2015

Keeping Track—The Fusion of Large, Automatically Collected Transport Data in Capturing Long-Term System Change

Oliver Lock, *Veitch Lister Consulting*
Gregory D. Erhardt, *RAND Europe*

OLIVER LOCK
Transport Analyst
Veitch Lister Consulting
oliver.lock@veitchlister.com.au

GREGORY D. ERHARDT
Doctoral Researcher
University College London
greg.erhardt.13@ucl.ac.uk

KEEPING TRACK – THE FUSION OF LARGE, AUTOMATICALLY-COLLECTED TRANSPORT DATA IN CAPTURING LONG-TERM SYSTEM CHANGE

This paper discusses the challenges and opportunities of using large, automatically collected data sources for the purpose of measuring and monitoring transit system performance over time, demonstrating factors such as changes in reliability and ridership.

It discusses work undertaken within the initial stages of ongoing research for the analysis of transport in San Francisco using the city's public transport vehicle location and passenger counting data, expanded with openly-available timetable information. It provides examples of open-source software tools developed for the fusion, analysis and visualisation of these large data sources. Such tools enable explorative and retrospective analyses of the city's transport network and provide opportunity of further work and research into network performance and integration of multiple other large data sources.

The paper investigates and highlights methods whereby data can be manipulated in order to assist in the creation of long-term data inventories of travel behavior and transport performance. These methods can be applicable to the data landscape for planning and modelling of transport systems in Australian cities.

Key words: Transit Performance Monitoring, AVL-APC, GTFS, Public Transport

1. Introduction

An emerging topic in travel forecasting is that of the use of 'big data' sources in transport planning and modelling. This data can be defined loosely as datasets that are too large and complex to manipulate or interrogate with typical methods or tools. As this topic is emerging in travel forecasting, standard practices are yet to be fully defined. Further to this, these data sources can be proprietary in nature, costly and their use can cause privacy concerns.

Examples of such data sources, ever-increasing in volume, are those automatically generated within many transit systems around the world. These can be found collected by automatic fare collection systems, passenger counting systems and vehicle location technologies. While the implementation of these technologies by transit agencies is not necessarily new, after now many years of being able to archive this data, there is further means to utilise it beyond its initial purpose – to study changes in the network over time and in response to specific transport investments.

As these technologies are continuously collected they have an advantage over traditional travel data collection methods, which are generally sampled at specific points of time and, due to costs, at a less frequent rate. In performing studies over long periods of time, with continuous data coverage, transport research projects can potentially be provided with enough data to achieve statistically sound results. Furthermore, analysis can be more directly applied to times before-

and-after new transport projects are introduced.

Despite these advantages, there are also evident barriers to using the data generated by these technologies. Much of this data can be difficult to fully utilise due to sheer volume (Sutton, 2004) and data validation that may be required (Furth et al., 2006, Hammerle et al, 2005). Due to large costs, the technologies are generally only able to be deployed on a proportion of the vehicles. Furthermore, the accuracy of these counts is ultimately reliant on the accuracy of the sensors, which may perform poorly in conditions such as overcrowding (Furth et al, 2006). These limitations make it clear that research that establishes methods for data cleaning, expansion and representation is required to assist transit agencies in making the most of these sources.

2. Literature Review

Recent initiatives have begun assessing performance measures generated from these data sources in high levels of detail. Data systems have been developed that encapsulate data processing and reporting (Liao & Liu, 2010; Liao, 2011), apply data mining methods in an effort to improve operational performance (Cevallos & Wang, 2008), and examine bus bunching (Byon et al., 2011; Feng & Figliozzi, 2011). Initial attempts have been made to visualise the data at a network level (Berkow, El-Geneidy, Bertini, & Crout, 2009; Mesbah, Currie, Lennon, & Northcott, 2012).

The research outlined in this paper by Erhardt et. al (2014) distinguishes itself from previous projects in two ways. Firstly, it does not assume full data coverage, and rather establishes a methodology to 1) expand the data to the whole schedule, 2) impute the missing values and 3) apply weighting methods to represent total ridership. One of the main advantages of taking this approach in that it allows the analysis to be performed without data needing to be collected for the whole system, which can reduce costs and ensures areas with partial or missing data coverage are not entirely excluded.

Secondly, the analysis enabled by this research is able to be performed over a long time period – from 2008 to present. The tool developed focuses on providing the capability to compare one period of time with another, placing the focus on changes rather than current conditions. This is as opposed to other applications, where the focus has been on providing a snapshot of current operating conditions over shorter periods of time (Liao & Liu, 2010; Feng & Figliozzi, 2011; Wang, Li, Liu, He, & Wang, 2013; Chen & Chen, 2009). The tool developed allows the ability to examine changes in specific spatial areas of the cities, which can be traced to individual projects or policy changes.

