

Validating T-BEST Models with 100% APC Counts*

*Center for Urban
Transportation Research
University of South Florida, Tampa
February 2007*

Contract Number: BD549-19

*This was the original project title that was motivated by the desire to find a transit agency with the highest penetration rate of automated passenger counters (APC) for boarding data.

DISCLAIMER

The opinions, findings, and conclusions expressed in this publication are those of the authors and not necessarily those of the State of Florida Department of Transportation.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. NCTR-576-07, BD549-19	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle Validating T-BEST Models with 100% APC Counts		5. Report Date February 2007	
7. Author(s) Xuehao Chu		6. Performing Organization Code	
9. Performing Organization Name and Address National Center for Transit Research (NCTR) University of South Florida 4202 E Fowler Ave., CUT100, Tampa, FL 33620-5375		8. Performing Organization Report No.	
12. Sponsoring Agency Name and Address Office of Research and Special Programs U.S. Department of Transportation, Washington, D.C. 20590 Florida Department of Transportation 605 Suwannee Street, MS 30, Tallahassee, FL 32399		10. Work Unit No.	
15. Supplementary Notes Supported by a grant from the Florida Department of Transportation and the U.S. Department of Transportation		11. Contract or Grant No. DTRS98-G-0032	
13. Type of Report and Period Covered		14. Sponsoring Agency Code	
16. Abstract <p>The current version of the Transit Boarding Estimation and Simulation Tool (T-BEST) uses a set of boarding equations that were estimated with a small sample of boarding data. In addition, it uses an overall weekday peak that combines the morning and afternoon peaks. This report documents a research effort in re-estimating a set of boarding equations for T-BEST to achieve two objectives. One objective was to use boarding data that are far more reliable than what were used in estimating the current set of boarding equations. The other objective was to make improvements in the boarding equations. Both objectives have been successfully achieved. The first objective has been achieved through using boarding data from a fleet of bus vehicles that had an APC penetration rate around 75 percent. The second objective has been achieved through many improvements in the boarding equations. Among these improvements is the separation of the current overall weekday peak into a morning peak and an afternoon peak. The re-estimated boarding equations perform well in terms of both how well the models fit the data and how observed and in-sample predictions compare. However, serious over-predictions can occur at a small number of stops, and these cases of over-prediction need to be dealt with individually as part of the validation process in each application of T-BEST.</p>			
17. Key Words Transit, Direct Boarding, Transfer Boarding, Accessibility, Stops, Routes, Bus Service, Service Planning		18. Distribution Statement Available to the public through the National Technical Information Service (NTIS), 5285 Port Royal Road, Springfield, VA 22161, 703-487-465, and through the NCTR website at http://www.nctr.usf.edu	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of pages 16 (without appendices)	22. Price

EXECUTIVE SUMMARY

Problem Statement and Objectives

The Public Transit Office of the Florida Department of Transportation (FDOT) in recent years has invested heavily in developing a comprehensive transit ridership forecasting model system for short-term service planning. The Transit Boarding Estimation and Simulation Tool (T-BEST) is the third generation of this effort. T-BEST makes several advances. It models and forecasts transit boarding at the individual stop level. It separates direct boarding from transfer boarding for both modeling and forecasting. It explicitly treats inter-relationships in a transit network through measures of accessibility to opportunities for potential activity participation. More important, these modeling advances provide a significant level of practical flexibility for transit service planning that has not been available before. T-BEST can be used to assess the boarding impacts of a variety of service changes, including operating strategies, schedule changes, alignment changes, system changes, and fare policy. The current user guide at www.tbest.gov provides detailed information about the modeling and forecasting framework of T-BEST as well as its flexibility for assessing the boarding impacts of service changes.

T-BEST's current boarding equations were estimated with a small sample of boarding data collected through automated passenger counters (APC) in the Jacksonville area of Florida. The small sample has greatly reduced the reliability of the current boarding equations in T-BEST. This small sample also has made it difficult to separate the morning and afternoon peaks on weekdays. The current research project was designed to re-estimate a set of boarding equations for the next version of T-BEST with two specific objectives. One objective was to use boarding data that are far more reliable than what were used for estimating the current set of boarding equations. The other objective was to make improvements in the boarding equations.

Findings and Conclusions

Both objectives have been successfully achieved. The first objective has been achieved through using boarding data from a fleet of bus vehicles that had an APC penetration rate around 75 percent at an agency that has had many years of experience in archiving and using APC data. The second objective has been achieved through many improvements in the boarding equations. These improvements include the following:

- Some improvements are structural: the current weekday peak period has been split into a weekday morning peak and a weekday afternoon peak.
- Some improvements are statistical: except for the afternoon peak period, boarding equations are estimated without the restrictive assumption that the mean and variance of model error terms are equal.
- Some improvements involve adding additional desirable variables: 1) the effect of park-n-ride lots on direct boardings has been taken into account for the weekday morning peak; and 2) the effect of daily service span on both direct and transfer boardings has been added for Saturdays and Sundays.
- Some improvements involve how certain variables enter the questions: 1) the socio-demographic characteristics of population enter the equations as quantity rather than as

shares; and 2) daily service span, service frequency, and accessibility to population and employment from boarding at a subject stop enter the equations in a log form.

- Some improvements involve how individual variables are computed, including the radius of stop buffers, how overlapping stop buffers are split to avoid double-counting, how some of the accessibility measures are computed, how components of impedance are weighted, the impedance threshold within which accessibility measures are computed, and the distance threshold within which people may transfer from one route to another.

