Research in Transportation Economics 59 (2016) 236–241

Contents lists available at ScienceDirect

Research in Transportation Economics

journal homepage: www.elsevier.com/locate/retrec

Workshop 5 report: Harnessing big data

Gabriel E. Sánchez-Martínez [a], Marcela Munizaga [b], *

[a] Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[b] Departamento de Ingeniería Civil, Universidad de Chile, Chile

ARTICLE INFO

Article history:
Received 20 October 2016
Accepted 25 October 2016
Available online 2 November 2016

JEL codes:
C610
C630
L910
L920
R410
R420

Keywords:
Big data
Measurement
Implementation challenges
Analysis tools
Transit best practices

ABSTRACT

A group of researchers, consultants, software developers, and transit agencies convened in Santiago, Chile over 3 days as part of the Thredbo workshop titled “Harnessing Big Data”, to present their recent research and discuss the state of practice, state of the art, and future directions of big data in public transportation. This report documents their discussion. The key conclusion of the workshop is that, although much progress has been made in utilizing big data to improve transportation planning and operations, much remains to be done, both in terms of developing further analysis tools and use cases of big data, and of disseminating best practices so that they are adopted across the industry.

1. What is big data?

Among the first tasks addressed by the workshop was defining big data. Participants suggested various definitions and characteristics of big data, but there was agreement on two broad characteristics: size and complexity. Big data refers to large amounts of data that began existing recently, as a consequence of increased automation, improvements in sensing, storage and communication technologies. It's size can be overwhelming, often exceeding our capabilities to understand it and make use of it with traditional tools and methods. This has in some cases encouraged transportation scientists to collaborate with computer scientists. Big data is a relative term, both because what is considered big data today may not be in the future and because scientists of different fields may have different baselines for how large their datasets are before it becomes difficult to analyze them. For instance, big data in some sub-disciplines of physics and astronomy can be much larger than big data in transportation.

2. Applications

Transportation scientists and practitioners utilize big data in a variety of applications, ranging from high-level planning to performance measurement and delivery of new transportation services. Big data can improve usual tasks and analysis by replacing qualitative perceptions with objective quantitative measures. It also unlocks new ways of analyzing, monitoring, and delivering public transportation. A discussion of the sources of ITS data and possible applications is presented by Hemily (2015).

Big data can be leveraged to evaluate urban planning policies that might affect economic growth, quality of life, environmental sustainability, and socioeconomic equity. It can enhance public transportation planning at several levels, including service planning and operations planning, by increasing the accuracy and precision of demand and running time estimates, not only in terms of their typical values but also of their stochasticity. This can lead to more effective decisions regarding facility location (for example, the location of a new rail station or maintenance yard), network planning (for example, the alignment of a rail line extension, and the modification of the bus network in response to this extension), project evaluation (for example, comparing estimated costs and

* Corresponding author.
E-mail addresses: gsanmar@mit.edu (G.E. Sánchez-Martínez), mamuniza@ing.uchile.cl (M. Munizaga).

http://dx.doi.org/10.1016/j.retrec.2016.10.008
0739-8859/© 2016 Elsevier Ltd. All rights reserved.

benefits of alternative sets of bus, bus rapid transit, light rail, and heavy rail services), resource allocation (for example, assigning buses to routes to maximize service quality), and scheduling (for example, optimizing vehicle and crew schedules to increase the robustness of operations to delays). Gschwender, Munizaga, and Simonetti (2015) present a successful use case of smartcard data for policy and planning.

Big data can be used to measure the performance of delivered service, from the operator and passenger perspectives, both ex-post and in real-time. An application already common in leading transit agencies is the development and publishing of performance scorecards that reflect service quality or unreliability as experienced by passengers, through key performance measures such as excess waiting time. These performance reports enable management to detect in real-time when operations are disrupted, to track performance over time, and to measure the effect of reliability schemes and infrastructure investments on service quality. For instance, performance reports may show reliability problems in a subway worsening over time as the signaling system ages, and then improve following maintenance or a hardware upgrade. Valdez, Potter, Cook, and Miles (2015) explore the use of data from crowdsourcing applications, and conclude that this is a fruitful area for further research, with many challenges to be faced.

