties, such as the smoothness, uniqueness, or compatibility of the objective functions, design parameters, and constraints. While this makes GAs attractive and popular for complex problems, it also causes GAs to suffer from combinatorial explosive solution costs due to huge solution search spaces often associated with large-scale problems. In the current TN literature, most GA applications are limited to small- or medium-sized network problems. Recently, Agrawal and Mathew (2004) applied a GA approach to a large-scale transit network. However, the travel demand (about 900 origin-desti-
nation pairs) was relatively small, and the search method required multiprocessor parallel processing due to intensive computation involved.

Table 1 summarizes the main features of some of the approaches reported in the literature. In the table, H&M indicates a combination of both mathematical programming methods and heuristic search schemes; MATH stands for mathematical optimization; H&M/AI means a combination of H&M and artificial intelligence techniques; and multiconstraints indicates use of multiple constraints such as maximum/minimum route length, maximum number of routes, minimum frequency, etc. Due to space limitations, the merits, solution strategies, and applicability to practical problems of the individual approaches are not discussed. Detailed information about various optimization approaches may be found in Fan and Machemehl (2004), which provides an extensive review and comparison of various optimization methods for TN design, and Zhao and Gan (2003), among others.

### Table 1. Main Features of Some Approaches Used in Transit Network Design

<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Year</th>
<th>Author</th>
<th>Optimization Objectives</th>
<th>Design Variables</th>
<th>Solution Approaches</th>
<th>Constraints</th>
</tr>
</thead>
<tbody>
<tr>
<td>1979</td>
<td>Dubois et al.</td>
<td>Gen. user cost</td>
<td>Route &amp; frequency</td>
<td>H&amp;M</td>
<td>Operating cost</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1979</td>
<td>Mandl</td>
<td>Gen. time</td>
<td>Route</td>
<td>H&amp;M</td>
<td>Coverage &amp; directness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1986</td>
<td>Ceder &amp; Wilson</td>
<td>Gen. user &amp; operator cost</td>
<td>Route, frequency &amp; vehicle scheduling</td>
<td>H&amp;M</td>
<td>Multiconstraints &amp; fleet size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1988</td>
<td>Van Nes et al.</td>
<td>No. of direct trips</td>
<td>Route &amp; frequency</td>
<td>H&amp;M</td>
<td>Operating cost &amp; fleet size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1991</td>
<td>Baaj &amp; Mahmassani</td>
<td>Multi-objects</td>
<td>Route &amp; frequency</td>
<td>H&amp;M/AI</td>
<td>Multiconstraints</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>Bookbinder et al.</td>
<td>Disutility function-transfer inconvenience</td>
<td>Timetable/headway (offset time)</td>
<td>MATH</td>
<td>Heuristic guidelines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1994</td>
<td>Shih &amp; Mahmassani</td>
<td>Multi-objects</td>
<td>Route &amp; frequency</td>
<td>H&amp;M</td>
<td>Multiconstraints</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>Pattnaik et al.</td>
<td>Gen. user &amp; operator time</td>
<td>Route network</td>
<td>GA</td>
<td>Frequency &amp; load factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2002</td>
<td>Chakroborty &amp; Dwivedi</td>
<td>Transfer directness</td>
<td>Route network</td>
<td>GA</td>
<td>Coverage &amp; user cost</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>Chien et al.</td>
<td>Total operator &amp; user cost</td>
<td>Route shape &amp; headway</td>
<td>MATH</td>
<td>Route length, waiting time, load factors, etc.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>Ngamchai &amp; Lovell</td>
<td>Total operator &amp; user cost</td>
<td>Route network &amp; headway</td>
<td>GA</td>
<td>Heuristic guidelines</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>Agrawal and Mathew</td>
<td>Total operator &amp; user cost</td>
<td>Route network &amp; headway</td>
<td>GA</td>
<td>Multiconstraints</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>Zhao &amp; Ubaka</td>
<td>Transfer directness</td>
<td>Route network</td>
<td>MATH</td>
<td>Multiconstraints &amp; directness</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The development of the combined simulated annealing and fast descent (SAFD) method in this study has been motivated by the lack of optimization procedures that are capable of tackling large-scale TN problems and finding global optimal solutions in terms of both user and operator costs. Unlike other search algorithms such as various heuristic methods and genetic algorithms, which do not theoretically guarantee good performance to ensure a global optimum, simulated annealing is supported by a solid theory. Under fairly general conditions, it has been shown that a global optimal will be obtained with probability 1 (Hajek 1988). The simulated annealing search scheme used in this study is based on the integrated simulated annealing, tabu, and greedy search method developed by Zhao and Gan (2003), originally designed for finding optimal TN route layouts to minimize passenger transfers.