This paper will focus on the presenting an outline of methodologies established to expand, impute and weight the transit data, as well as on the tools that enable the exploratory data analysis of a wide number of transit performance variables. This study uses San Francisco as a case study, taking advantage of the automated vehicle location (AVL) and automated passenger count (APC) data currently available on the city transit system.

2. Methodology

This section will go through the steps required to transform the data into an appropriate format for analysis. Firstly, it will discuss the data sources that were used themselves – AVL, APC and GTFS data. Secondly, it will go further to discuss how this data is cleaned and expanded into aggregates and meaningful performance measures and integrated into a visualisation tool.

2.1 AVL/APC Data

The data source was generated by the San Francisco Municipal Transport Agency (SFMTA), which includes an approximately 25% representative sample from 2008 of the city's bus network equipped with these technologies. Table 1 demonstrates the main variables utilised within this dataset. Equipped buses are rotated across the system each month, according to a deployment plan. These are executed in a way specific individual trips are sampled on an at least once per month regular basis, so a bus running at a specific time on a specific route is captured (SFCTA, 2013).

Variable Name	Description	Example
STOPA	Stop Sequence	49
QSTOP	Unique Stop ID	6125
ANAME	Stop Name	POST ST & GRANT AVE
HR,MIN,SEC,MONTH,DAY,YEAR	Clock Hour, Minute, Second of Stop Arrival	16:35:44 05/09/2013
ON	Passengers boarding total	0
OFF	Passengers alighting total	0
LOAD	Load departing stop	2
ROUTE	Route ID	3
LAT / LONG	GPS latitude / longitude	37.79048 / 122.39881
MILES	GPS-based miles since last stop	4.5
TRIP	Trip Departure Time (from leaving initial position)	1330
DOW	Schedule day of week	Weekday = 1, Sat = 2, Sun =3
DIR	Direction (Inbound towards Downtown or Outbound)	1 = Inbound, 0 = Outbound
VEHNO	Bus Vehicle Number	2801
SCHTIM	Scheduled time	9999
SRTIME	Scheduled recovery time	99.9
ARTIME	Actual recovery time	99.9
DHR,DMIN,DSEC	Clock hour, minute, second of departure	16:36:44
FULL	Capacity of Coach	63
OVER	Load/Capacity	63
MAXVEL	Maximum velocity since last stop	0
TIMESTOP	Time bus stops	16:35:44
DOORCLOSE	Time bus closes doors	16:35:44
PULLOUT	Time bus moves away from stop	16:35:44

Table 1. Sample of variables available within the SF Muni AVL/APC data.

2.2 GTFS Data

The General Transit Feed Specification (GTFS) is a format developed as a consistent worldwide standard for public transportation schedules and their geographic information. This standard allows an open collaboration between public transit agencies and software developers, allowing applications to be written that utilise data in an interoperable way, meaning across different modes within one city, but also that same application compatible with another city's data. Depending on the agency, schedules can potentially be regularly updated, giving researchers open access into data that describes how the service changed over time. Table 2 displays the minimum requirements given by for an organisation's feed (Google Developers, 2012).

Filename	File description
agency.txt	The details of agencies that provide data in this feed.
stops.txt	Individual locations where vehicles pick up or drop off passengers.
trips.txt	Trips for each route. A trip is a sequence of two or more stops that occurs at a specific time.
stop_times.txt	Times that a vehicle arrives at and departs from individual stops for each trip.
calendar.txt	Dates for service IDs using a weekly schedule. Specify when service starts and ends, as well as days of the week where service is available.

Table 2. Sample of variables available within the SF Muni AVL/APC data.

2.3 Setting up the environment

Data analytics is inherently exploratory, which makes a setup for the rapid iteration of the data highly desirable. The data was stored in a Hierarchical Data Format (HDF5). This is best suited for write-once, read-many datasets. While data can be added to the file, the format is not well-suited for real-time data analysis - a consideration for further development of similar projects.

Visualisation was key in maintaining context among different time scales. With this particular dataset, the frequent zooming into small details was required. Python was chosen due the large amount of libraries available. The library 'pandas' combines high performance array-computing features and 'NumPy' allows flexible manipulation capabilities of spreadsheet platforms commonly used for data analysis, such as Excel, as well as querying functionality similar to that used in relational databases, such as SQL.

Using a data-frame format similar to statistical software R, it provides sophisticated functionality to make it easy to reshape, slice and dice, perform aggregations and select subsets of data. At this stage, the main user-interface of the tool takes form in IPython Notebook where methods and functions can be called and variables changed so that the analyst-user can perform a specific set of queries and create plots with the data. For the data visualisation component, IPython was used for graphs and load profiles, while the software Processing was used to visualise outputs of the tool in the form of an executable application.

2.4 Data processing

The raw AVL/APC data was cleaned and aggregated into meaningful sets using 'sfdata_wrangler', developed by Erhardt (2014). This software was developed in an open-source framework in the Python environment.

