Open Government Data and Public Transportation

Kenneth Kuhn, University of Canterbury

Abstract

Governments are increasingly making public transportation data available to the public on the Internet. The data can be used to explore and characterize current and historical service levels or to forecast operations in the immediate future. This paper considers, as an example, real-time bus location data provided by the San Francisco Municipal Transit Agency. General techniques for making use of such data to benefit both providers and users of public transportation are described. There is a brief discussion of why the advantages of making data available often outweigh the disadvantages.

Introduction

Many governments are increasing opportunities for the general public to access government data over the Internet. One of the most famous examples is the www.data.gov Web site, set up by the national government of the United States to allow users to "easily find, download, and use datasets generated by the Federal government" (www.data.gov/about). The www.data.govt.nz and www.datasf.org Web sites link to datasets from the governments of the nation of New Zealand and the City and County of San Francisco, respectively. Several of the data sets made available recently relate to public transportation.

Google Transit allows the public to access transit route and schedule data hundreds of cities worldwide have provided through either the maps.google.com or www.google.com/transit Web sites. The Web sites provide directions on how to

take transit to complete a trip based on desired origin, destination, and either time of arrival or time of departure. Directions are typically viewed on a Google Maps interface familiar to most. The interface allows for fast and simple map panning and zooming, while at a higher level ensuring maps are reproducible and can be embedded into other Web pages. The display can be customized to reflect the device used to access the transit data and the preferred language of the user. Transit route and schedule data are provided to the public in a common format, making it relatively easy for those interested to download, understand, and manipulate the data behind Google Transit. In particular, it is relatively simple to make and then explore maps showing transit routes alongside one of the many other available data sets formatted for use within Google Maps. This can be done by those with a casual interest in public transportation service analyses or by developers interested in creating a commercial application based on the available data. All of this is accomplished with public agencies responsible only for the initial step of providing route and schedule data in the established format.

The Bay Area Rapid Transit system in the San Francisco Bay Area and the TriMet public transportation system in the Portland, Oregon, area make predictions of vehicle arrival times at stops available online. The San Francisco Municipal Transit Agency (SF MUNI) recently made such predictions available, along with the real-time locations of transit vehicles. The data are discussed in more detail in later sections of this paper and on Web sites accessible from www.datasf.org. The open, online posting of forecast and especially real-time vehicle location data will lead to a variety of applications involving analyzing historical or real-time transit service levels, as well as forecasting future operations.

This paper considers bus location data from San Francisco and suggests, by way of example, some ways in which the data could be processed to provide useful information. The following section describes the data, focusing on data collection and initial processing steps that will be important for a variety of applications. The next section focuses on analyzing historical or real-time service levels. Discussion regarding forecasting future operations follows. A brief subsequent section argues that the benefits of posting public transportation data, even real-time vehicle location data, on the Internet likely outweigh the costs.

Example Data and Pre-Processing

This section describes the collection and initial processing of example data from SF MUNI. San Francisco, like many cities, has collected information on the real-time positions of its transit vehicles for a number of years using an automatic vehicle location (AVL) system. There has been a great deal of interest in the transportation engineering research community regarding the use of AVL data, especially for predicting vehicle arrival and departure times at stops. For example, Maclean and Dailey (2002) report on the construction of one system that provides predictions of transit vehicle stop departure times to potential riders with mobile phones. Shalaby and Farhan (2004) describe another system that uses AVL and passenger count data to forecast stop arrival times for both potential customers as well as those controlling the public transportation system.

Recently, San Francisco began to allow anyone with an Internet connection to load Web pages that contain data describing, for each transit vehicle on a user-specified SF MUNI route, an identification tag referred to as a BusID, the latest recorded position of the vehicle by latitude and longitude (lat/lon), and a timestamp indicating when the lat/lon data were collected. To inform the following discussion of the issues associated with open transit data, data regarding the SF MUNI bus route "1" or "1 – California" available on the Internet were studied. A program was written in the computer programming language Ruby to periodically load Web pages and then save relevant data (i.e., to *scrape* data off the Internet) for subsequent analysis. The collected data contained the positions of the buses in operation on the SF MUNI 1 – California route every minute for three weeks in February 2010.