**Solution Methodology**

For simplicity, the following assumptions were made in this study:

1. The demand pattern, expressed in a transit origin-destination (OD) matrix, remains the same during the period of study.
2. Passengers’ choices of routes are based on the shortest travel time. Terminal times are not included, although may be added easily.
3. Transit vehicles have the same seating capacity.
4. Passengers arrive at transit stops randomly (uniform distribution); therefore, the average waiting time to board a vehicle ($t_{\text{wait}}$) is one half of the headway ($h$), i.e.,

$$t_{\text{wait}} = h/2$$ (1)

The following simple (yet widely used) relationships between TN parameters are employed:

$$L \equiv \frac{h \cdot q_{\text{max}}}{V_{\text{Seat}}} \leq L_{\text{max}}$$ (2)
where:

- \( L \) is vehicle load factor
- \( q_{\text{max}} \) represents the critical link passenger flow of a given route
- \( V_{\text{Seat}} \) indicates vehicle seating capacity
- \( L_{\text{max}} \) signifies a user-defined maximum allowable load factor
- \( 2R_L \) is the round-trip in-vehicle travel time
- \( R_{\text{Fleet}} \) represents the route vehicle fleet size

More complex relationships between various TN parameters may be found in Ngamchai and Lovell (2003), Shih and Mahmassani (1994), Bookbinder and Désilets (1992), among others. The above simplifications do not prevent the proposed SAFD method from solving problems with complex TN parameter relationships such as nonlinear, nonconvex, or stochastic function relationships. Like genetic algorithms, the SAFD search method relies only on the evaluation of objective or constraint functions themselves. Therefore, difficult issues in traditional nonlinear search methods, such as function smoothness, convexity, uniqueness, etc., are not of concern. Theoretically, the proposed SAFD approach should be able to solve transit network optimization problems with dynamic demand iteratively as long as the transit demand (OD matrix) may be obtained after each transit route network and headway update. The challenge of solving dynamic demand problems is to have effective and reliable models to generate a new OD matrix after a new route layout is produced.

**Representation of Transit Service Area, Route Network, Transit Demand, and Headways**

A transit service area is represented by a set of street nodes, denoted as \( N^{(n)} \{n_1, n_2, \ldots, n_n\} \), that are connected to each other by a set of street segments, \( A^{(m)} = \{a_1, a_2, \ldots, a_m\} \). Together, these street node sets and street segment sets are referred to as the *street network* of the transit service area. A street segment \( a_i \) in \( A^{(m)} \) may
be defined by its two end-street nodes \( n_{i_1}, n_{i_2} \), i.e., \( a_i = a_i(n_{i_1}, n_{i_2}) \) and \( n_{i_1}, n_{i_2} \in N(n_i) \), \( (i = 1, 2, ..., m) \). A street segment length is measured by in-vehicle travel time between its two end nodes. A path or a route between any two street nodes is defined as a sequence of nodes, \( p = p(p_1, p_2, ..., p_k) \), and there is one street segment connecting two neighboring nodes in the path. In this study, only undirected networks are considered, but the methodology may be easily extended to directed networks. It is also assumed that the street network is connected, meaning that any two nodes in the street network are connected by at least one path. A TN \( T \) consisting of \( l \) routes may be represented by a set of route/path vectors

\[
T^{(l)} = T^{(l)} \{ r_1, r_2, ..., r_l \}, \quad r_j = r(n_{j_1}, n_{j_2}, ..., n_{j_{s(j)}}), \quad (j = 1, 2, ..., l)
\]

where:

\( n_{jk} \) is the \( k \)-th node \((k = 1, 2, ..., s(j))\)

\( s(j) \) represents the number of transit stops on transit route \( r_j \)

The above TN vector set \( T^{(l)} \) may also be expressed as a TN matrix

\[
T = T[ t_{ij} ], \quad t_{ij} = \begin{cases} 
1, & \text{if node } j \text{ is on route } i, \\
0, & \text{if node } j \text{ is not on route } i,
\end{cases} \quad i = 1, 2, ..., l, \quad j = 1, 2, ..., n
\]

In this study, for the purpose of representation uniqueness, it is assumed that transit stops coincide with street nodes. Vehicle headways of a TN system may expressed in a vector form

\[
h = (h_1, h_2, ..., h_l)
\]

where:

\( h_q \) \((q = 1, 2, ..., l)\) is the vehicle headway of route \( q \)

Using relationship (3), a headway vector may be derived from vehicle assignment vector
\[ v = (v_1, v_2, \ldots, v_l) \]

where:

\[ v_q \ (q = 1, 2, \ldots, l) \]
is the number of vehicles operating on route \( q \)

Transit demand is given by an OD matrix

\[ O = [o_{ij}] \]

where:

\[ o_{ij} \]
is the number of trips originating from node \( i \) and destined for node \( j \)

**Representation of Search Spaces for Transit Routes and Transit Network**

The solution search spaces are locally and iteratively defined, and the size of a local search space may be flexible depending on available computing resources. A local path space is defined by a master path, a key-node representation of the master path, and a set of paths that are in the neighborhood of the master path. A master path is a path from which a local path space is generated. Key-nodes are a set of nodes on a master path based on which paths in the local path space are generated. Figure 1 illustrates a master path (solid line) and its three key-nodes \( n_1, n_2, \) and \( n_3 \).