The estimated boarding equations perform well in terms of both how well the equations fit the data and how observed and in-sample predictions compare. However, serious over-predictions can still occur at a small number of stops, and need to be dealt with individually as part of the validation process in each application of T-BEST.

Benefits

The new boarding equations are expected to be implemented into T-BEST 3.0. In addition to providing additional flexibility in evaluating the boarding impact of transit service changes, the new equations are expected to increase the reliability of these boarding evaluations. Transit agencies in Florida and nationally will benefit from the additional flexibility and reliability not only in developing state-required Transit Development Plans but also in other service planning activities.

TABLE OF CONTENTS

Disclaimer	ii
Technical Report Documentation Page	iii
Executive Summary	iv
Problem Statement and Objectives	iv
Findings and Conclusions	iv
Benefits	v
Table of Contents	vi
List of Tables and Figures	vii
Introduction	1
T-BEST Methodlogy	1
Current	1
Improvements	3
Adopted	3
Not Adopted	3
Model Structure	3
Network Relations	4
Direct Boarding	5
Transfer Boarding	6
Data	7
Agency Selection	7
Data Acquisition and Processing	7
Directly Derived Variables	8
T-BEST Generated Variables	9
model Estimation	10
Requirements	10
Statistical Requirements	11
Theoretical Requirements	11
Results	12
Application	14
Conclusions	15

LIST OF TABLES AND FIGURES

Table 1. Bus Stops with Zero Boardings by Time Period	11
Table 2. Results.....	13
Table 3. Observed and In-Sample Predictions of Direct Boardings during Morning Peak.....	15

INTRODUCTION

FDOT's Public Transit Office in recent years has invested heavily in developing a comprehensive transit demand forecasting model system. Transit Boarding Estimation and Simulation Tool (T-BEST) is the third generation of this effort. T-BEST can be used to assess the patronage impacts of a variety of service changes, including operating strategies, schedule changes, alignment changes, system changes, and fare policy. The current user guide at www.tbest.gov provides detailed information about the modeling and forecasting framework of T-BEST as well as its flexibility for assessing the patronage impacts of service changes.

T-BEST's prior transit demand models were estimated with boarding data from a small sample collected through automated passenger counters (APC) in the Jacksonville area of Florida. The small sample of boarding data has greatly reduced the reliability of the current demand models in T-BEST. This small sample also has made it difficult to separate the morning and afternoon peaks on weekdays. The value of T-BEST can be greatly enhanced with better data on boarding volumes.

This report documents a research effort to re-estimate T-BEST models with the following two objectives. One is to separate the morning and afternoon peaks on weekdays. The other is to use boarding data from a significantly greater APC sample. The rest of this report is organized into three sections. The data section describes data collection and processing efforts. The estimation section describes the modeling efforts. The last section concludes the report.

T-BEST METHODOLOGY

Current

The methodology underlying T-BEST has been developed to ensure that the final boarding equations are sensitive to a wide range of socio-economic and supply attributes. Based on the current user guide, the following features of T-BEST are particularly noteworthy:

1. *Forecasting Stop-Level Boardings:* T-BEST provides forecasts or predictions of stop-level boardings. Thus, ridership in the context of T-BEST is defined as the number of boardings at each stop that is specific to a direction and a route.
2. *Direct vs Transfer Boardings:* T-BEST incorporates separate equations for estimating and distinguishing between direct boardings and transfer boardings at each stop. At any given transit stop, one may have patrons who begin their trip at the designated stop and other patrons who are transferring from a different route in the middle of their trip/journey. By distinguishing between direct and transfer boardings, T-BEST is able to:
 - a. provide a quantitative perspective on the extent of trip linking that is occurring
 - b. provide a framework for analyzing the impacts of transfer points and transfer opportunities on ridership

3. *Time of Day Based Analysis*: T-BEST includes separate ridership estimation equations by time period within a week. The time periods that have been incorporated into the current version of T-BEST include:
 - a. Weekday peak period (covering both the AM and PM peaks)
 - b. Weekday off peak period
 - c. Weekday night period
 - d. Saturday (all day)
 - e. Sunday (all day)
4. *Spatial Accessibility (Socio-Economic Characteristics)*: T-BEST accounts for spatial accessibility in computing boardings at individual stops. Presumably, ridership is dependent on the number of people of various characteristics (defined by age, working status, race/ethnicity, income, car ownership, etc.) who can access the transit system. T-BEST considers circular buffer areas around individual stops to identify the market that has access to the transit system.
5. *Time-Space Network Connectivity*: In addition to considering spatial accessibility at the origin stop, one needs to consider the overall connectivity and time-space accessibility that a system provides to accurately compute ridership at any stop. People are more likely to use a transit system (stop) that is well connected and from which many destinations offering a range of activity opportunities can be reached. However, it is likely that riders will not be willing to tolerate trip lengths or durations and transfers beyond a certain threshold level. Thus, one needs to consider the activity opportunities (measured in terms of population and employment) that can be reached within a certain time frame and number of transfers when modeling the number of boardings at any stop. In addition, this network accessibility needs to be computed and accounted for along the temporal dimension. The network connectivity and range of reachable destinations may be different at different times of the day due to supply differences by time of day. T-BEST incorporates a powerful, comprehensive, and sophisticated methodology for accounting for time-space network connectivity and accessibility, thus making it the ideal tool for transit ridership forecasting.
6. *Competing and Complementary System Effects*: Within a transit system, there are bound to be competing and complementary system effects that affect ridership. For example, any stop is likely to have a series of neighboring stops that are competing for the same market/riders. If indeed, neighboring stops have overlapping market area buffers, then it is important to consider such competing effects in computing stop-level ridership. Similarly, there may also be complementary effects that affect and enhance ridership at a stop. For example, if a stop is a transfer point where two or more routes meet, then the number of boardings at the stop may be enhanced by virtue of the transfer opportunities present there. T-BEST explicitly accounts for both of these effects in computing stop-level ridership.
7. *GIS-Based Software Tool*: T-BEST has been developed so that the user can interface with the software largely through an interface that provides full GIS functionality. A user needs to have ArcView 8.3 or later residing locally on the machine to use T-BEST. A modest investment in ArcView 8.3 will allow the user to tap the full potential of T-BEST. Socio-economic scenarios, supply attributes, and route and stop configurations can be changed