An immediately noticeable flaw in the data posted by SF MUNI is that provided lat/lon coordinates are often some distance away from the transit route specified. AVL systems typically identify the positions of transit vehicles via radio triangulation, often using global positioning system (GPS) satellites. In urban environments, radio waves reflect off buildings and other objects, introducing error in position estimation. Ochieng and Sauer (2002) report that in a trial in downtown London, roughly 30 percent of reported GPS fixes were not within 10 meters of true locations. Another concern is that AVL systems may incorrectly identify which buses are on which routes, due to hardware, software, or human error. This issue will be discussed further subsequently. Finally, drivers do not always follow anticipated routes, for instance, when maintenance work closes a portion of road on a bus route.

For several of the foreseeable uses of transit vehicle location data, it makes sense to pre-process available data, replacing reported lat/lon pairs with likely positions on the route of interest and discarding data points that refer to buses that are likely not on the route of interest. Pre-processing of this sort may provide a more realistic picture of actual transit vehicle locations on the route of interest. Ideally pre-processing would be based on controlled studies comparing actual positions of transit vehicles with data posted on the Internet. The Kalman filter is often used to refine GPS position estimates, and Cathey and Dailey (2003) have applied it to AVL data in this context, obtaining vehicle speeds and other relevant information in the process.

A simpler method for pre-processing the raw data was used here. Lat/lon data on the real-time locations of buses were compared to each link of the chosen bus route (route data being available on the Internet). The points on the bus route closest to the reported positions were then noted, along with the distances between location data points and the bus route. The data points furthest from the bus route in question were considered outliers and discarded. The calculations involved finding lines perpendicular to transit route links and took negligible amounts of computation time. It seems reasonable to expect such a method will be used by those with a casual interest in studying public transport operations based on open data.

Figure 1 shows an example of the output of this procedure. In the section of San Francisco shown in Figure 1, the 1 bus route runs on Clay Street towards downtown (at right) but returns on Sacramento Street, a block to the south, away from the downtown area. Small question marks indicate the positions of data points that were among the 15 percent of data points furthest away from the route of the 1. These data points were subsequently removed from the collected data. Triangles indicate the positions of all other data points, and straight lines connect these data points with the route of the 1. The street map, like the bus position data, is available to the general public on the Internet (at www.openstreetmap.org).

After this preprocessing, a series of latitude and longitude pairs all located on the bus route of interest were obtained. It was then possible to reduce these data to normalized, one-dimensional measures of how far along the route different buses were at different times. Following the language of Cathey and Dailey (2003), the one-dimensional measure is labeled "distance-into-trip."

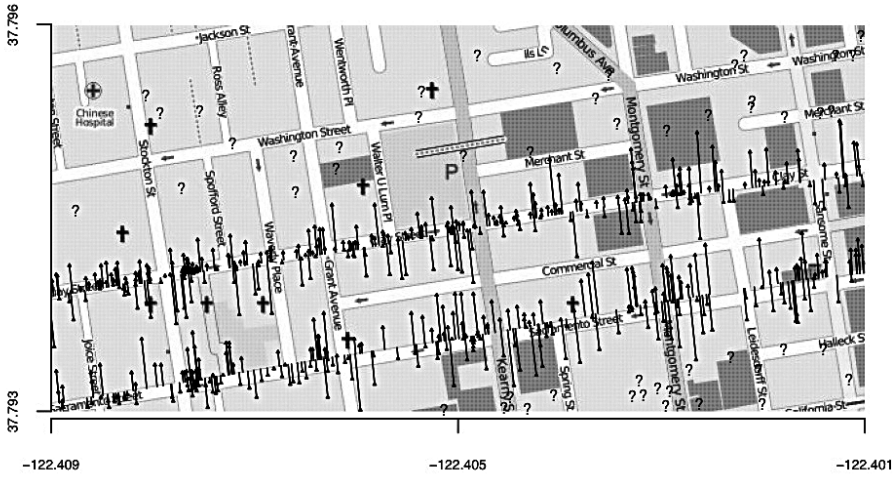

Figure 1. Matching Data Points to Positions on SF MUNI Bus Route 1

Figure 2 contains two plots of distance-into-trip data, showing bus operations on the 1 over a portion of its route and at a particular stop. On the left is a time-distance diagram. Small triangles show measures of distance-into-trip plotted against the times these measures were recorded. Straight lines link data points associated with the same BusID. Individual transit stops are associated with different distances-into-trip. The locations of three actual stops on the SF MUNI bus route 1 are identified by dotted horizontal lines in the time-distance diagram in Figure 2. The many uses of time-distance diagrams have been noted by Bruun et al. (1999).