**Figure 1. Three-Key-Node Representation of a Transit Route**
Based on these key-nodes, local node spaces may be defined and a local path space is derived from the local node spaces. An $i$th order local node space of a master node is defined as the set of nodes that may be connected to the master node with $i$ or fewer street segments. As an example, in Figure 1, the first order node space of the key-node $n_1$ is comprised of the master node $n_1$ itself (black circle) and its immediately adjacent nodes ($n_{11}$, $n_{12}$, $n_{13}$, and $n_{14}$). The second order node space of the key-node $n_3$ includes all the nodes that may be connected to the master node $n_3$ with two or fewer street segments (i.e., all the black, white, and gray circles around node $n_3$). Clearly, a local node space is a subspace of the street node space $N(n)$. As the order $i$ increases, a local node space will approach the original street node space $N(n)$. Therefore, the order of a local node space provides a measurement of the degree of localization.

The procedure to generate a local path space from a master path has three steps:

1. Select $s$ nodes from the node set of the master path $p = p(n_1, n_2, ..., n_r)$ as the key-nodes.
2. Generate a sequence of $s$ local node spaces of a given order from these $s$ key-nodes.
3. Define the local path space as the set of paths consisting of piecewise shortest path segments that start from each node in the first local node space, sequentially pass through one node in each of the intermediate local node spaces, and end at each node in the last local node space.

The local network search spaces of a transit network $T^{(i)} = \{r_1, r_2, ..., r_l\}$ is defined as the set of all local path spaces on the local node spaces (of a given order) of all the routes. In general, a route derived from a smaller number of key-nodes will result in better route directness and a smaller local path search space, but its flexibility is also limited. A route with a larger number of key-nodes is relatively more flexible to reach more neighboring nodes and therefore may cover more trips. However, it is also associated with a larger local path search space thus requiring more computing resources.

**Simple Constraints for Transit Route Network**

It is well known that for a discrete system, identifying and incorporating as many appropriate constraints as possible will significantly reduce the size of the search space. The following constraints are applied in this study:
• Maximum in-vehicle travel time (or route length) constraints for individual transit routes. It is known that a lengthy transit route not only results in difficulty in maintaining schedule, but also presents a safety hazard due to possible driver fatigue.

• Constraint on total transit vehicle fleet size. Since total vehicle fleet size of a transit network is closely related to operator cost, this may be considered an operator cost constraint. The optimization problem may be stated as finding a route network and route headways that result in the optimal service coverage and minimum user cost for a given fixed-operator budget (reflected by a given fleet size).

• Minimum and maximum headway constraints on individual routes. Route headways should be neither too small for operational reasons nor too large to result in long waiting times; the latter adversely affect ridership. Determination of headway is also constrained by vehicle load factor, which limits the number of passengers in a transit vehicle to ensure passenger comfort.

**Route Directness Constraints**

Route directness is defined as

\[
d(r) = \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} w_{ij} \left( \frac{u_{ij}^{(S)}}{u_{ij}^{(R)}} \right)
\]

where:

- \( q \) is the number of nodes on route \( r = r(n_1, n_2, ..., n_q) \)
- \( u_{ij}^{(R)} \) signifies the user travel cost between nodes \( n_i \) and \( n_j \) along the route
- \( u_{ij}^{(S)} \) is the user travel cost along the shortest path of the street network between these two nodes
- \( w_{ij} \) is a weighting factor
For geometry-based route directness, $w_{ij} = 2/(q^2-1)$, and for ridership-based route directness

$$w_{ij} = \left(\frac{o_{ij} + o_{ji}}{2}\right)^{g-1} \sum_{i=1}^{g} \sum_{j=i+1}^{g} (o_{ij} + o_{ji})$$

where:

- $o_{ij}$ and $o_{ji}$ are coefficients of the demand matrix $O$

In general, a larger route directness value implies better route directness but may result in higher transit operating cost, while a smaller route directness value may mean possible loss of ridership and, in turn, higher operating cost. A description of the physical meaning of route directness is found in Zhao and Gan (2003).

**Network Directness Constraints**

The meaning of network directness is similar to that of the route directness except that the directness measurement is based on the geometry or ridership characteristics of the entire network instead of individual routes.