and edited on the fly, thus making T-BEST a truly user-friendly transit ridership forecasting tool.

8. *Performance Measures:* T-BEST includes estimates of several performance measures in its output. Performance measures such as route miles, service miles, service hours, boardings per service mile or hour, and average boardings per service run are provided by T-BEST at the individual route-level and for the system as a whole. These performance measures can be used to assess the impacts of various socio-economic and supply scenarios on system performance.

Improvements

Working with FDOT and others involved in developing T-BEST, the research team identified and considered a number of potential improvements for the re-estimated boarding equations. Some of these were eventually adopted, but others were not.

Adopted

Three improvements were adopted in the new boarding equations:

- Separating the current weekday peak period into a weekday morning peak and a weekday afternoon peak.
- Considering the effect of park-n-ride lots on direct boardings in the weekday morning peak.
- Considering the effect of daily service span on both direct and transfer boardings on Saturdays and Sundays.

Not Adopted

The research identified two additional sets of potential improvements for T-BEST. One set is additional service attributes beyond frequency and daily service span, particularly service reliability at the stop level, at the route level, and at the origin-destination level. The importance of service reliability is beyond argument, particularly in circumstances where headways are relatively large. The other set is stop attributes, such as the presence of benches, shelters, etc. Neither of these has been adopted for the next version because few agencies have data on or even a standardized basis for measuring these attributes.

Model Structure

T-BEST models direct boardings separately from transfer boardings. For modeling, all stops are divided into those that provide transfer opportunities and those that do not provide transfer opportunities. For a given stop, transfer opportunities exist when at least one stop on a different route is located within walking distance of that given stop. Model estimation is done in two steps. In the first step, the model for direct boardings is estimated using data from stops without transfer

opportunities. In the second step, the estimated model for direct boardings is first applied to all stops to predict direct boardings. For those stops with transfer activities, the predicted direct boardings is subtracted from the observed total boardings, and the difference is used as the dependent variable for estimating the model for transfer boardings.

Network Relations

Inter-relationships within a transit network really occur at the stop level. At a given stop along a particular route, boarding is influenced by whether there are other stops, either along the same route or other routes, within walking distance, from which potential users can get to the same destinations or different destinations. These other stops are referred to as the neighboring stops of the subject stop. More important, boarding at this stop is influenced by the opportunities that can be reached by potential users from each of these neighboring stops. If a potential user could reach a movie theater from any neighboring stop but not from the subject stop, the chance that this user would board at the subject stop is minimal. If a potential user can reach a movie theater from the subject stop with less time than from all neighboring stops, the chance of the subject stop being used is high. The stops accessible from the neighboring stops are referred to as the accessible stops. Among other factors, accessibility to opportunities around these accessible stops for potential activity participation can be critical in modeling and forecasting patronage at the stop level.

Neighboring Stops

For a given stop (along a particular route in a particular direction), its neighboring stops are other stops within its buffer or whose buffers overlap with its buffer. These neighboring stops represent alternative points at which potential transit riders in the subject buffer may board a transit vehicle either on the subject route, in the subject direction of the subject route, or on other routes. The neighboring stops for a given subject stop fall into one of four groups: N_0 through N_3 .

- One set of neighboring stops are those on the same route and in the same direction as the subject stop. Some of these may be upstream of and some downstream of the subject stop. For either upstream or downstream, there may be multiple stops, depending on the density of stops in the subject direction along the subject route. While all of these potential neighboring stops can influence boarding at the origin stop, only the closest downstream stop is to be included in N_1 .
- The second set of neighboring stops are those along the same route but in the opposite direction. There may be multiple of these potential neighboring stops. For actual measurement, however, only one is required. When there are multiple stops, the one closest to the subject stop is to be chosen as the N_2 neighboring stop.
- The N_3 neighboring stops are those along other routes that are located within the subject buffer or within buffers that overlap the subject buffer. In any direction along any of these other routes, there may be multiple potential N_3 neighboring stops. Again for computationally purposes, only one such stop from each combination of direction and route is to be included in N_3 . If two other routes intersect the subject route at the subject stop, for

example, N_3 would have four stops in most cases. It may have fewer than four if one or both of these intersecting routes are one-way.

- The last set of neighboring stops, N_0 , is a subset of N_3 . They are neighboring stops on other routes and are located within the subject buffer. The reason to exclude those N_3 neighboring stops located outside the subject buffer is that people that alight at them would need to walk more than the radius of a buffer to transfer at the subject buffer.