This was first utilised for the creation of a typical day for each month. The script is able aggregate the raw AVL/APC data into meaningful data frames for these monthly averages – providing summaries at system, route and stop-level for multiple time periods of the day. The 25% sample was combined with the GTFS schedule to provide estimates of trips that were under-sampled, oversampled or not sampled at all within particular months of the data. This process can be found outlined further in Erhardt et al. (2014), and summarised in Figure 1. Example results of these are outlined in the following section of the paper.

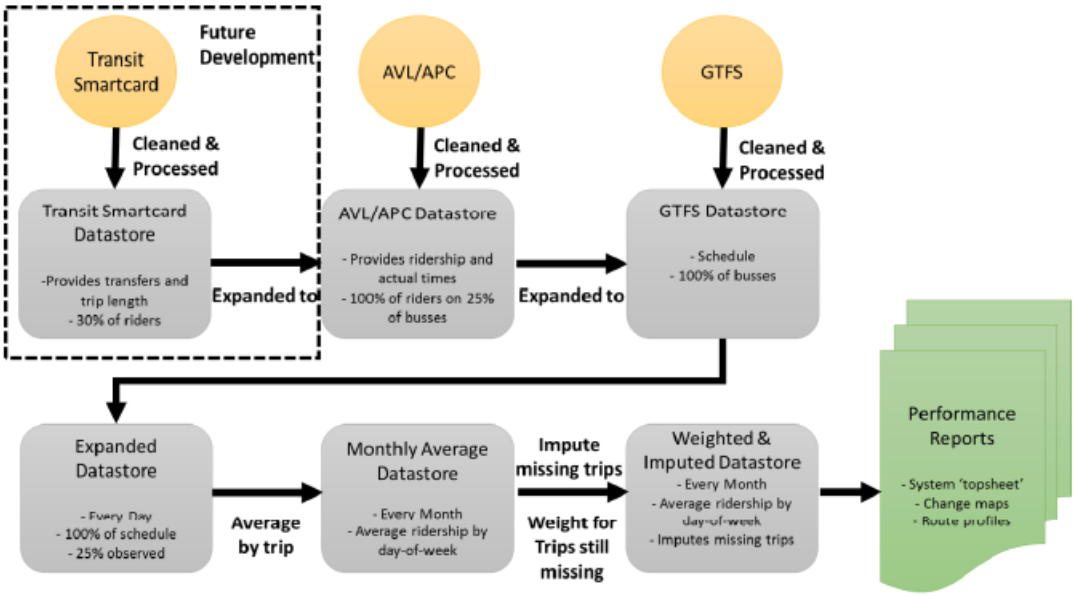

Figure 1. Data Processing Flow (Erhardt, Lock, Arcaute, & Batty, 2014)

2.5 Choosing a past transit policy

In order to test the data-mashing tool and its capability to perform an analysis of change of between one period of time and another, a past policy was selected. The policy selected was identified as one of the most major network-wide changes to occur in the period of time of data availability. In April 2009, SFMTA proposed system-wide service cuts of approximately 10% to the San Francisco Municipal Railway – including the bus system. This took effect from May 9, 2010 to September 4, 2010. The budget cuts reduced the frequency of service along specific routes and shortened hours, but did not change routings. On September 4, 2010, approximately 60% of the services eliminated were restored, with specific focus on evening and night-time service frequencies, as well as the last scheduled trips (SFMTA, 2010a).

3. Results

The following section covers a number of the results of the tool applied to a past policy analysis scenario. These show a number of the main capabilities of the tool.

3.1 Identifying service changes

The data and tool was able to identify spatial and temporal variation of the service changes on the network. Figure 2 highlights the changes in frequency within the timetable information as a series of generated facetted maps. Figure 4 shows clearly where routes were consolidated between the two time periods.

The tool was also able to effectively identify service changes network-wide before and after the policy change. Figure 2 highlights a basic line chart output of the tool, which shows the change in frequency of buses over the time period, using the combined timetabling and vehicle location data. Integrating passenger counts, Figure 3 shows waiting time per passenger, thus being able to measure passengers' experience by taking advantage of all three data sets. The figure indicates that the largest changes in waiting time were experienced by passengers after 19:00 and particularly after 22:00.