The practice of measuring performance can potentially improve operations by increasing the attentiveness of management personnel. Performance reports can be internal to the agency or public. In some cases, agencies have both detailed reports for management and simplified customer facing reports. Customer facing score cards can be an important part of an agency's public relations strategy. They can improve public perception by making the agency appear more technologically advanced and by showing that the agency cares about its users. Managers may hesitate about publishing scorecards because they can make problems more apparent and can provide material for news articles reflecting negative opinions of an agency's competency. However, this pressure can encourage managers to pay more attention to operations, and can ultimately lead to improved service quality. Public performance data has also enabled public transport and data enthusiasts to develop useful web and mobile phone applications at no cost to the agency, which can be considered a positive public involvement outcome. Mobile phone applications that help passengers plan their trips and give them estimated vehicle arrival times at the stop where they are waiting are a prominent example. These applications typically receive periodic network and schedule updates, as well as real-time vehicle positions, from data feeds published by the agency. The increasing opening of data and publishing of performance scorecards show that many agencies are convinced that the benefits of public data and performance reports outweigh the risks.

Real-time monitoring of transit operations enables sophisticated real-time control and incident management. Giesen, Julio, and Lizana (2015) present a real time prediction algorithm of bus speeds based on machine learning techniques. Location sensors installed in vehicles can stream data to a control center, where controllers can see how well drivers follow the schedule and the headways between vehicles. Control centers can respond to incidents requiring emergency maintenance or dispatching of police or paramedics. They can also regulate vehicle movement to reduce waiting times and improve reliability, by instructing drivers to execute control actions such as holding, short-turning, deadheading, and expressing. Typically, only vehicle locations are sent in real-time, but it is possible to also send passenger counting and fare collection data to capture demand and estimate loads. Raw data may be combined with inference models to make estimates of system state for aspects not directly observed, such as vehicle loads and number of people waiting at stops and platforms. These estimates can help control centers make smarter decisions. (Sánchez-Martínez, 2015).

Performance monitoring based on big data has proven useful in the context of contracting out operations. It is reasonable and increasingly common for agencies to expect private operators to utilize automatically collected data for planning and controlling operations. Agencies can collect vehicle location, fare collection, and passenger counting data to evaluate performance and determine how well service was delivered. Service quality expectations can be formalized through monetary penalties and incentives awarded to those operators not meeting or exceeding, respectively, these expectations. Data collected by the agency over time can be shared with bidding operators to encourage the development of realistic schedules and cost estimates, thereby minimizing the risk of bids based on unrealistic schedules that could not be delivered in practice and would result in poor service quality and tensions between the agency and the operator. One outstanding challenge is identifying whether poor performance is due to the operator or to external factors, such as traffic, that are beyond the control of the operator, in order for the agency to have realistic expectations, which are necessary to award incentives and penalties fairly. If operators perceive a high probability of being penalized for poor performance in spite of their efforts to operate well, they may include the expected penalties in their bid, thus passing the risk to the agency and undermining the incentive provisions in the contract.

Although it may involve advanced analysis not traditionally carried out by agencies, big data can help identify the factors that worsen transit service quality. For instance, a detailed analysis of bus running times can help an agency identify sources of delays and running time variability, by decomposing the effects of dwell times and vehicle movement between stops and through intersections. The results of such analyses shed light on potential priority schemes such as dedicated bus lanes and signal priority, pinpointing where and when investments in these kinds of transit priority may be most effective (Machlab, 2014). Arriagada, Gschwender, and Munizaga (2015) and Danes and Muñoz (2015) present models to identify the variables that contribute to increased travel time variability, which is related to bus bunching and poor service quality. Sánchez-Martínez, Koutsopoulos, and Wilson (2015) propose a framework for using automatically collected data and simulation modeling to maximize service performance in a set of high-frequency bus routes. Their approach can be used to detect routes operating with insufficient or excessive resources. They present an application using data from the Boston metropolitan area.

In the line of proposing methods and metrics that are useful for transport planners, Viggiano, Koutsopoulos, Attanucci, and Wilson (2015) propose a method to identify opportunities to improve the public transport system through route modifications as well as a specific metric based on the comparison between the observed travel time and the best possible travel time.