Accessible Stops

With the four sets of neighboring stops determined, five sets of accessible stops are defined: S_0 through S_4 . Assume that stop s serves direction d along route r .

- Set S_0 includes stops that can reach any of the N_0 neighboring stops on other routes that are located within the subject buffer. The purpose of S_0 is to capture passengers riding toward stop s through other routes. That is, S_0 represents feeders for potential transfer boarding at stop s . S_0 is used later to measure the transfer potential for stop s . This transfer potential will be used in modeling transfer boarding but not in modeling direct boarding.
- S_1 includes stops downstream of stop s that can be reached from stop s through route r via the transit network. The purpose of S_1 is to capture the opportunities for potential activity participation that are accessible for a potential user who boards at stop s or its N_1 neighboring stops.
- Set S_2 includes stops in the network upstream of stop s through route r that can be reached from the N_2 neighboring stop. S_2 captures the opportunities for potential activity participation in the opposite direction of traveling at stop s through the same route as boarding at stop s .
- Set S_3 includes stops that can be reached from any of the N_3 neighboring stops. S_3 captures the opportunities for potential activity participation along other routes for people in the origin buffer. These three sets of accessible stops are used later to measure the accessibility to these opportunities for potential users in the stop s buffer.
- Set S_4 includes stops in S_3 that overlap stops in S_1 . That is, people in the origin buffer can access some of the opportunities around each of the S_4 stops from boarding at the origin stop or at any of the N_3 neighboring stops. Overlapping stops refers to stops where the buffers overlap.

Direct Boarding

Direct boarding for a given stop s and time period n is hypothesized to have the following equation:

$$D_n^s = g\left(R_n^s, C^s, A_{1n}^s, A_{2n}^s, A_{3n}^s, A_{4n}^s, X_n^s\right), \quad n = 1, \dots, N \quad (1)$$

where

- s = index for any origin stop.
- n = index for any time period.
- N = number of time periods.
- D_n^s = direct boardings at stop s during period n for the direction and along the route that define stop s .
- R_n^s = number of bus runs (frequency) departing at stop s during period n for the direction and along the route that define stop s .
- C^s = vector of buffer characteristics for stop s . These characteristics include the amount of population and employment as well as their characteristics.
- A_{1n}^s = vector of accessibility to employment and population in the buffer areas of S_1 stops during period n .
- A_{2n}^s = vector of accessibility to employment and population in the buffer areas of S_2 stops during period n .
- A_{3n}^s = vector of accessibility to employment and population in the buffer areas of S_3 stops during period n .
- A_{4n}^s = vector of accessibility to employment and population in the overlapped buffer areas S_3 stops and S_1 stops during period n .
- X_n^s = vector of other stop and route characteristics during period n .

Transfer Boarding

Transfer boarding for a given stop s and time period n has the following equation:

$$T_n^s = g\left(R_n^s, P_{0n}^s, A_{1n}^s, A_{2n}^s, A_{3n}^s, A_{4n}^s, Y_n^s\right), \quad n = 1, \dots, N \quad (2)$$

where

- T_n^s = transfer boardings at stop s during period n for the direction and along the route that define stop s .
- P_{0n}^s = transfer potential from upstream boarding at S_0 stops toward stop s during period n .
- Y_n^s = vector of other stop and route characteristics for period n .

The amount of population and employment and their characteristics in the buffer of a subject stop are not directly relevant to transferring users. As a result, related variables are now replaced by the variable measuring transfer potential. It is possible that transit users may want to avoid

transferring in buffer areas with certain characteristics, particularly in certain time periods. One good example is crime occurrence at night. Data on such characteristics are rarely available, however. The vector of other stop and route characteristics in these equations may differ from those in the equations for direct boardings because some of these are irrelevant to transferring users. A good example is the presence of special generators.

DATA

Agency Selection

Selecting a transit agency for data collection is a critical job once the data requirements for developing the model are identified. The primary focus was to select such an agency that has APC's installed on the majority of their fleet of vehicles. Also of importance were the duration of archived data available, reliability of the collected data, and the size of the fleet operated by the agency.

The starting point was the Federal Transit Administration's year 2000 survey of transit agencies on their current and planned use of advanced public transit technologies including APCs. A shortlist of seven candidate agencies was then created based on the criteria that the number of vehicles in the fleet should be 100 or more and more than 50% of the fleet should be installed with APCs. All these seven transit agencies were then contacted with detailed information about the type and the quality of data desired for the project.

Out of these candidate transit agencies, TriMet in Portland, Oregon stood out as one of the richest source of APC data. TriMet operates buses, light rail as well as streetcar, thus providing the flexibility of adding the multi-modal dimension in the boarding equations. Along with the APCs, Tri-Met also has on-board Automatic Vehicle Location (AVL) systems on their vehicles, making it a reliable source of stop-level boarding data properly indexed to the route, direction, and the geographical location of the stop. TriMet operates a large fleet of vehicles consisting of over 650 buses, 100 light rail cars, and 7 Streetcars. TriMet services cover around 575 square miles of the urban portion of the tri-county area (Clackamas County, Multnomah County, and Washington County) in Portland, Oregon with a total population of 1.3 million in year 2000. TriMet carries more than 300,000 riders on a typical weekday. These highlighting features of TriMet made it attractive as a source for data required for generating reliable and representative model equations. The major advantage of selecting TriMet is that they have APCs installed on approximately 75% of the bus fleet and approximately 25% of their light rail cars. The level, the extent, and the quality of data collected and maintained by TriMet is worth a mention here. TriMet has a large group of staff dedicated for data storage and maintenance.