**Optimization Objective Functions**

The objective function in this study is the total user travel cost, which is the summation of the user travel times of all the trips between the OD pairs in the transit service area

$$U(T, O, h) = \sum_{i,j=1, i \neq j}^{n} \left[ o_{ij} \cdot U_{ij}(T, h) \right]$$

where:

- $o_{ij}$ is the number of trips originating from node $i$ and destined for node $j$
- $h$ represents a headway vector
- $U_{ij}(T, h)$ is the user travel time of one trip between nodes $i$ and $j$ in TN $T$
For zero-, one-, and two-transfer trips, user travel time may be expressed as

\[ u_{ij}^{(0)} = t_{\text{wait}}^{(r)} + t_{\text{invh}}^{(r)} \]  
\[ u_{ij}^{(1)} = t_{\text{wait}}^{(r1)} + t_{\text{invh}}^{(r1)} + t_{\text{wait}}^{(r2)} + t_{\text{invh}}^{(r2)} + t_{\text{Penl}}^{(r1r2)} \]  
\[ u_{ij}^{(2)} = t_{\text{wait}}^{(r1)} + t_{\text{invh}}^{(r1)} + t_{\text{wait}}^{(r2)} + t_{\text{invh}}^{(r2)} + t_{\text{Penl}}^{(r1r2)} + t_{\text{wait}}^{(r3)} + t_{\text{invh}}^{(r3)} + t_{\text{Penl}}^{(r2r3)} \]  

where:

\[ u_{ij}^{(k)} (k = 0, 1, 2) \] represents the travel time of a \( k \)-transfer trip between a demand node pair \( i \) and \( j \)

\[ t_{\text{wait}}^{(q)} \text{ and } t_{\text{invh}}^{(q)} \] are, respectively, waiting time and in-vehicle travel time on the transit route \( q \) (\( q = r, r1, r2, r3 \))

In equations (9) and (10), \( t_{\text{Penl}}^{(r1r2)} \) is the penalty for transfers between routes \( r1 \) and \( r2 \) expressed in equivalent in-vehicle travel time, while \( t_{\text{Penl}}^{(r2r3)} \) in equation (10) is the penalty for transfers between routes \( r2 \) and \( r3 \). All the travel time components in expressions (8), (9), and (10) are functions of TN matrix \( T \) and/or route headway vector \( h \). A detailed description of the above travel time components can be found in, for example, Shih and Mahmassani (1994). For uncovered trips, including those that require too many transfers and thus are unlikely to occur, the corresponding travel times are represented by a fictitious travel cost penalty value. The penalty cost associated with trips involving \( k \) transfers may be chosen from the following:

\[ u_p^{(0)} = t_{\text{wait}}^{(\text{max})} + t_{\text{invh}}^{(\text{max})} \]  
\[ u_p^{(1)} = 2 t_{\text{wait}}^{(\text{max})} + 2 t_{\text{invh}}^{(\text{max})} + t_{\text{Penl}}^{(\text{max})} \]  
\[ u_p^{(2)} = 3 t_{\text{wait}}^{(\text{max})} + 3 t_{\text{invh}}^{(\text{max})} + 2 t_{\text{Penl}}^{(\text{max})} \]
where:

$$u_p^{(k)} (k = 0, 1, 2, 3)$$ represents the maximum possible cost of a $k$-transfer trip

$$t_{\text{wait}}^{(\text{max})}, t_{\text{invh}}^{(\text{max})}, \text{and } t_{\text{penl}}^{(\text{max})}$$ represent, respectively, the maximum possible waiting, in-vehicle travel, and transfer penalty times for all demand trips in the service area.

Based on the above definitions and notations, the total user-cost objective function may be expressed as

$$U^{(k)}(T, O, h) = \sum_{i,j=1, i\neq j}^n \left[ o_{ij} \cdot U_{ij}^{(k)}(T, h) \right]$$

where:

$$U^{(k)} (k = 0, 1, 2)$$ is the total user-cost function based on $k$-or-less transfer trips, and the corresponding travel time is defined as

$$U_{ij}^{(q)}(T, h) = \begin{cases} u_{ij}^{(0)} & \text{if demand } o_{ij} \text{ is satisfied with a zero-transfer trip} \\ u_{ij}^{(q)} & \text{otherwise (} q = 0, 1, 2, 3 \text{)} \end{cases}$$

$$U_{ij}^{(1)}(T, h) = \begin{cases} u_{ij}^{(0)} & \text{if demand } o_{ij} \text{ is satisfied with a zero-transfer trip} \\ u_{ij}^{(1)} & \text{if demand } o_{ij} \text{ is satisfied with a one-transfer trip} \\ u_{ij}^{(q)} & \text{otherwise (} q = 0, 1, 2, 3 \text{)} \end{cases}$$

$$U_{ij}^{(2)}(T, h) = \begin{cases} u_{ij}^{(0)} & \text{if demand } o_{ij} \text{ is satisfied with a zero-transfer trip} \\ u_{ij}^{(1)} & \text{if demand } o_{ij} \text{ is satisfied with a one-transfer trip} \\ u_{ij}^{(2)} & \text{if demand } o_{ij} \text{ is satisfied with a two-transfer trip} \\ u_{ij}^{(q)} & \text{otherwise (} q = 0, 1, 2, 3 \text{)} \end{cases}$$
It may be seen from the structure of the total user-cost function $U^{(k)}$ defined in (15) through (18) that minimization of this function has two effects: minimizing the total user cost based on $k$-or-less transfer trip coverage, and maximizing $k$-or-less transfer trip coverage. The penalty term $u^{(q)}_p$ regulates the balance between these two effects. In general, the larger the penalty value, the greater relative importance is given to service coverage. Based on the above descriptions, a TN design optimization problem may be stated as follows:

Minimize:

$$U^{(k)}(T, O, h) = \sum_{i,j=1, i \neq j}^n o_{ij} \cdot U^{(k)}_{ij}(T, h)$$

Subject to:

- Route length constraints: $R^{(q)}_{\text{min}} \leq R^{(q)}_L \leq R^{(q)}_{\text{max}}, \ (q = 1, 2, ..., l)$

- Route headway constraints: $h^{(q)}_{\text{min}} \leq h_q \leq h^{(q)}_{\text{max}}, \ (q = 1, 2, ..., l)$

- Route load factor constraints: $\frac{h_q \cdot q^{(q)}_{\text{max}}}{V^{(q)}_{\text{Seat}}} \leq L^{(q)}_{\text{max}}, \ (q = 1, 2, ..., l)$

- Route directness constraints: $d_q \geq d^{(q)}_{\text{min}}, \ (q = 1, 2, ..., l)$

- Total TN fleet size constraints: $\sum_{q=1}^l R^{(q)}_{\text{Fleet}} \leq N^{(T)}_{\text{Fleet}}$

where:

- $R^{(q)}_{\text{min}}$ and $R^{(q)}_{\text{max}}$ represent the minimum and maximum route length constraints for route $q$

- $h^{(q)}_{\text{min}}$ and $h^{(q)}_{\text{max}}$ are the minimum and maximum route headway constraints for route $q$

- $d^{(q)}_{\text{min}}$ is the minimum route directness constraint for route $q$

- $N^{(T)}_{\text{Fleet}}$ is the system fleet size
Simulated Annealing Algorithm Search for Optimal Transit Network

The simulated annealing (SA) search scheme is a stochastic process designed to avoid being trapped into poor local optima. Under fairly general conditions and for large problems, it may always be expected to find a global solution faster than a random search method. In an SA search, a solution and its associated local space will replace the incumbents with probability 1 if it has a better goal value or with some probability between 0 and 1 otherwise. The probability to accept a worse solution is proportional to the difference between its goal value and the current best goal value. A slightly worse solution has a higher probability of being accepted than a much worse solution. In the long run, as the number of search iterations becomes sufficiently large, the search process may escape from any local optimum, and eventually should visit a global optimal solution (Hajek 1988). For a TN design optimization problem, the simulated annealing procedure involves the selection of a solution candidate TN $T$ from a local network space based on a given initial TN $T_0$. The network $T$ is accepted if the associated objective function $U(T, O, h) < U(T_0, O, h)$. Otherwise, $T$ is accepted with a probability

$$ p = \exp(-\Delta/t), \Delta = [U(T, O, h) - U(T_0, O, h)] $$

25

where:

$t (t > 0)$ is the temperature of the annealing process

$U$ is the total user-cost function defined in (15)

The term “temperature” is borrowed from annealing in solids, and has no physical meanings. The value of a “temperature” $t$ in equation (25) merely reflects the fact that the likelihood of accepting solutions with worse goal values is regulated by $t$. A larger $t$ results in a larger probability of accepting worse solutions, while a smaller $t$ reduces such chances. In practice, a large initial $t$ is often chosen to increase the chance of escaping from a poor local minimum, which is gradually reduced in the search process by a factor $\tau$ to enhance the selectivity of the search toward improved solutions.
One difficulty with the SA search method described above is that the search process may repeat the same solutions or a sequence of solutions many times before moving to other search regions. To alleviate this problem, a tabu list is established to keep track of solutions evaluated recently to prevent them from entering the solution search process again. A detailed description of the simulated annealing search process may be found in Zhao and Gan (2003).

**Fast Descent Method Search for Optimal Transit Headways**

In the SA search process, the TN headway vector \( h \) in equation (25) remains a passive network parameter (i.e., \( h \) does not play an active role in the SA search process). The TN headway vector \( h \) will be modified only if its vector components violate their associated constraints, such as (21), (22), etc. This is due to the fact that for large-scale TN problems, simultaneous search for both optimal route networks and headways may be computationally intractable due to the huge solution population size to expose any meaningful characteristics (e.g., promising search directions, local minima, etc.). The search for a better headway \( h \) is performed in a separate process called the fast descent (FD) search. The basic idea behind this method is to find the vector components (or directions) of a cost function such that appropriate adjustments of these components will lead to the fast descent of the cost function. The procedure for applying a FD search method to search for optimal route headways is outlined as follows:

1. If the maximum fleet size constraint is not violated, find the route (the FD component) in the network for which a decrease in its headway (as a result of increasing the number of vehicles operating on this route) will result in the FD of the total user-cost function. Update existing route headway with the new headway. Repeat this step until the maximum fleet size constraint is violated. Go to step 2.