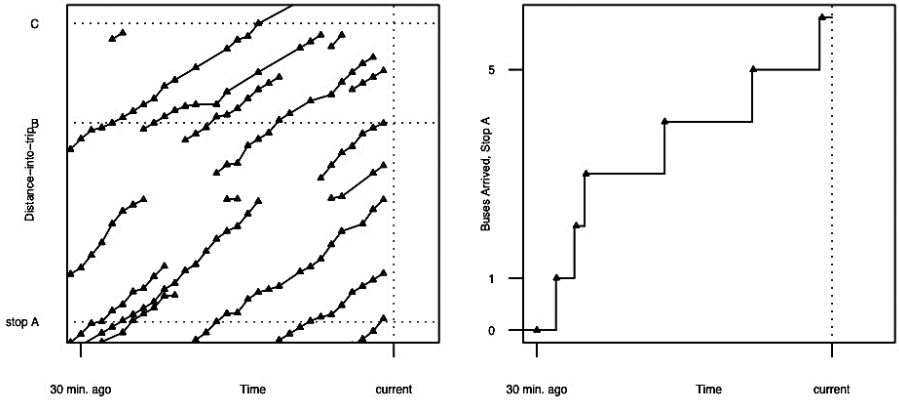

Figure 2. Visualizing Distance-into-Trip Data

On the right of Figure 2, the data from the time-distance diagram is processed to show how many buses have passed the location of one bus stop as a function of time. The small triangles here mark the times each bus was first actually observed at a location downstream of the selected stop, while the lines link estimated stop arrival times based on the time-distance diagram. The computational burden of the steps required to create graphs of the type shown in Figure 2 is minimal, and such graphs could be created in real-time as data are collected. To make this point, time in Figure 2 is measured in terms of minutes prior to "current" time.

Time-distance diagrams based on the data scraped off the Internet often showed buses appearing and disappearing in the middle of their route. This may reflect errors in the association of buses to routes, or prolonged periods of time when buses were either not reporting location data or reporting data significant distances away from their route. Other times, data points that appeared to track a single vehicle trajectory were associated with multiple BusID tags.

Although posted data are imperfect, plots like those shown in Figure 2 are meaningful and relatively easy to create. The forms of both plots are well known in transportation engineering. It would be possible for someone with training in this area to scan graphs like those provided and immediately recognize when and where there are problems associated with the actual provided bus service. For instance, if bus bunching were a problem, the time-space diagram would show the trajectories of buses in a bunch in close proximity to one another, with large headways on the time axis separating different bunches of trajectories. The characteristic step shape of the graph at the right of Figure 2 would become more irregular, with long headways alternating with sharp jumps up the graph whenever a bunch of buses arrived. If data are available regarding passenger arrival times at transit stops (for instance, data on turnstile movements at subway train stations), it would be relatively simple to compare these data to plots like those at the right in Figure 2 to determine where and when transit passengers were or are waiting for service. Monitoring graphs like those presented in Figure 2 in real-time could aid tactical or operational decision making, while performing analyses of historical data could aid strategic planning. The reader is referred to Bruun et al. (1999) for further discussion of the potential of time-distance diagrams in particular.

Analyzing Service Levels

Figure 2 contains estimates of the times transit vehicles arrived at stops based on empirical observations of when the vehicles were last observed traveling towards

stops and first observed traveling away from stops. Bus arrival times at stops can be used to study the headways between transit vehicles. For instance, Figure 3 shows histograms of the headways at the three stops depicted in Figure 2. Such a figure could be studied to see how frequent and how regular public transport service is at different stops. Figure 3 indicates that service at stop C was somewhat less regular than service at stops A and B. Scheduled headways varied between 3 and 15 minutes during the period when data were collected but were identical across the three stops.