Data Acquisition and Processing

A data request was sent to TriMet to match up the requirements with the data available for use. It was made known by the agency that the APC/AVL data are kept live for a period of sixth months before being archived and removed from the system. TriMet has a policy of having major service changes at nearly three months apart while having minor changes at any times it is deemed necessary. These time periods are called booking periods. At the time the data request was sent,

the two latest booking periods for which data were available were: March 06, 2005 to June 04, 2005 and June 05, 2005 to September 03, 2005. Having stop-level boarding data available for a period of six months is a vantage point for the present analysis. TriMet vehicles run over 100 routes serving nearly 500,000 stop arrivals everyday. This gives a good idea as to the extent of area covered and the volume of the data to be procured. The huge dimension of the data and the amount of man hours required for extracting the data from their system necessitated a personal visit to TriMet's office in Portland, Oregon.

The raw input data acquired were used to generate variables and information for model estimation. Some of these variables were directly derived from these raw input data, while others were generated through applying a modified version of T-BEST to TriMet.

Directly Derived Variables

The raw input data collected were used to directly generate many of the variables for model estimation. Understanding and processing the data was a long learning process, taking a six-month period from September 2005 through March 2006.

Schedule Data

There were sixteen different schedules over the six months of the two booking periods. Different schedules consisted of different sets of stops and routes. No single schedule encompassed all the stops for which APC data were available.

The schedule data were used to generate data on the number of vehicle arrivals for each stop (frequency) and data on vehicle travel time between consecutive stops. Vehicle travel time and frequencies were averaged over all the schedules.

Routes and Stop Data

The main complicating factor in dealing with route data was that some of the routes deviated from the primary paths for some of the vehicle trips. This demanded coding of route variants as separate routes for T-BEST. Route variants were classified using ArcGIS. Different trip numbers for each schedule were allotted to route variants based on the starting and ending stops of the route.

The complicating factor in dealing with stop data was that T-BEST requires sequential stop IDs. The presence of multiple schedules and no schedule containing all stops served in the six-month period created problems. Among all the trips that took place for each route by direction, the trip that served the maximum number of stops was selected. Other stops were manually coded into T-BEST. A variable in the original TriMet data measuring distance along a route was used to generate stop IDs.

APC Data

The APC data contained information related to the actual physical location of a transit vehicle, the status of APC (present, not present, and not working), and the actual number of boardings and alightings every time the vehicle stopped and the doors opened.

At least four factors complicated the processing of APC data. The first two relate to both bus and light rail. Not all vehicles arrived with APCs. The data did not have valid APC records for some of the vehicles with APCs. The other two factors relate to light rail only. Trains typically arrived with two cars, but only one car had APCs when any APC was present. Train cars have two doors on each side, but only one door had an APC when any APC was present on a train car.

For each time period, boarding at each bus stop was determined by multiplying the total number of bus arrivals during this period by the average boarding among all bus arrivals with APCs and with valid APC data. Boarding at each light rail station was processed in additional steps. Boarding from the APC door was doubled to get boarding for the whole car, and boarding for the whole car is further multiplied by the number of cars in a train to get boarding for the entire train.

The streetcars did not have any APCs. TriMet provided boarding data collected from a small sample of ride-check data during the period for which APC data are available for bus and light rail.

Population and Employment Data

Population and their socio, demographic, and economic data were from the 2000 Census. Socio and demographic data were available at the block level, while economic data were available at the block-group level.

TriMet also provided employment in a point layer shapefile format. While the employment data were specific to individual addresses, they were in ranges. The median of the range was used as the employment level for each address. The employment data also gave the SIC codes, which were used to identify commercial, industrial, and service employment.

T-BEST Generated Variables

Modified T-BEST was used to generate several variables for model estimation. These are transfer potential P_0 and accessibility variables A_1 through A_4 . It was a long and difficult process taking more than six months from April 2006 to October 2006. The first obstacle was to get T-BEST successfully running. While there may be other reasons, the main difficulty was that the original version of T-BEST could not handle a large system like TriMet. After many modifications by the developer through a lengthy trial and error process, applying modified T-BEST was finally successful. There were many other reasons for the difficult process, as summarized next.

Some of these reasons directly relate to adjustments to the data collected from TriMet. For example, streetcar stations were taken out of the estimation database because all of them had transfer opportunities. As a result, they cannot be used in estimating the direct boarding models. Another example is the separation of light rail stations from bus stops. T-BEST was originally

designed to reflect differences in modal technologies in different constant terms. After some trial and error, it was decided to model bus stops separately. An additional example is how park-n-ride lots are related to stops. Initially these lots were treated just like any other special generators and were related to the closest stop for each route and direction. But this approach did not make sense because it would mean that a park-n-ride lot would serve both directions of a bus route and that a park-n-ride lot at a light rail station also serves any intersecting bus routes.

Other reasons relate to adjustments to T-BEST programming by the developer, including:

1. Changed the initial splitting of buffers from among all overlapping buffers to only those in the same direction of a route.
2. Changed how A_4 , which is the overlapped areas between A_1 and A_3 buffers, is calculated.
3. Changed the coefficient in the impedance function so that the friction factor declines faster with impedance.
4. Increased the impedance cutoff threshold from 45 minutes to 75 minutes.
5. Reduced the radius of buffers from $\frac{1}{4}$ to $\frac{3}{16}$ of a mile.
6. Reduced the radius from $\frac{1}{8}$ to $\frac{1}{16}$ of a mile for defining S_0 .
7. Eliminated outbound transfer opportunities at the first stop on a route and inbound transfer opportunities on the last stop of a route.