2. Find two routes in the route network for which decreasing the headway for one route (by increasing the number of vehicles on the route) and increasing the headway for the other route (by reducing an equal number of vehicles on this route) will result in the FD of the total user-cost function. Update existing route headways with the new headways. Repeat this step until a (local) minimal cost value is reached. Start a new round of SA search iteration for a better transit route network layout.
In this study, the SA and FD search processes are integrated together in an iterative manner (i.e., an FD search process for better route headways will be executed after one or more SA search iterations for better route network layouts).

**Numerical Experiments**

The first experiment was based on a real network in Switzerland (Mandl 1979). This problem was also used by Shih and Mahmassani (1994) and Baaj and Mahmassani (1991) as a benchmark to test their approaches to TN design optimization. Mandl’s problem consisted of a street network of 15 nodes with a total demand of 15,570 trips per day. The length of a street segment was defined in terms of in-vehicle travel time in minutes. The maximum route length was constrained to 40 minutes. In Table 2, the first row identifies the source of the solutions to the benchmark problem, which include Mandl’s problem as well as its variations constructed by Baaj and Mahmassani (1991) and Shih and Mahmassani (1994). The second row identifies solutions to the benchmark problem. The problems/solutions differ in their number of routes, total fleet size, and/or the search method used. Mandl (1979), Baaj and Mahmassani (1991), and Shih and Mahmassani (1994) all used a transfer penalty of 5 minutes, which was also assumed in the proposed SAFD method. The methods used to obtain the results are indicated in the third row. For each solution, the unshaded column provides the statistics for the layouts produced in the original studies, and the shaded column gives the statistics for the results produced from the SAFD method developed in this study. All the SAFD results in this table were generated with the total user-cost objective function $U^{(1)}$ defined in equation (15). The operator cost is reflected by the TN fleet sizes shown in the 10th row. It may be seen that the percentages of zero-transfer trips are higher for all solutions produced from this study. Except for Mandl’s original results, all solutions provided 100 percent trip coverage with trips requiring zero or one (one-or-less) transfer. The savings in total user travel time in hours from the SAFD method are shown in the 8th row.

The second experiment involved a large-scale TN optimization problem based on the service area of the Miami-Dade Transit Agency (MDTA), which encompasses a region of about 300 square miles with a population of about 2.3 million. MDTA ranks as the 16th largest transit agency in the United States. At the time of this research, MDTA operated 83 transit routes, including a rail rapid transit system of 22.5 route miles (Metrorail), a 4.5-mile downtown automated circulation system (Metromover), and 81 bus routes with about 4,500 transit stops. The street net-
Table 2. Comparison of Results from Different Methods

<table>
<thead>
<tr>
<th>Problem Source</th>
<th>Mandl’s method</th>
<th>Baaj &amp; Mahmassani</th>
<th>Shih &amp; Mahmassani</th>
</tr>
</thead>
<tbody>
<tr>
<td>Route layout case</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Search method</td>
<td>Mandl SAFD</td>
<td>B&amp;M SAFD</td>
<td>B&amp;M SAFD</td>
</tr>
<tr>
<td>0-transfer trips (%)</td>
<td>69.94 95.31</td>
<td>78.61 95.18</td>
<td>79.96 95.44</td>
</tr>
<tr>
<td>1-transfer trips (%)</td>
<td>29.93 4.69</td>
<td>21.39 4.82</td>
<td>20.04 4.56</td>
</tr>
<tr>
<td>2-transfer trips (%)</td>
<td>0.13 0.00</td>
<td>0.00 0.00</td>
<td>0.00 0.00</td>
</tr>
<tr>
<td>Total user travel cost (min)</td>
<td>219094 185158</td>
<td>205646 190998</td>
<td>203936 189460</td>
</tr>
<tr>
<td>Savings in total user travel cost (hr)</td>
<td>- 565.60</td>
<td>- 244.13</td>
<td>- 457.93</td>
</tr>
<tr>
<td>Travel directness</td>
<td>1.335 1.128</td>
<td>1.253 1.115</td>
<td>1.141 1.141</td>
</tr>
<tr>
<td>Total fleet sizes</td>
<td>99 99</td>
<td>89 89</td>
<td>77 77</td>
</tr>
<tr>
<td>Number of routes</td>
<td>4 4</td>
<td>6 6</td>
<td>8 8</td>
</tr>
<tr>
<td>Average transfers</td>
<td>1.302 1.047</td>
<td>1.214 1.048</td>
<td>1.046 1.190</td>
</tr>
<tr>
<td>Network directness</td>
<td>0.811 0.970</td>
<td>0.912 0.959</td>
<td>0.954 0.856</td>
</tr>
</tbody>
</table>

1Mandl’s method
2Simulated annealing and FD method
3Baaj and Mahmassani’s method
4Shih and Mahmassani’s method