Big data is also being employed to analyze demand and understand passenger behavior. Data from fare collection, automatic passenger counters, and Bluetooth, Wi-Fi, and cellphone signal sensors is being used to study how individuals move throughout the city using the transit network and other modes. Some analyses are aggregate, focusing on the concentrations of demand in time and space, while others are disaggregate, focusing on how individuals travel and in particular how they change their travel behavior after a change in fare structure or network (Pan, Sun, & Looi, 2015; Shimamoto & Kondo, 2015). Machine learning has been applied to identify clusters of passengers that behave similarly, for example regular weekday commuters vs. occasional travelers, and to assign individual passengers to clusters (Goulet-Langlois, 2015; Ortega-Tong, 2013). This has numerous applications, ranging from network planning (for example, introducing a new bus line that directly connects two locations between which transit travel previously required a transfer) to marketing (for example, optimizing the location of adverts so that they are more relevant to the demographics commuting through a station). A more advanced application involves predicting and encouraging behavioral change. For example, an agency can promote an alternative network path to manage crowding, or it can estimate the changes in ridership due to an upcoming fare increase. Nakamura, Uno, Nakamura, Schmöcker, and Iwamoto (2015) explore the effect of loyalty programs on public transport users travel behavior using smart card data and a stated preference survey.

Aside from enhancing public transit operations in the aforementioned ways, big data has led to an increase in competition and innovation in the urban transportation industry. New business models have been created that blur the line between public transport, taxi, and private auto modes. Examples include car sharing and bike sharing programs, and demand responsive services involving the use of mobile phone applications from its users (to request the service) and from its drivers (to provide service and collect fare). There is currently a push for research and development on autonomous vehicles being done by the industry and academia, related to how big data can be leveraged to improve safety and increase the productivity of driverless fleets.

3. Data sources

In the context of public transportation, big data is obtained from a wide variety of sources, including sensors installed on vehicles and stations, crowdsourcing, video surveillance, Wifi, Bluetooth, and cellphone networks, and various forms of published urban data.

Transit agencies gather a lot of data from automatic vehicle location (AVL), automated fare collection (AFC), and automatic passenger counting (APC). AVL, which in some cases refers specifically to bus tracking systems, can more broadly refer to different technologies that identify the location of a vehicle and either collect it or transmit it in real time. Bus tracking systems often rely on satellite location services such as GPS, and are sometimes combined with odometers and inertial navigation systems for dead-reckoning, to improve tracking accuracy when vehicles are under bridges, crossing tunnels, or surrounded by tall buildings. Train tracking systems are often based on track circuit occupancy, but sometimes involve radio-frequency identification (RFID) and satellite-based location.

AFC systems consist of fare boxes installed in vehicles and fare gates installed in stations that interact with fare cards issued to passengers for the purpose of collecting fare. A variety of fare media exist, including cash, paper tickets, smart cards, and contactless bank cards. Fare boxes and gates usually record transactions that are sent to AFC servers either in real time or in periodic batches. These transactions include timestamps and identifiers for the fare card and the fare box or gate, from which the time and place of each transaction can be obtained. In the case of fare boxes without location data, it is often necessary to find the location based on the vehicle's AVL data. Cards are often taken as proxies for passengers, although it is possible that multiple people share a card and that a single passenger has multiple cards. Some transit systems require that passengers tap in and out of the system (e.g., Sydney), giving origin destination pairs directly, while others only require tapping in (e.g., Santiago). In the latter case, destinations can be inferred based on the sequence of taps.

APC systems comprise a variety of technologies that count passengers in the transit system. Some agencies have vehicles with equipment that counts boardings and alightings. Some trains have weight sensors on each car for the braking system that can also be used to roughly estimate loads. Fare gates often record passengers passing through, even when it is to exit on a system that does not require tapping out. Video surveillance feeds from both stations and vehicles can be processed by algorithms that count passengers. Telecommunication antennas (for Bluetooth, Wi-Fi, and cellphone) can also be used to gather data on the number of connected devices, which can be used to estimate the number of people within range of the antenna. People counting based on video surveillance and telecommunication antennas can be applied at the broader city level, outside of stations, bus stops, and public transportation vehicles.

Data can be collected from sensors not directly related (but relevant) to public transportation. Traffic sensors that count flow of vehicles through road links and intersections can be useful to study how traffic delays bus services. Weather data, in particular temperature and precipitation, can be used to study how adverse weather influences public transportation demand.

Although data often comes from automated sensors, it can also be produced by public transportation agencies and urban planning organizations. Geospatial data about zoning, land use, household density, and job density is useful for network planning and macroscopic demand modeling. Public transport agencies are increasingly publishing data about their network and schedules in machine-readable formats, in particular the General Transit Feed Specification (GTFS), which allows passengers to plan trips using Google Maps and other third party applications. Surveys are useful to gather data on aspects seldom available through automatically collected data streams, such as sociodemographics, home and work locations, and ownership of automobiles.