Additional reasons relate to identifying and making corrections in T-BEST programming. The following are some examples:

1. How A_2 is determined;
2. A_4 was greater than A_1 or A_3 ;
3. How impedance components were weights;
4. Whether a stop has transfer opportunities.

It is important to point out that while light rail and streetcar stations were excluded from developing boarding equations, the entire transit network was used in generating transfer potential and the accessibility measures.

These issues were identified and dealt with at different points of time during the six-month period. While the reasons vary for many of these changes and adjustments, they were made after lengthy discussions among the research team and the developer. For every change that was made, much effort was spent on identifying the change, re-running T-BEST, generating output files, organizing an estimation dataset, and estimating models.

MODEL ESTIMATION

Requirements

There are two basic requirements that the estimated boarding equations should meet. One is statistical, and the other is theoretical. Each is discussed below.

Statistical Requirements

Model estimation requires selecting a statistical model that matches the nature of data at hand. Boardings at individual stops are a type of count data. Count data have two distinguishing features. One feature is that they are integers, and the other is that boardings are zero for a large portion of stops. Table 1 shows the number and percent of stops with zero boardings for each time period.

Table 1. Bus Stops with Zero Boardings by Time Period

Interval	AMPEAK	MIDDAY	PMPEAK	NIGHT	SATURDAY	SUNDAY
0-Boarding Stops	5,059	4,098	4,808	5,687	2,168	2,080
All Stops	9,926	9,671	9,877	9,874	7,734	6,844
% 0-Boarding	51	42	49	58	28	30

The commonly used linear regression model is inappropriate for count data. Rather, count data typically are modeled with Poisson and related statistical models. Poisson is the simplest but has a restrictive assumption that the mean and variance of the error terms are the same. Negative Binomial relaxes this assumption. More advanced models within this group deal with special features of count data. One special feature relates to whether the occurrence of zeros is actual behavioral or the result of sampling. The occurrence of zeros in the current dataset is unlikely to have resulted from randomness in data collection for two reasons. One reason is that the TriMet boarding data are from buses that have a high APC penetration rate. The other reason is that the boarding data cover a period of 6 months. As a result, this research focuses on Negative Binomial with Poisson as the backup in case Negative Binomial fails to converge.

Theoretical Requirements

Estimated boarding models should meet three simple theoretical requirements. The first requirement relates to direct boardings. People board transit to go somewhere for something. If the accessibility to population and employment from boarding a particular stop is zero, direct boardings at that stop should be zero. Mathematically, this requires that the sum of A_1 to employment and A_1 to population enters the model in a log form. The second requirement relates to transfer boardings. People who board at a stop to transfer to a particular route alight at a nearby stop along another route. If nobody alights at any nearby stops along other routes, transfer boardings at a subject stop should be zero. Mathematically, this requires that P_0 from upstream boardings along other routes enters the model in a log form. The third requirement relates to buffer population and employment for predicting direct boarding. If nobody lives or works in the buffer of a subject stop, direct boardings at that stop should be zero. Mathematically, this requires that the sum of total population and total employment enters direct boarding models in a log form. While the first two requirements were implemented when possible, the third requirement was not implemented. The research team decided that it was more important to use components of population and employment in the models than total population and total employment.

Results

Table 2 shows the estimation results for both direct boarding and transfer boarding for each time period for bus stops only. Poisson was used to estimate the models for the PMPEAK period, while Negative Binomial was used for all other time periods. For the Negative Binomial models, the parameter “Alpha” reported in the table indicates the degree to which Poisson’s restrictive assumption is violated. The following highlights observations from these estimated models.

The models fit the data well. One important indicator for model fit is improvements in log-likelihood between a simple model with constants only and the full model. The improvement in log-likelihood between “Restricted log likelihood” to “Log likelihood function” ranges from 25 percent to 57 percent in the direct boarding models. The improvement is much greater in the transfer boarding models, ranging from 70 percent to 84 percent. The only exception is for the afternoon peak models, which are Poisson-based.

Service frequency in a log form has a positive and statistically significant coefficient as expected. It appears that it has a larger impact on direct boarding than on transfer boarding on weekdays, but that it has a smaller impact on direct boarding than on transfer boarding on weekend days. Daily service span also has a positive and statistically significant effect on direct boardings as expected for both Saturdays and Sundays.

Buffer characteristics are included in the direct boarding only. Different population segments perform differently in different time periods. One surprising result is that commercial employment has a significantly greater impact on direct boarding than service employment.

The results on other stop and route characteristics are mostly expected. The number of park-n-ride lot spaces is considered for direct boarding in the morning peak and has a positive and statistically significant effect. Route types are considered for both direct and transfer boardings. Relative to other route types, radial and express routes attract additional direct boardings, particularly express routes. On the other hand, crosstown and circular routes attract more transfer boardings than both radial and express routes. In addition, the number of nearby stops on other routes that people may transfer from to a subject stop also is considered for the transfer models, and has a positive and statistically significant effect for all time periods.

Transfer potential P_0 has a positive and statistically significant effect on transfer boardings for all time periods. The effect appears to be greater on weekdays than on weekend days.