Table 3. Comparison of SAFD Search Results with the Existing Network

<table>
<thead>
<tr>
<th>Network Parameters</th>
<th>Existing Network</th>
<th>User-Cost Objective Function $U^\theta$</th>
<th>User-Cost Objective Function $U^t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max. headway cases (min)</td>
<td>$h_{\text{max}} = 20$</td>
<td>$h_{\text{max}} = 20$</td>
<td>$h_{\text{max}} = 30$</td>
</tr>
<tr>
<td>Transfer penalty time (min)</td>
<td>5 10</td>
<td>5 10</td>
<td>5 10</td>
</tr>
<tr>
<td>One-or-less transfer trips (%)</td>
<td>55.17 55.17</td>
<td>82.35 82.35</td>
<td>82.48 82.48</td>
</tr>
<tr>
<td>Two-or-less transfer trips (%)</td>
<td>65.23 65.23</td>
<td>86.48 86.48</td>
<td>86.60 86.60</td>
</tr>
<tr>
<td>Total covered trips (%)</td>
<td>65.69 65.69</td>
<td>86.51 86.51</td>
<td>86.64 86.64</td>
</tr>
<tr>
<td>Average user cost (minutes)</td>
<td>42.8 47.1</td>
<td>39.7 42.5</td>
<td>39.8 42.6</td>
</tr>
<tr>
<td>Total fleet sizes</td>
<td>600 600</td>
<td>600 600</td>
<td>600 600</td>
</tr>
<tr>
<td>Average transfers (two-or-less)</td>
<td>1.934 1.934</td>
<td>1.673 1.673</td>
<td>1.678 1.678</td>
</tr>
<tr>
<td>CPU time (hours)</td>
<td>- -</td>
<td>0.3270 0.266</td>
<td>0.264 0.223</td>
</tr>
</tbody>
</table>

124
work used in this experiment consisted of 4,300 street segments and 2,804 street nodes, and the longest bus routes were about 32 miles. Total length of the transit system was about 1,300 route miles, not including some small loops at the ends of some routes or in shopping centers.

The OD matrix was generated from the 1999 validated Miami-Dade travel demand model, which provided the daily number of passenger trips between each pair of traffic analysis zone centroids. These trips were manually distributed to the surrounding street network nodes with considerations given to land-use patterns, proximity, and street network connectivity. Total demand was 161,944 daily transit trips. They were distributed, sparsely and unevenly, between about 120,000 demand (OD) pairs.

Operator cost is reflected by the network fleet size of 600 transit vehicles, which is about the same as that operated by MDTA. The total number of transit lines in the example remains the same as the existing system, and Metrorail and Metromover alignments are fixed in the optimization process. Other constraints and data used in this example include: for bus routes, the minimum and maximum headways are $h_{\text{min}} = 4$ min and $h_{\text{max}} = 20$ min; for Metrorail, $h_{\text{min}} = 6$ min and $h_{\text{max}} = 10$ min; and for Metromover, $h_{\text{min}} = 1$ min and $h_{\text{max}} = 6$ min. The minimum and maximum route lengths (in-vehicle time) are $R_{\text{min}} = 10$ min and $R_{\text{max}} = 90$ min; average in-vehicle travel speeds are 21 mph for bus and 31 mph for Metrorail. Since no data were available regarding the appropriate value of transfer penalty time in Miami-Dade County, results obtained from two sets of transfer penalty values are presented. One is $t_{\text{pen}} = 5$ minutes, which is the same as that used in the first example. The other is $t_{\text{pen}} = 10$ minutes. Ideally, the value of transfer penalty value should be determined through transit user surveys since transfer penalty reflects transit users’ tolerance to experiences of unpleasants or inconvenience during vehicle transfers. Penalty values are likely to vary across different geographic areas and change with demographics, socioeconomics, topography, climate, quality of transfer facilities, etc.

All the numerical results were obtained on a personal computer with a 2.8GHz CPU and 1GB RAM memory. Table 3 presents the results from the proposed SAFD method. There are two sets of results, one based on the zero-transfer total user-cost function $U^{(0)}$ and the other based on the one-or-less transfer total user-cost function $U^{(1)}$. For references, TN parameters for the existing network are also included. The first row in the table identifies the objective functions used to generate the corresponding numerical results. The second row provides the maximum
route headway constraints for different test cases, while the third row indicates the transfer penalty values for various test cases. For various objective functions, headway constraint, and transfer penalty combinations, statistics for the TNs obtained from the optimization process are provided. Results in Table 3 show that, user cost appears to be more sensitive to the value of transfer penalty time than to other TN parameters. This implies that an accurate transfer penalty value is needed to obtain a good estimate of user cost and improvement in transfer facilities and that transfer quality will help reduce the transfer penalty.