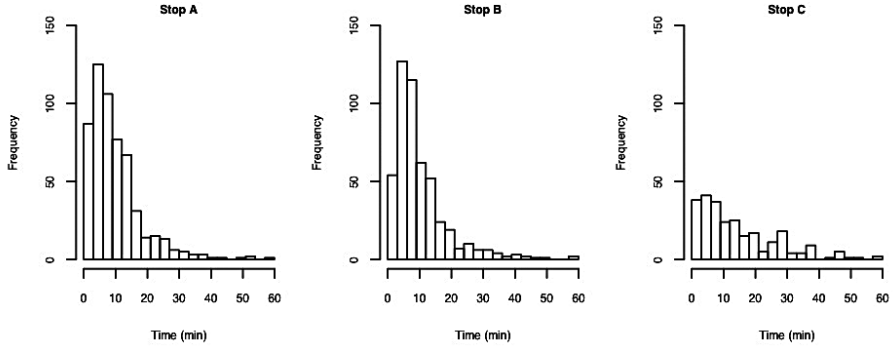

Figure 3. Observed Headways at Three Stops

By linking data points associated with the same identification tag, as in the time-distance diagram in Figure 2, it also becomes possible to study the magnitude and regularity of bus travel times on different sections of the roadway network. Such analyses would prove useful if a public agency, not necessarily the agency managing the public transport system, was considering implementing traffic management initiatives to control transit vehicle or general traffic speeds. Figure 4 contains histograms of the transit times different buses took traversing the two links connecting the three stops shown in Figure 2.

Figure 4 makes clear that travel times are significantly longer and less consistent on travel between stops A and B, as opposed to between stops B and C. Such data might help convince decision makers to invest in traffic management initiatives on the roadways linking stops A and B. It is worth noting here that recent analysis shows that the variance of the travel time on a trip across multiple sections of a roadway network is grossly underestimated by the sum of link-specific travel time variance estimates (i.e., ignoring covariance terms) (Nicholson et al. 2010). Technical points like this are likely to be overlooked by casual policy analysts, possibly leading to erroneous findings.

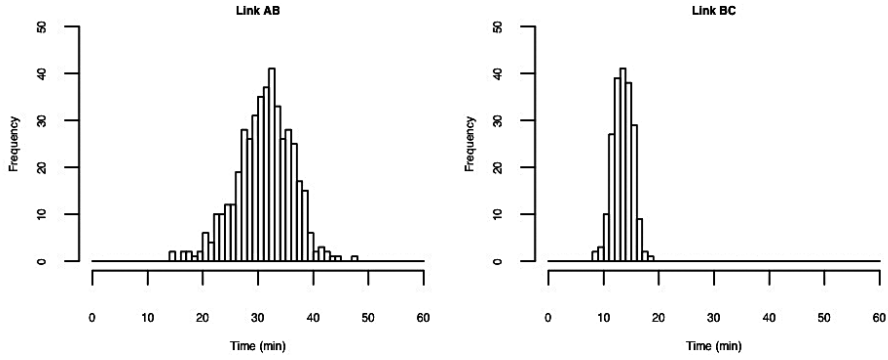

Figure 4. Observed Travel Times on Two Links

As more local authorities post data on the Internet, it would be possible to compare public transportation service at stops or over routes in different cities. Data on the service levels offered by public transportation systems also could be compared to other data sets to explore potential correlations. For example, weather data could be used to study how precipitation impacts transit service on different routes or in different cities. Public health, land use, and demographic data already support a wide range of interesting studies. For instance, one work by Yi et al. (2008) describes how maps of cancer incidence and age-adjusted mortality rates can be created quickly, easily, and for free using open government data and open-source mapping tools. There are numerous studies that could be done comparing public health, land use, and demographic data to public transportation data. The online publication of real-time transit vehicle location data will allow such studies to explore actual, rather than scheduled, transit service. As more open government data are created, the possibilities for further research expand combinatorially.

Forecasting Future Operations

Some of the most interesting applications of open public transportation data involve predicting vehicle arrival and departure times at stops or otherwise forecasting future operations. As was mentioned previously, significant research effort has been directed towards establishing techniques for generating and displaying predictions of arrival and departure times for buses. To inform discussion, this section will describe two methods to forecast the times particular buses take to traverse the portion of the route of the SF MUNI 1 – California line between stops A and C shown in Figure 2.