Table 2. Results

Variables	AMPEAK		MIDDAY		PMPEAK		NIGHT		Saturday		Sunday	
	Coeff.	t-ratio	Coeff.	t-ratio	Coeff.	t-ratio	Coeff.	t-ratio	Coeff.	t-ratio	Coeff.	t-ratio
Direct Boarding Models												
Constant	-5.27213	-30.9	-5.04569	-37.3	-4.616480	-49.6	-5.34199	-29.89	-4.78386	-27.6	-5.38740	-17.2
ln(frequency)	1.84369	28.4	1.76518	37.3	1.583610	56.8	1.63576	32.25	1.37427	21.5	1.29677	20.4
Daily service span in hours									0.02764	2.3		
ln(daily service span)											0.46251	2.8
Workers	0.00656	11.4										
Population in origin buffer			0.00080	2.3								
0-vehicle population in origin buffer	0.00314	2.0					0.00603	3.10	0.00434	2.8		
1-vehicle population in origin buffer											0.00220	3.2
Poverty population in origin buffer											0.00864	4.9
Black population in origin buffer	0.00377	3.6	0.00330	3.4	0.002772	8.7	0.00482	4.44	0.00244	2.5		
Hispanic population in origin buffer			0.00305	3.2	0.002654	10.1	0.00441	4.05	0.00437	4.3	0.00262	2.4
Multi-family population in origin buffer			0.00246	4.4	0.001597	12.2	0.00161	2.48	0.00237	4.4		
Service employment in origin buffer	0.00294	5.5	0.00373	6.7	0.001847	14.0	0.00270	4.92	0.00231	4.9	0.00249	5.0
Commercial employment in origin buffer	0.00510	4.5	0.01310	12.0	0.009622	43.7	0.01596	13.51	0.01553	14.8	0.01541	14.2
Number of park-n-ride lot spaces	0.03194	3.5										
Stop on an express route	0.90536	3.8					1.08191	3.55				
Stop on a radial route	0.18548	3.9	0.22337	4.9	0.134750	2.8			0.28680	2.8		
Stop on a crosstown route					0.153351	3.1			0.15262	1.5		
Stop near a college or university							0.42702	1.60	0.65219	3.1	1.00474	4.6
$A_2+A_3-A_4$ to population	-0.00005	-7.9	-0.00005	-5.8	-0.000018	-5.9	-0.00007	-4.44	-0.00003	-2.7	-0.00005	-3.3
A_4 to population			0.00012	2.9								
$A_2+A_3-A_4$ to employment	-0.00001	-2.1	-0.00002	-2.8	-0.000028	-12.3	-0.00003	-2.90	-0.00006	-6.6	-0.00005	-5.5
A_4 to employment			-0.00005	-1.7								
ln(A_1 to population+employment)	0.12143	8.2	0.09104	11.9	0.173492	16.5	0.12902	8.52	0.06556	10.5	0.07839	11.7
Alpha	1.29663	15.7	1.41534	25.8			1.41604	15.61	1.44934	33.0	1.47726	29.9
Observations		4805.0		4977.0		4776.0		5048.00		4268.0		3789.0
Log likelihood function		-6895.3		-9041.3		-11673.1		-5812.10		-10751.2		-8976.7
Restricted log likelihood		-9139.3		-15537.2		-17728.9		-8414.60		-25159.8		-19155.1
ρ squared		0.25		0.42		0.34		0.31		0.57		0.53
Transfer Boarding Models												
Constant	-3.53130	-16.4	-3.55439	-17.5	-4.427520	-88.3	-2.72408	-20.56	-5.43045	-17.1	-5.01103	-15.8
ln(frequency)	1.37322	12.5	1.55455	25.5	1.483530	108.6	1.19469	22.72	1.54214	20.3	1.48016	17.2
Stop on a crosstown route	0.25909	2.7			0.516167	34.1	0.28101	2.79				
Stop on a circular route					0.617693	21.0						
Stop near a regional mall			0.90669	2.0	0.477329	17.2			1.40059	3.8	1.20080	3.5
Stop near a college or university			1.20916	2.8	1.464340	39.7						
ln(P_0 from boardings on other routes)	0.18918	10.2	0.11710	12.1	0.332583	58.1	0.03576	4.80	0.35078	12.6	0.38455	12.6
Inbound stops on other routes	0.06537	8.4	0.07315	7.2	0.030397	37.2	0.09582	11.57	0.07799	7.3	0.06981	5.5
A_1 to population	0.00006	2.1					0.00062	5.74				
A_4 to population					0.000025	5.5	-0.00062	-3.65				
A_1 to employment	0.00069	4.2					0.00043	1.87			0.00072	3.0
$A_2+A_3-A_4$ to employment	-0.00001	-4.8	-0.00002	-5.8	-0.000001	-3.9	-0.00002	-6.93	-0.00003	-8.0	-0.00003	-6.8
A_4 to employment	-0.00077	-4.8	-0.00008	-8.2	-0.000056	-31.6	-0.00063	-2.64	-0.00021	-10.7	-0.00094	-3.8
ln(A_1 to population+employment)			0.04807	3.3	0.052761	20.7			0.05132	2.5		
Alpha	7.09098	28.3	7.97665	28.9			6.91829	27.66	6.64712	28.5	6.38468	26.5
Observations		4588.0		4158.0		4617.0		4576.00		2963.0		2673.0
Log likelihood function		-6678.0		-7014.2		-34093.5		-6780.90		-7012.1		-5950.9
Restricted log likelihood		-22078.7		-37921.3		-53264.3		-25016.90		-51888.0		-37284.2
ρ squared		0.70		0.82		0.36		0.73		0.86		0.84

More important, the effect of the accessibility measures on direct and transfer boardings is as expected.