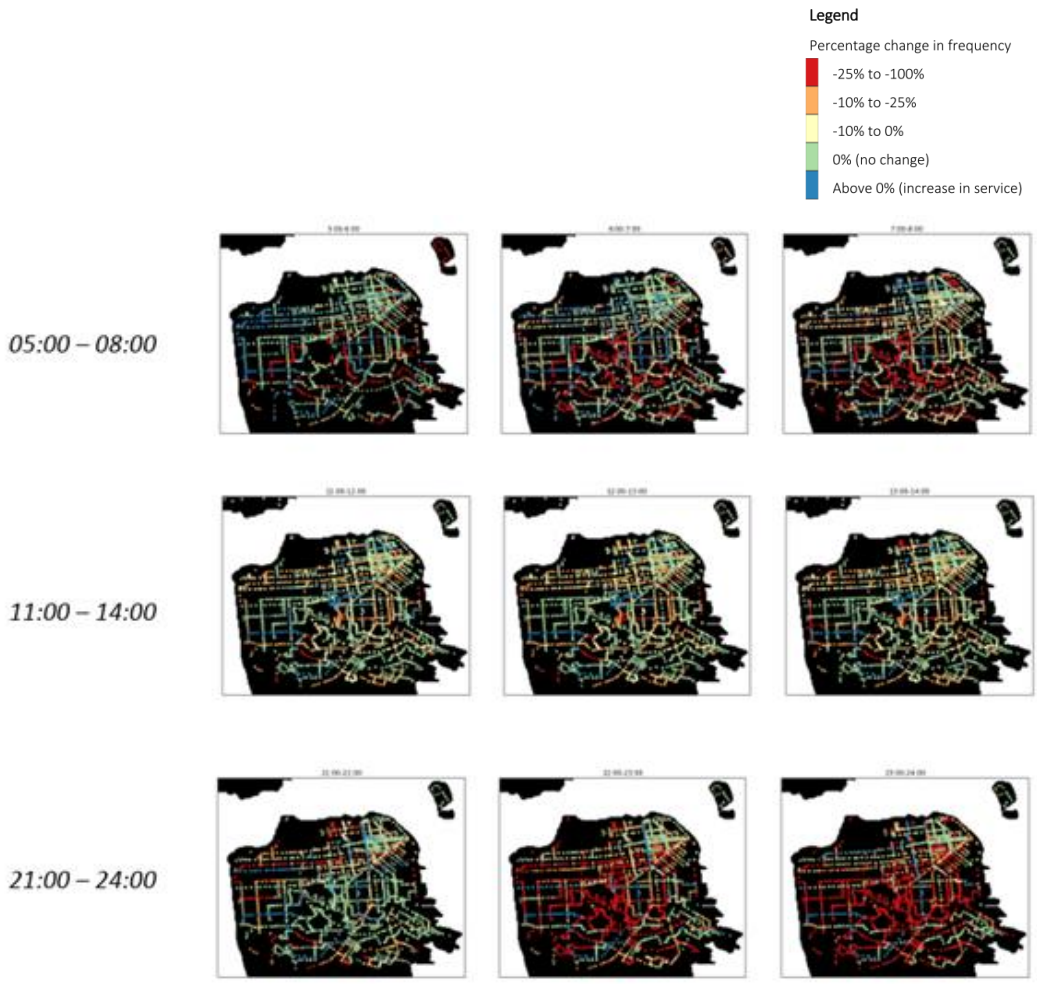

Figure 2. Facetted maps illustrating spatial and temporal changes in scheduled bus service frequency before and after policy change.

Figure 3. Differences between the network in July 2009 and July 2010, red circles indicating stops present in 2009 that were not present in 2010 (indicated in blue).

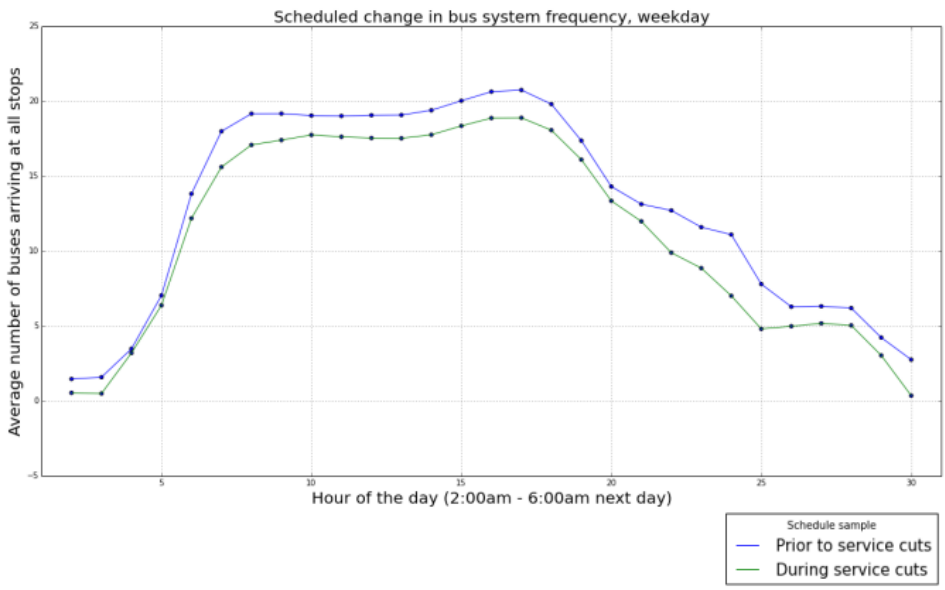

Figure 4. Number of buses scheduled to be arriving at each hour of the day, averaged across all bus stops in the system (weekday schedule).

