Data mining of transit vehicle location, passenger count, and fare collection databases for intelligent transportation applications

Supporting efficient transit service by maximizing the use of available information

The operational performance of transit systems is currently evaluated on an as-needed basis rather than as a systematic component of transit system management. Evaluation typically focuses on either running time or travel patterns. However, the availability of detailed data from automatic vehicle location (AVL), automatic passenger count (APC), and electronic fare collection systems opens the door to systematic evaluation of transit system performance supporting a variety of transit planning and decision making applications.

In this project, researchers Henry Liu of the Department of Civil Engineering and Chen-Fu Liao of the Minnesota Traffic Observatory developed a data analysis framework to process vehicle location, passenger count, and electronic fare collection data, generate route-based performance measures, and utilize these measures for transit planning. The framework is shown in Figure 1.

![Figure 1. Transit performance analysis framework.](image)

Data processing

Data pre-processing and analysis are carried out in an SQL database. Raw data from AVL, APC, and electronic fare collection systems, along with route schedule data, are pre-processed to remove outliers and improve data quality. Due to the nature of GPS position data and the polling rate of AVL systems, a filtering algorithm is used to prepare the bus position data for analysis. Following pre-processing, data can be cross-validated through a data-matching process.

The prepared data are now ready for the main processing and data mining operations. The data processing system integrates the prepared data to generate
measures such as bus running time variations, schedule adherence, delays, boarding/alighting activities, and dwell time. The analysis methodology improves data quality though data mining and fusion processes.

The data analysis model presents opportunities to study transit travel patterns that may identify problems with transit service delivery. Future development of the data processing system is targeted at predicting potential transit service problems based on archived or real-time transit data.

Integration with arterial traffic management systems
In addition to data from transit vehicles, transit performance analysis can also include other data affecting bus travel time, such as current traffic conditions or historical traffic patterns. One source of such data is the SMART-SIGNAL system (Systematic Monitoring of Arterial Road Traffic and Signals) recently developed by the same research team. Installed alongside signal-control hardware, SMART-SIGNAL collects and archives event-based traffic signal data at multiple intersections and automatically generates real-time performance measures including travel time, number of stops, queue length, intersection delay, and level of service. The system is currently deployed along an 11-intersection segment of a major arterial roadway in the Twin Cities.

Data available from SMART-SIGNAL includes signal timing and phasing, arterial traffic volume, speed, delay, and travel time. Integration of the transit database model and the arterial travel time model implemented by SMART-SIGNAL will enable traffic managers to develop robust and effective operation plans and to support other Intelligent Transportation Systems applications. Potential applications include real-time bus arrival time prediction, running time estimation, schedule adjustments, recommendations for Transit Signal Priority (TSP) deployment, and real-time service management in the future. Data visualization based on system performance can also be used as an effective tool to support transit planning, scheduling and decision making.

System development and experimental results
Bus stop data analysis
Researchers obtained one month of raw AVL data (4.2 million records) and APC data (3.4 million records) collected in 2008 from Metro Transit, the main transit agency in the Minneapolis-St. Paul metropolitan area. At the time of this research, approximately 30% of Metro Transit buses were equipped with APC systems that collected boarding and alighting counts at each bus stop.

Each bus stop is situated within a Time Point (TP) zone, with zone boundaries typically set approximately 200 feet before and after the stop; TP zone configuration typically includes an intersection, and the bus stop may be located on the near side (bus stops before intersection) or the far side. Time Point Time is the time spent
inside the zone boundaries. Dwell time is defined as the time between doors-open and doors-close at the bus stop.

Bus TP time, calculated from AVL/APC data, includes:

1. Bus travel time from the check-in point to the actual stop location and bus doors open.
2. Bus dwell time at stop. Dwell time is a function of several factors: number of passengers boarding/alighting, bus type, bus floor configuration, payment type, wheelchair deployment, and passenger load.
3. Time between last bus door close and actual departure time. This interval includes bus holding time at time point if the bus is running ahead of schedule, and possible traffic signal delay (at near-side stops).
4. Time between pulling away from the stop and crossing the check-out boundary of the TP zone.

Current APC-equipped buses record the arrival and departure time when crossing time point boundaries. The bus door opening and closing time are not recorded by the AVL/APC system. Bus TP time analyses at time points for nearside and far-side stops, and for early and late buses are considered separately. Link travel time between two consecutive time points is also analyzed. Statistical model based on the empirical data can be developed for link travel time estimation. A time point based dwell time model and route-based simulation for route productivity, run time and recovery time analysis as well as scheduling planning and reliability analysis can be developed as valuable tool for transit agencies. Stop level dwell time can also be modeled when the stop-level data become more available. Developing a TP-based dwell model and route based simulation tool will allow transit agencies to optimize running time, route productivity and reliability.

A Graphical User Interface (GUI) prototype was also developed to automate data analysis and visualization for selected route, direction and service type. Time point level analysis, inter-TP travel time and route performance analysis are available from the AVL/APC data processing system. Figure 2 illustrates the visualization of route data.
Transit user data analysis

In order to investigate the access behavior of transit user, we obtained a dataset of over 20,000 U-Pass and MetroPass users from the University of Minnesota's Parking and Transportation Services (PTS). MetroPass holders are transit users working at University of Minnesota. Addresses of MetroPass users were geocoded to the 2005 Twin Cities Metro street map using GIS software to compute the traveling network distance from each registered address to corresponding locations of all tap-on transactions.

Distribution of aggregated traveling distance to access transit is plotted in Figure 3. The distribution histogram shows that 62% of the MetroPass users travel less than ½ mile to access transit to work. Almost half of the MetroPass users (45%) travel less than ¼ mile, most likely by walking, to access transit for employment. About 32% of the MetroPass users travel more than one mile, most likely by automobile, to take transit services.
Figure 3: Histogram of travel distance from home to transit stop for MetroPass users.