Big data can be crowd-sourced. In the context of transportation, this often happens through web and smartphone applications such as Google Maps, Waze, Moovit, Twitter, and Facebook. These applications provide some value to users, such as navigation or social media, and in turn can collect location data. Waze allows its users to report construction work or accidents causing traffic delays. Twitter and Facebook allow its users to post comments, which can sometimes be related to transportation; for example, users may alert others of a station fire causing train delays. Sometimes these comments are posted faster than through official channels. They can be mined for keywords, as a possible way of detecting relevant events.

4. Missing information

Big data often covers many aspects of transportation systems, but its users would often wish to know more. In particular, they often lack information about the sociodemographics and personal preferences and lifestyles of passengers, including household income, car ownership, detailed home and work locations, etc. In some countries detailed land use and zoning data is not available, making it more difficult to infer trip purpose. A detailed and unified description of urban infrastructure related to public transport (for example, number of lanes on streets, identification of segregated busways or priority lanes, and location of traffic signals) is sometimes missing. Knowledge about events that affect transportation systems, such as construction work, accidents, rail signal failures, and public demonstrations is sometimes available anecdotally or in free text, but rarely in machine-readable formats.

Tamblay, Galilea, and Muñoz (2015) propose a method to complement an OD matrix obtained from smart card transactions by Munizaga and Palma (2012), using land use information to incorporate the access and egress segments of the trips.

5. Challenges to access and utilization

There are a variety of challenges hindering access and utilization of big data for transportation applications, including technical, political, legal, and commercial.

One of the major challenges is ownership, which is sometimes tied to the commercial sensitivity of sharing data. AVL, AFC, and APC data are often owned by either the operator or the agency (which in some cases are the same entity). When agencies contract out operations and the contracts do not have data ownership provisions, operators may not be willing to share their data, because it can be used by competitors to estimate costs and bid against them for the same contract. This is unfortunate for the agency, planning organizations, and researchers because they lose the opportunity of gaining insights from analyzing the data, not to mention being unable to measure the performance of the private operator and encourage competition in the market. In some cases, operators who own the data are willing to share it with researchers or agencies under a non-disclosure agreement, which can also inhibit research utilizing the data. Network and schedule data in GTFS format is generally available to the public. Data collected through manual surveys is often owned by the entity that collects it. Data collected from traffic sensors and signals is often managed by a government agency, but usually by a different agency than the one in charge of public transportation, or sometimes by a different division of the same agency. Communication, cooperation, and sharing of data between the people in charge of traffic and signals and those in charge of public transport is uncommon. Data collected from usage of smartphone apps is typically owned by the entity that owns the app, which is often a business but at times is a public agency or a research group. When owned by a business, data on user locations and on the travel information being looked at or requested is not often available to the public due to its perceived commercial value.

Worries and legal constraints about privacy can also be a barrier. Big data projects can capture large amounts of personal or commercially sensitive data. The perceived legitimacy of institutions collecting and analyzing personal data is important to secure citizen participation in schemes. Care must be taken to safeguard this information and to make it available only in anonymized or aggregated forms to prevent its misuse (e.g. by terrorists or criminals). Additionally, it is important that users feel in control of their personal information (e.g. through opt-in or opt-out schemes), and that they perceive that their personal information contributes to the creation and fair distribution of the value created through big data (Valdez et al., 2015). Agency managers sometimes do not want to publicize that they collect and analyze transportation data at the individual level, for fear that it could lead to public concerns based on a misunderstanding of what data is collected and how it is used. However, agencies must usually disclose it in legal terms. Fare transaction data is usually anonymous, unless the card is registered, in which case the contact information may be associated. The latter is useful to understand the true origins and destinations of trips, starting from home instead of the local bus stop or rail station. Many passengers are willing to share their data when they understand it is being used for research and to improve planning. The willingness to share contact information or fill out a survey increases when passengers are offered the opportunity of entering a lottery to win prizes (Viggiano, Koutsopoulos, & Attanucci, 2014). Many applications do not require personal details. To minimize the risk of sharing data that could be tied to an individual, agencies and researchers can create an encrypted version of the identifier on fare cards that allows tracking an individual without knowing the actual identifier. On applications that don't require disaggregate data, aggregation over time and space can further improve privacy. For example, data on the number of passengers traveling between zones every hour of the weekday, averaged over a month, can be quite useful for planning.