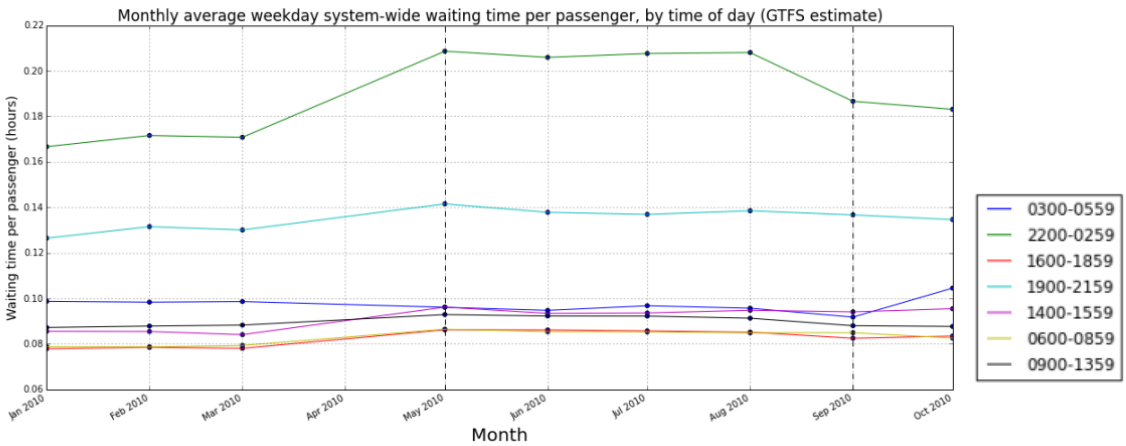

Figure 5. Monthly average estimated waiting time per passenger, as calculated through the scheduled data linked with AVL/APC data.

3.2 Identifying impacts

A number of variables were available for performance reporting analysis including on-time performance, speed, waiting time, crowding and ridership. The results shown here will focus on one specific variable, showing how the tool is able to report passenger crowding. This considers stop-level crowding, meaning the number of passengers arriving at an individual bus stop on arrival in relation to the capacity of the bus arriving at that individual stop.

In order to understand crowding along these routes, a function was also developed to draw load profiles of routes from the data, shown in Figure 6. The lines of the load profiles at the top show the capacity, which represents the crush load of a bus. In this example, July 2009 and 2010 were chosen to be compared. The red line represents July 2009's capacity, prior to the budget cuts, while the dashed blue is the capacity in 2010.

The lines with the dots indicate the passenger load of the buses at these two separate times, following the same colour conventions. The bar graphs represents passengers getting on (vertical, upwards), or off (vertical, downwards), with the same red/blue colour matching. Interpreting these graphs, one can see that the closer the dotted line gets to the dashed line, the higher probability that the bus can get overcrowded during this time.

While these are still made from aggregations of a month's worth of data, which is unlikely to capture extreme events, one can identify where incidents of crowding may have been much more likely to occur. For example, if a trip is consistently at 75% capacity, this indicates potentially high levels of crowding. Figure 6 shows a decrease in patronage after the service cuts, however also with a larger reduction in capacity. The diagram indicates that route 1-California shows one stage where the line was, before the cuts, at approximately 50% capacity later became approximately 70% capacity. The ability to generate these kind of comparative plots for multiple time periods can be useful observing changes in patronage, analyzing the effectiveness of the service changes and also as a tool for future service planning. The tool is flexible in its querying, where the user could quickly adapt this to include several more dates or routes to perform analysis of any specific change that needs to be investigated.