An example of a naïve approach to forecasting would be to predict that the travel time of a bus on a particular portion of its route will match the travel time of the last bus to have traversed the same portion of the route. Slightly better results would likely result if a few recent travel times were considered and averaged, possibly weighted according to how long ago they were recorded. An alternate naïve approach would be to predict the travel time using the travel time recorded at the same time-of-day on a similar day. Again, multiple data points, related by time-of-day, could be studied to improve results. Combining the two approaches described above would allow for consideration of seasonality (time-of-day dependence) as well as more localized variations in traffic conditions. One framework generalizing the naïve approaches described above is the k-Nearest Neighbour (kNN) method. This method will be described and used here, based on previous research employing the technique for traffic flow forecasting (Smith et al. 2002).

Whenever an estimate of travel time is requested, information would be provided on the current state of the bus and the bus route. Such information would include current time-of-day and day of the week. Such information also could include data such as prevailing weather conditions and how far ahead or behind schedule the preceding vehicle is, if such information were deemed relevant to travel time prediction. Observed state information is compared to similar state information associated with empirical observations of travel times. The k observed travel times deemed to have the most similar state information are then selected. The kNN method then estimates the travel time of the bus in question by averaging the selected travel times, possibly weighted according to proximity (in terms of state information) to current conditions.

The kNN method is relatively easy to implement, requiring limited application-specific expertise, and makes no parametric assumptions on variables of interest. The method does require definitions of system state and the metric for evaluating the differences between states. Exploratory analysis of historical data should be used to ensure the data used to define states are relevant to the data being forecast. For instance, here, data from one work week of operations of the SF MUNI 1 bus route were set aside for initial data analysis. As Figure 5 makes clear, no correlation was evident between how far ahead or behind schedule one bus was and the travel time for the following bus on the link chosen for analysis. (This finding helps explain why bus bunching was not observed in Figure 2.)

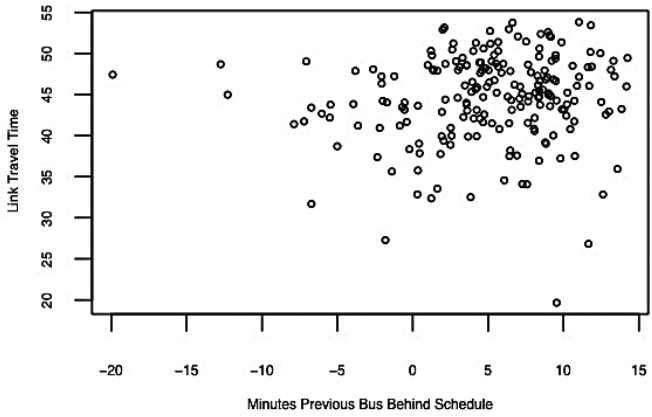

Figure 5. Link Travel Times and Previous Bus Tardiness

In the example work shown here k was set to 10 i.e., 10 previously-observed data points were averaged (with equal weighting) to create travel time estimates. The number 10 was chosen arbitrarily. Condition states were based on date and time-of-day data. Only data from the same day of the week were considered when coming up with a travel time estimate. For state pairs associated with the same date, the "distance" between the states was defined as the number of minutes' difference in time-of-day. For states whose dates were different but fell on the same day of the week, the measure described above was multiplied by 2. Again, the chosen approach is somewhat arbitrary. Exploratory analyses, like the interpretation of Figure 5 above, can be used to inform model selection.

A number of alternate approaches for travel time estimation are available based on neural networks (Huisken and Berkum 2003), time series analysis (Smith et al. 2002), and Kalman filters (Shalaby and Farhan 2004). The Kalman filter method is used here to enable a comparison between different predictive techniques. The chosen method is described as the "Link Running Time Prediction Algorithm" in the work of Shalaby and Farhan (2004).

A brief description of the selected Kalman filter algorithm follows; for further details, the reader is referred to Shalaby and Farhan (2004). Terms representing filter gain (g), loop gain (a), filter error (e), and predicted travel times (p) are calculated iteratively. At time step $t+1$, the terms are calculated using formulae (1) through (4):

$$g(t+1) = (e(t) + VAR) / (e(t) + 2\,VAR) \quad (1)$$

$$a(t+1) = 1 - g(t+1) \quad (2)$$

$$e(t+1) = VAR\ g(t+1) \quad (3)$$

$$p(t+1) = a(t+1)\ x(t) + g(t+1)\ y(t+1) \quad (4)$$

The input data include *x(t)*, the actual travel time of the previous bus at time step *t*, *y(t+1)*, the actual travel time of the bus observed at time step *t+1* on the previous day, and VAR a measure of the variance of the input and output data based on observations of travel times at time t on the previous three days. This approach is (again) based on estimating travel times based on the most recently-observed travel times and observations of travel times reported at the same time-of-day on preceding days.