- The accessibility to downstream employment via the subject route (A_1) is positive and statistically significant for both direct and transfer boarding; however, the impact is much higher on direct than on transfer boarding.
- The accessibility to employment at alternative destinations through the opposite direction of the subject route or through other routes has been combined ($A_2+A_3-A_4$), and has an expected negative and statistically significant effect on both direct and transfer boarding. Relatively, however, the effect of this combined accessibility to alternative employment is larger on direct boarding than on transfer boarding.
- How the accessibility to employment that can be reached both via the subject stop and via other routes/stops (A_4) may impact boarding may go either way on a theoretical ground. With a few exceptions, A_4 to population is statistically insignificant for both direct and transfer boardings. A_4 to employment is mostly insignificant for direct boardings, but is statistically significant negative effect on transfer boardings.

APPLICATION

For forecasting purposes, the direct-boarding model for a given period would first be applied to all stops to forecast direct boarding. For any given stop along a subject route, the forecast direct boarding at all stops along other routes that feed into the subject stop is then used to measure the potential for transfers at the given stop. The next step would be to forecast transfer boarding at stops with transfer opportunities. Total boarding would be the sum of the two.

The re-estimated boarding equations should be applied with care. Using direct boarding for the morning peak as an example, the following illustrates how these equations should be used. Direct boarding for the morning peak is expected to be equal to the product of (frequency)^{1.84369}, (A_1 to population+employment)^{0.12143}, and the exponential function of the following linear combination:

$$\begin{aligned}
 & -5.27213 \\
 & +0.00656 * \text{Workers} \\
 & +0.00314 * \text{0-vehicle population in origin buffer} \\
 & +0.00377 * \text{Black population in origin buffer} \\
 & +0.00294 * \text{Service employment in origin buffer} \\
 & +0.00510 * \text{Commercial employment in origin buffer} \\
 & +0.03194 * \text{Number of park-n-ride lot spaces} \\
 & +0.90536 * \text{Stop on an express route} \\
 & +0.18548 * \text{Stop on a radial route} \\
 & -0.00005 * A_2+A_3-A_4 \text{ to population} \\
 & -0.00001 * A_2+A_3-A_4 \text{ to employment.}
 \end{aligned}$$

The models predict reasonably well. As an example, Table 3 compares the number of stops with certain boarding ranges between observed direct boardings and in-sample predictions of direct boardings during the morning peak period.

Table 3. Observed and In-Sample Predictions of Direct Boardings during Morning Peak

Interval	Observed	Predicted
[0,1)	2,748	2,573
[1,2)	685	857
[2,3)	347	513
[3,4)	245	303
[4,5)	175	160
[5,10)	351	240
[10,20)	186	112
[20,30)	44	14
[30,40)	13	13
[40,200)	11	17
[200,500)	0	2
[500,1000)	0	1
Total	4,805	4,805

However, serious over-predictions can occur at a small number of stops. In the example given here, the maximum observed boarding is smaller than 200, but the predicted boarding is greater than this maximum for three stops. Among these three stops, the predicted boarding is in the range from 200 to 500 for two stops and is over 500 for one stop.

CONCLUSIONS

This report has documented a research effort in re-estimating a set of boarding equations for the Transit Boarding Estimation and Simulation Tool (T-BEST). One objective of this research effort was to use boarding data that are far more reliable than what were used for estimating the current set of boarding equations. The other objective was to make improvements in the boarding equations.

Both objectives have been successfully achieved. The first objective has been achieved through using boarding data from a fleet of bus vehicles that had an APC penetration rate of approximately 75 percent. The second objective has been achieved through many improvements in the boarding questions. These include the following:

- Some of these improvements are structural: the current weekday peak period has been split into a weekday morning peak and a weekday afternoon peak.
- Some improvements are statistical: except for the afternoon peak period, boarding equations are estimated without the restrictive assumption that the mean and variance of error terms are equal.
- Some improvements involved adding additional desirable variables: 1) The effect of park-n-ride lots on direct boardings has been taken into account in the weekday morning peak; and 2) The effect of daily service span on both direct and transfer boardings has been added for Saturdays and Sundays.

- Some improvements involve how certain variables enter the questions: 1) The socio-demographic characteristics of population enter the equations as quantity rather than as shares; and 2) Daily service span, service frequency, and accessibility to population and employment from boarding at a subject stop enter the equations in a log form.
- Some improvements involve how individual variables are computed, including how overlapping stop buffers are split to avoid double-counting, how some of the accessibility measures are computed, how components of impedance are weighted, the threshold of impedance within which accessibility measures are computed, and the distance threshold within which people may transfer from one route to another.

In addition, some of these improvements were made by design at the very beginning of the research effort. These include splitting the morning and afternoon peaks, the inclusion of park-n-ride lot spaces, and the inclusion of daily service span on weekend days. Other improvements, however, were identified and considered through a trial and error process as part of data processing and model estimation.

The estimated boarding models perform well in terms of both how well the models fit the data and how observed and in-sample predictions compare. However, serious over-predictions can still occur. Being based on Poisson-type models, the estimated boarding equations are not going to prevent such over-predictions from occurring. It appears that such serious over-prediction occurs only at a small number of stops, and need to be dealt with individually as part of the validation process in each application of T-BEST.