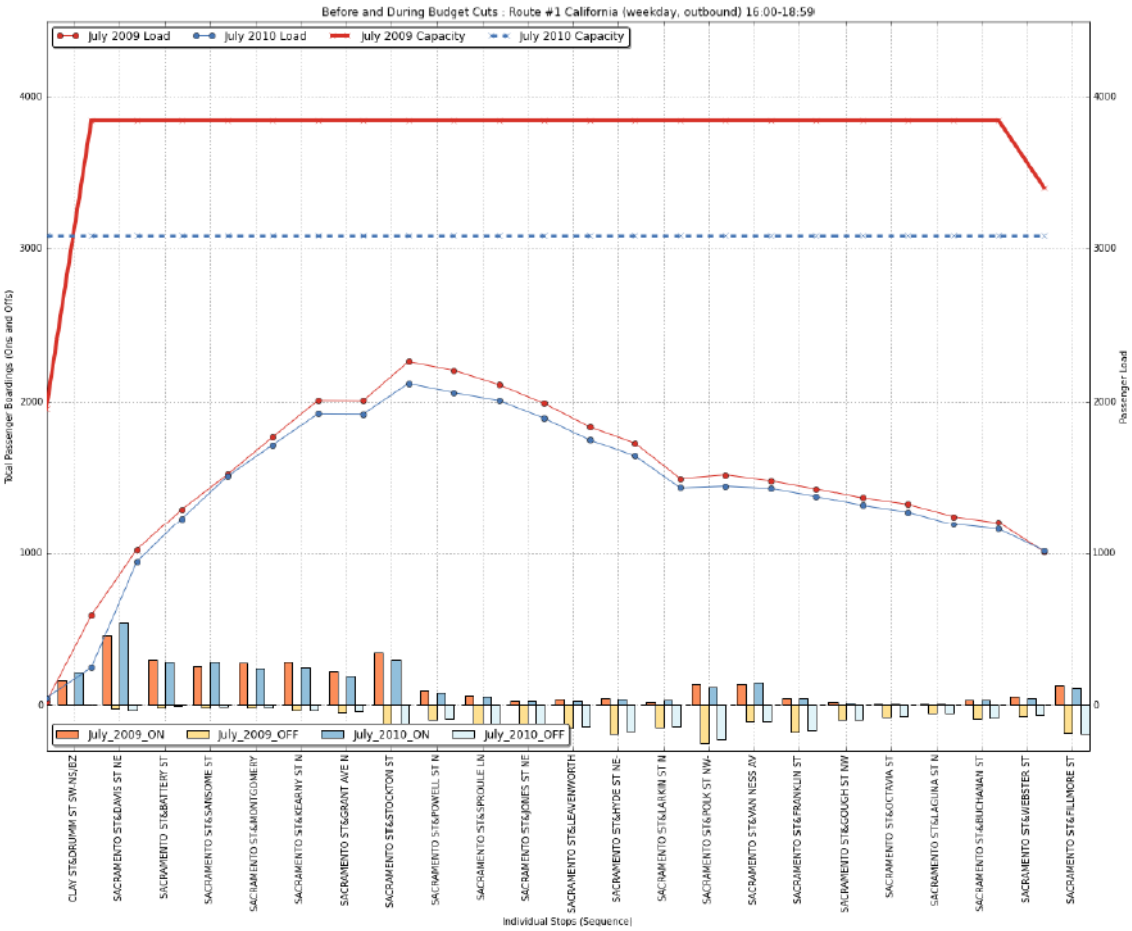

Figure 6. Load profile of Route #1 California, on a weekday, outbound in the PM peak between 16:00 and 18:59.

As well as comparative analysis of routes, network-wide analysis of the data was enabled as part of this research for a number of variables in the form of an interactive visualisation tool. The visualisation tool displays the individual stops recorded in the AVL/APC data on a map of San Francisco, colour-coded to match certain performance indicators and scaled to highlight changes in passenger load. Colour schemes are chosen from 'colorbrewer' (Harrower and Brewer, 2011). The colour schemes chosen were suited for diverging, quantitative data classes – wherein critical data break points need to be emphasized.

The tool is designed for the comparative visualisation of two time periods. It consists of three views – a 'before' month view, an 'after' month view and a difference view. Users can select the two months and then the absolute or relative (percentage) change in the values chosen. The tool allows rapid interaction with many variables, choosing from the performance measures as defined in the previous section – crowding, speed, waiting time and reliability. The time of day, grouped into early morning (3:00-5:59), morning peak (6:00-8:59), day (9:00-13:59), afternoon (14:00-15:59), afternoon peak (16:00-18:59), evening (19:00-21:59) and night (22:00+).

While the performance variables are indicated by coloured circles, the load variables increase the size of these circles. As the change in load can be both positive and negative, an increase in load was indicated by a filled circle, while a decrease is indicated by an outlined circle.

In this example, Figure 7 shows that the tool is able to clearly identify a number of routes which experienced changes in passenger crowding during the policy change. It is able to identify specific routes that experienced the high change in crowding, as well as showing how significant that change was in relative and absolute terms.

Used in conjunction, these analytical tools presented enable a number of capabilities for transport data analysis. Following the examples outlined in this section, it is clear that the tool is able to generate a large number of analytical tasks from the combination, expansion and imputation of numerous large transport datasets. These range from performance reporting, line charts, load profiles and network-wide spatial analysis.

While this section has shown examples mainly from frequency and crowding, in the current stages there are several different performance variables are available to be queried – such as speed, headway, passenger waiting time and on-time performance.