Here, travel time data on the section of the SF MUNI 1 between stations A and C in Figure 2 were collected during the first two working weeks of February 2010. Data from the initial week were used to generate initial historical data sets required by the kNN and Kalman filter algorithms. Figure 6 plots observed travel times from the second work week plotted against predicted travel times for both tested algorithms.

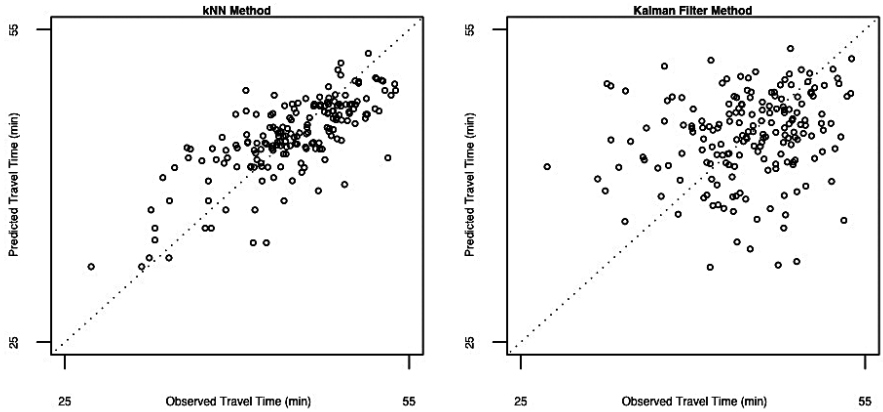

Figure 6. Predicted and Observed Travel Times

Figure 6 shows that the kNN algorithm produced somewhat more accurate estimates of travel time. It must be said that the kNN algorithm used significantly more input data when estimating travel times. There is a multitude of ways to set up kNN and Kalman filter algorithms for travel time prediction, and the results presented here should not be used to justify favoring one approach over the other. If the goal were to develop a travel time prediction algorithm for actual application, further

research should be done exploring residual data values to find opportunities to increase accuracy.

The Case for Open Government Data

The preceding sections described how real-time vehicle location data could be processed to analyze public transportation service levels or forecast future operations. It is here argued that it is in the government's interest to place such data on the Internet. In particular, analyses are likely to be more diverse and applications more efficient when raw data are made available to all. Robinson et al. (2009) have identified specific technical areas where it is preferable to have numerous private actors, rather than one government agency, processing data. Three of these areas of clear significance for public transportation are "advanced search," "mashups with other data sources," and "data visualization."

In the context of this paper, advanced search relates to the ability of interested parties to find data on the service levels offered by a particular set of public transport routes at a particular set of stops and over a particular set of time periods. It would be in the commercial interest of private companies to have efficient algorithms for selecting data relevant to user searches. It is worth noting that users will have significantly different search requests, due to the multitude of meaningful ways vehicle location data can be used (some of which were described above). Allowing many actors to access raw data increases the chances that a user with an unusual search request will be able to find a way to satisfy this request.

A mashup is the result of matching or comparing two or more distinct data sets. In the context of public transportation data, mashups could be especially useful, for instance, linking route maps of two different public transportation systems or comparing public health and transit data. Government agencies can create their own mashups, but opening up data will create a more diverse set of results. It is here possible to leverage the public's interest in transit. There are already large numbers of interesting mashups based on the limited public transportation data currently available on the Internet. For example, on www.thestar.com/staticcontent/822896, a map shows district-specific rates of driver's license suspensions for drunk driving alongside discs identifying areas within 1 kilometer of subway stations in Toronto. The mashup provides weak evidence for the hypothesis that those who live within walking distance of the subway are less likely to drink and drive and could spur further research or changes to public policy.

There are large numbers of interesting ways in which private citizens and companies have visualized available public transportation data without creating mashups. For instance, www.swisstrains.ch shows the real-time positions of trains in Switzerland on a Google map. This Web site, like the previously-discussed Google Transit, is popular in part because data are shown on a Google Maps interface that many find familiar and highly usable, allowing quick and easy panning and zooming. Again, it is in the commercial interest of parties processing the government data to provide high-quality data visualization tools. The relative advantage of private sector provision of data visualization services will become more apparent as more complex data are made available, requiring a greater integration of advanced search and data visualization.