From Table 3, it may be seen that results obtained from this study have significant improvement over the existing one. The zero-transfer trips increased from 14.38 percent based on the existing network to 32.37 percent with objective function $U^{(0)}$ and $h_{\text{max}} = 20$, an improvement of about 125 percent. The one-or-less transfer trip coverage increased from 55.17 percent to 86.41 percent with objective function $U^{(1)}$, $h_{\text{max}} = 30$, and $t_{\text{Pen}} = 5\text{min}$, an improvement of about 57 percent. Assuming most transit riders are only willing to transfer once per trip, the one-or-less trip coverage shown in the fifth row would be the actual total service coverage of the corresponding TN. The remaining trips either require two or more transfers or are not satisfied. The second from the last row in Table 3 presents the average transfers per trip for the two-or-less transfer trips that involve the use of Metrorail and Metromover. The high level of service of the rail lines is more likely to encourage people to use transit even if the trips may require two transfers. The eighth row in Table 3 shows the per user cost based on two-or-less transfer trips. As expected, larger transfer penalty time values result in higher user costs.

The modeled network is not a prefect description of the actual network even though care has been taken to prepare the input data as accurately and completely as possible. Consequently, some of the differences in the statistics from the exiting network and the results generated from the SAFD may be attributed to the inaccuracy in the modeled network. However, the main purpose of the second example is not to show the superiority of the SAFD results over the existing network, but rather to demonstrate that for a large-scale transit network with a given transit demand pattern as well as a constraint set, the proposed method is able to improve the initial network configuration in a reasonable amount of time.
Conclusions
A mathematical stochastic method for large-scale TN optimization has been described. A stochastic local search method based on simulated annealing and fast descent search has been developed and has been shown to be capable of tackling large-scale transit network design optimization problems and producing results in a reasonable amount of time. The solution methodology is generally applicable to a wide range of practical TN problems, does not favor any particular transit network configurations, and gives reasonably good results in a reasonable amount of time. The methodology also allows results to improve and approach the global optimum as the computer resource or power increases.

Future improvements to the solution method may include, for example, the development of time-dependent TN optimization methods to optimize a TN by taking into consideration changes in network conditions and OD distribution during different time periods of a day, inclusion of terminal access times in the user costs, and removal of simplifying assumptions regarding transit fare and fixed demand.

Acknowledgements
The author would like to thank the reviewers for their careful reading of this paper and for their helpful comments and suggestions.

References


Fan, W., and R. B. Machemehl. 2004. Optimal transit route network design problem: algorithms, implementations, and numerical results, Report No. SWUTC/04/167244-1, Austin, TX: Center for Transportation Research, University of Texas at Austin.


Shih, M. C., and H. S. Mahmassani. 1994. *A design methodology for bus transit networks with coordinated operations*. SWUTC/94/60016-1. Austin, TX: Center for Transportation, University of Texas at Austin.


About the Author

**FANG ZHAO** (*zhaof@fiu.edu*) received her Ph.D. in civil engineering from Carnegie Mellon University in 1991. Since 1992 she has been teaching in the Civil and Environmental Engineering Department at Florida International University. She is currently an associate professor and deputy director of the Lehman Center for Transportation Research. Her main research interests include public transportation, travel demand modeling, geographic information systems, and transportation data modeling. Dr. Zhao is a registered professional engineer in the State of Florida and serves on both the TRB Transit Planning and Development Committee and the New Technologies and Systems Committee.
SUBSCRIPTIONS

Complimentary subscriptions can be obtained by contacting:

Lisa Maitland, Assistant to the Editor
Center for Urban Transportation Research (CUTR)
University of South Florida
Fax: 813-974-5168
Email: jpt@cutr.usf.edu
Web: www.nctr.usf.edu/journal.htm

SUBMISSION OF MANUSCRIPTS

The Journal of Public Transportation is a quarterly, international journal containing original research and case studies associated with various forms of public transportation and related transportation and policy issues. Topics are approached from a variety of academic disciplines, including economics, engineering, planning, and others, and include policy, methodological, technological, and financial aspects. Emphasis is placed on the identification of innovative solutions to transportation problems.

All articles should be approximately 4,000 words in length (18-20 double-spaced pages). Manuscripts not submitted according to the journal’s style will be returned. Submission of the manuscript implies commitment to publish in the journal. Papers previously published or under review by other journals are unacceptable. All articles are subject to peer review. Factors considered in review include validity and significance of information, substantive contribution to the field of public transportation, and clarity and quality of presentation. Copyright is retained by the publisher, and, upon acceptance, contributions will be subject to editorial amendment. Authors will be provided with proofs for approval prior to publication.

All manuscripts must be submitted electronically, double-spaced in Word file format, containing only text and tables. If not created in Word, each table must be submitted separately in Excel format and all charts and graphs must be in Excel format. All supporting illustrations and photographs must be submitted separately in an image file format, (i.e., TIF, JPG, AI or EPS), having a minimum 300 dpi, and include all sources and written permissions. Each chart and table must have a title and each figure must have a caption.

All manuscripts should include sections in the following order, as specified:

- Cover Page - title (12 words or less) and complete contact information for all authors
- First Page of manuscript - title and abstract (up to 150 words)
- Main Body - organized under section headings
- References - Chicago Manual of Style, author-date format
- Biographical Sketch - of each author

Be sure to include the author’s complete contact information, including email address, mailing address, telephone, and fax number. Submit manuscripts to the Assistant to the Editor, as indicated above.