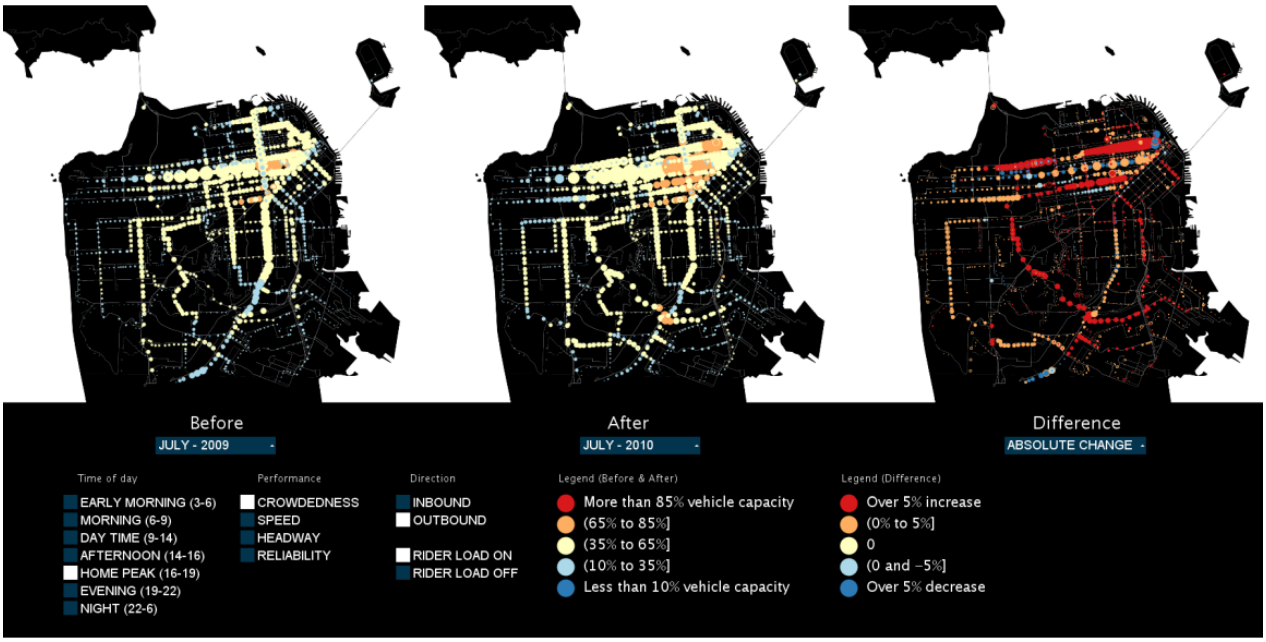

Figure 7. Comparison of outbound crowding (passengers at stop / capacity) between July 2009 and July 2010, in the PM peak between 16:00 and 19:00.

4. Discussion / Conclusion

The current product of this research is a data-mashing tool with multiple output features that can be used to measure transit system performance over time, as well as examples of its applications in a case study. The software and methodology used are adaptable to other cities with similar data available.

One of the initial challenges addressed in this research was that not all vehicles have location and passenger-counting technologies installed, so the tool had to be developed to match the entire schedule of trips (available within the GTFS data). After this process remaining missing data added the additional requirement of weights were needed to be introduced to compensate for missing data. In further research and development, integration of transit smartcard data, weighted to the passenger-counting data could further enrich the data set and potentially provide an effective understanding of the origin and destination of trips.

The outputs of the tool will need to be validated to compare with other available measures, such as officially published ridership totals. Once validated, there are a number of means in which these could enable alternate uses beyond its analytical capabilities, such as providing inputs or comparison with transport model results. Some examples of such applications include comparing public transport boarding/alighting with previously modelled results, examining whether responses to policy/investments in historic data match similar changes made in modelled data or as a means to assign more detailed information of public transport or traffic speeds to link speeds within a modelled network.

5. Further work

The expansion, analysis and visualisation strategies performed here are key in spreading capability of analysis of large datasets, particularly in areas or cities with inconsistent data coverage. At the time of writing, a number of the technologies outlined within this project are applicable to up-and-coming data availability in Australian cities and a number of them have been available for long enough periods to generate similar tools.

In particular, the GTFS scheduling data is now openly and publically accessible for all major Australian cities with the recent release of the data for Melbourne by Public Transport Victoria. An example of some potential similar applications is the real-time vehicle location data has also been recently released by Translink in Queensland in the form of GTFS-Realtime, which can provide similar utility to that of the combined AVL/GTFS data used in this study.

As a further example, Melbourne's tram network has been long equipped with AVL technology, which informs waiting-time applications. In order to apply similar methods in this study, integration of passenger counts would need to be obtained from transit smartcard data, which has its own privacy and data consistency challenges. For example, with the transit smartcard system in Melbourne (Myki), current tram passenger requirements only requires 'touching off' in certain zones network, which means the data would cover passenger boarding and not necessarily alighting. Further considerations would be how to capture instances of fare evasion, or when a user enters at a travel-free zone and linking the

touch-on time to specific location of the vehicle.

Further research could employ methods here in comparisons between cities, or developing tools / methods to a finer scale than monthly in order to analyse not just policy changes but disruptions such as transport strikes and major events. Further work can also be done in including other analytical libraries available in Python, such as those used in graph-theoretic network analysis.