Studies indicate that many public transportation customers are very interested in receiving information on system operations in the immediate future (Dailey 2001). This public interest suggests that when data are made available, private citizens and companies will create applications based on the data. Empirical evidence supports this conclusion. BayTripper (www.baytripper.org) is one of many currently-available applications based on SF MUNI data. These applications supplement public transportation service, for instance, allowing users to wait at home rather than at a bus stop for a delayed bus. To support such applications, public agencies are required only to make vehicle location data available. In particular, the government need not specify what data applications will provide to users, how to forecast transit vehicle movements, or how to display forecast data. In fact, the best results are typically achieved when the government makes raw data available and does not focus on encouraging one or two specific potential applications of such data (Robinson et al. 2009). Competition makes it likely that the best-designed applications become the most-used applications.

It is worth noting that all of the applications using transit data cited in this paper are free to the general public. Private citizens with a casual interest in studying public transportation service levels seek to attract attention and create policy change but not to make money. Companies providing applications that describe or predict transit operations will use automated analyses where the marginal costs of providing information to one additional user are essentially nil. Such companies will typically generate revenue via advertising or very small user fees.

Conclusion

The San Francisco Municipal Transit Agency recently made information on the real-time positions of its vehicles available to the general public on the Internet. This research describes how such data encourage private citizens and corporations to create products, especially Web and mobile phone applications, which enhance or supplement public transportation services. Further research monitoring the impacts of San Francisco's decision to post public transportation data on the Internet is warranted. It appears likely that a committed and savvy public transportation agency can extract significant value for a minimal investment posting operation data on the Internet.

References

Bruun, E. C., Vuchic, V. R., and Y.-E. Shin. 1999. Time-distance diagrams: A powerful tool for service planning and control. *Journal of Public Transportation* 2: 1-24.

Cathey, F. W., and D. J. Dailey. 2003. A prescription for transit arrival/departure prediction using automatic vehicle location data. *Transportation Research Part C* 11: 241-264.

Cherry, C., Hickman, M., and A. Garg. 2006. Design of a map-based transit itinerary planner. *Journal of Public Transportation* 9: 45-68.

Dailey, D. J.. 2001. Smart Trek: A model deployment initiative. Final Research Report WA-RD 505.1. Washington State Department of Transportation.

Huisken, G., and E. C. van Berkum. 2003. A comparative analysis of short-range travel time prediction methods. Transportation Research Board annual meeting, Washington D.C.

Maclean, S. D., and D. J. Dailey. 2002. Wireless Internet access to real-time transit information. *Transportation Research Record* 1791: 92-98.

Nicholson, A., Kuhn, K., and K. Munakata. 2010. Travel time reliability and correlation. Submitted to *Transportation Research Part A*.

Ochieng, W. Y., and K. Sauer. 2002. Urban road transport navigation: Performance of the global positioning system after selective availability. *Transportation Research Part C* 10: 171-187.

Peng, Z.-R., and R. Huang. 2000. Design and development of interactive trip planning for Web-based transit information systems. *Transportation Research Part C* 8: 409-425.

Robinson, D., Yu, H., Zeller, W. P., and E. W. Felten. 2009. Government data and the invisible hand. *Yale Journal of Law and Technology* 11: 160-175.

Shalaby, A., and A. Farhan. 2004. Prediction model of bus arrival and departure times using AVL and APC data. *Journal of Public Transportation* 7: 41-61.

Smith, B. L., Williams, B. M., and R. K. Oswald. 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C* 10: 303-321.

Yi, Q., Hoskins, R. E., Hillringhouse, E. A., Sorensen, S. S., Oberle, M. W., Fuller, S. S., and J. C. Wallace. 2008. Integrating open-source technologies to build low-cost information systems for improved access to public health data. *International Journal of Health Geographics* 7: 29.

About the Author

Kenneth D. Kuhn (*kenneth.kuhn@canterbury.ac.nz*) is a lecturer in the Department of Civil and Natural Resources Engineering at the University of Canterbury in Christchurch, New Zealand. His research explores the mathematical and economic foundations of transportation decision making.