In developing techniques for handling these big transport data sources further, as well as in identifying their challenges and opportunities, we enable planners and other professionals have the best information and decision-making tools available for designing effective urban transportation systems.

5. References

Berkow, M., El-Geneidy, A., Bertini, R., & Crout, D. (2009). Beyond Generating Transit Performance Measures: Visualisation and Statistical Analysis with Historic Data. Transportation Research Record: Journal of the Transportation Research Board, 2111(- 1), 158–168. doi:10.3141/2111-18

Byon, Y.-J., Cortés, C. E., C, M., Javier, F., Munizaga, M., & Zuniga, M. (2011). Transit Performance Monitoring and Analysis with Massive GPS Bus Probes of Transantiago in Santiago, Chile: Emphasis on Development of Indices for Bunching and Schedule Adherence. Presented at the Transportation Research Board 90th Annual Meeting. Retrieved from http://trid.trb.org/view.aspx?id=1092377

Cevallos, F., & Wang, X. (2008). Adams: Data Archiving and Mining System for Transit Service Improvements. Transportation Research Record: Journal of the Transportation Research Board, 2063(-1), 43–51. doi:10.3141/2063-06

Chen, W.-Y., & Chen, Z.-Y. (2009). A Simulation Model for Transit Service Unreliability Prevention Based on AVL-APC Data. In International Conference on Measuring Technology and Mechatronics Automation, 2009. ICMTMA '09 (Vol. 2, pp. 184–188). doi:10.1109/ICMTMA.2009.77

Erhardt, G. D. 2014. San Francisco Data Wrangler <sfdata_wrangler>. Github repository.

Erhardt, G.D., Lock, O., Arcaute,E. & Batty, M. 2014. A Big Data Mashing Tool for Measuring Transit System Performance. The Big Data and Urban Informatics Workshop. Chicago, Illinois, USA.

Feng, W. & Figliozzi, M. Using archived AVL/APC bus data to identify spatial-temporal causes of bus bunching. Compendium of papers of 90th Transportation Research Board, Annual Meeting, Washington, DC, 2011.

Feng, W., & Figliozzi, M. (2011). Using Archived AVL/APC Bus Data to Identify SpatialTemporal Causes of Bus Bunching. Presented at the 90th Annual Meeting of the Transportation Research Board, Washington, D.C. Retrieved from https://wiki.cecs.pdx.edu/pub/ItsWeb/TrbConferences/11-

4262_Using_Archived_AVLAPC_Bus_Data_to_Identify_Spatial-Temporal_Causes_of_Bus_Bunching.pdf

Further, P. G., Hemily, B.J., Muller,T.H. & Strathman, J. G. 2006. TCRP Report 113: Using archived AVL-APC data to improve transit performance and management, Washington D.C., Transportation Research Board.

Google Developers. 2012. General Transit Feed Specification Reference [Online]. Available: https://developers.google.com/transit/gtfs/reference [Accessed April 2014].

Hammerle, M., Haynes, M. & McNeil, S. 2005. Use of automatic vehicle location and passenger count data to evaluate bus operations. Transportation Research Record: Journal of the Transportation Research Board, 1903, 27-34.

Harrower, M. & Brewer, C. A. 2011. ColorBrewer. org: an online tool for selecting colour schemes for maps. The Map Reader: Theories of Mapping Practice and Cartographic Representation, 261-268.

Liao, C.-F. (2011). Data Driven Support Tool for Transit Data Analysis, Scheduling and Planning. Intelligent Transportation Systems Institute, Center for Transportation Studies, University of Minnesota. Retrieved from http://conservancy.umn.edu/bitstream/11299/149402/1/CTS11-15.pdf

Liao, C.-F., & Liu, H. (2010). Development of Data-Processing Framework for Transit Performance Analysis. Transportation Research Record: Journal of the Transportation Research Board, 2143(-1), 34–43. doi:10.3141/2143-05

Mesbah, M., Currie, G., Lennon, C., & Northcott, T. (2012). Spatial and temporal visualisation of transit operations performance data at a network level. Journal of Transport Geography, 25, 15–26. doi:10.1016/j.jtrangeo.2012.07.005

SFCTA 2013. San Francisco Congestion Management Program. The San Francisco County Transportation Authority.

SFCTA 2013. San Francisco Congestion Management Program. The San Francisco County Transportation Authority.

SFMTA. 2014a. About the SFMTA - History and Fleet : Muni Fleet [Online]. Available: http://sfmta.com/about-sfmta/our-history-and-fleet/sfmta-fleet [Accessed June 2014].

Sutton, J. C. 2004. Geographic information systems applications in transit, Transportation Research Board.

Wang, J., Li, Y., Liu, J., He, K., & Wang, P. (2013). Vulnerability Analysis and Passenger Source Prediction in Urban Rail Transit Networks. PLoS ONE, 8(11), e80178. doi:10.1371/journal.pone.0080178