Lack of interest in data and data-driven decision making by senior managers of transit agencies is another challenge. Agencies in which top managers champion data analytics have been the most successful in their data collection programs and are best positioned to conduct objective analysis based on demand and performance data. On the other hand, those without support from the top managers for big data projects make it difficult for researchers and analysts within the organization to encourage investment and hiring in information technology and telecommunications and to change policies and procedures based on what can be learned from data to gain efficiencies. A key challenge for researchers and others promoting big data in transportation is avoiding failures in data projects, such as making policy decisions without the due diligence of cleaning and validation, or without a transportation background, since failures may lead to losing trust and support from key people for future work with big data.

Pineda, Schwarz, and Godoy (2015) use exogenous data to validate the OD matrices obtained by Munizaga and Palma (2012). They compare the demand levels in the Metro network in Santiago, Chile obtained from the smartcard OD matrix with estimates obtained from a large survey (of about 150,000 observations) and with estimates based on load measurements. Pineda et al. conclude that the estimates obtained from the big data sources are highly reliable, a finding that has prompted the agency to consider cancelling or decreasing the frequency of the (currently) annual survey.

Even when data is available to researchers, several technical challenges can arise. Poor data quality due to sensor failures or mistakes in manual data entry can add a burdensome prerequisite of data cleaning before useful analysis can be carried out. Vehicle location data based on satellite tracking (e.g. GPS) sometimes has too long a polling interval, e.g. 30–90 s, which does not lend itself to identifying which stops were served or the arrival and departure times from stops (Bull & Herrera, 2015). Even on systems with more frequent polling or polling on geo-fences instead of intervals, there often are issues identifying when trips begin and end at terminals. Data on the location of bus stops is sometimes manually entered and can contain errors or be imprecise. Sometimes the data is not collected at all, or stored for a short period, or stored in proprietary systems that make accessing data difficult. Sometimes the databases can be accessed but they are poorly documented and not well understood. Another frequent problem is independent sets of data about stops, vehicles, routes, and trips in scheduling, AVL, AFC, and APC systems, which requires matching different identifiers, etc. The transit industry is challenged by a lack of standards regarding both hardware and software. Other technical challenges include recognizing and dealing with bad days and unusual events such as snowstorms, protests, and sport events to avoid mis-characterizing average performance, avoiding biases that may arise when scaling up samples of data, adjusting to ever-changing schedules and stop locations, and carrying out long-term panel analyses when passengers replace their fare media or share it with others (Hemily, 2015).

Another common challenge is the lack of expertise in data science. Many agencies lack staff skilled in database design and computer programming, and researchers themselves do not always have the data management knowhow required. A survey of tools used among the workshop participants revealed that a large variety of tools are being used, including Matlab, R, Stata, SPSS, Excel, and Access for statistics; Python (with modules such as sklearn for machine learning, and SciPy), C++, C#, Java, and shell scripts for programming; CPLEX, Gurobi, Minos, and GLPK for optimization; QlikView, iGraph, Tableu, D3, and CartoDB for visualization; Rapid Miner, SAP/BI, IBM/Cognos, Oracle/BI, MicroStrategy, and Hastus-Analytics for business intelligence; ArcGIS, QGIS, TransCAD, Google Earth, and OpenStreetMap for GIS and mapping; and PostgreSQL (often with PostGIS), Oracle, SQL Server, and TeraData for databases and data warehousing. In most cases the users of these tools have learned on their own in an ad-hoc fashion rather than through formal instruction. The analysis methods employed include origin, destination, and transfer inference, linear and logistic regression, simulation, clustering, and classification.

6. Increasing tangible realized gains

The big data movement is advancing rapidly, but more should be done to increase the tangible realized gains derived from big data applications in transportation. This takes effort on the part of both public transportation agencies and researchers. One such effort is making credible business cases for big data, showing how costs can be decreased or benefits increased, and to demonstrate, in quantitative terms when possible, the benefits of previous data projects. Data should be made public when there is no risk privacy being compromised, to encourage development of new tools and generate and public excitement and participation. When new tools are developed, user-friendly interfaces should be produced and visualizations should be emphasized to attract the interest of the public, managers, and policy makers. Agencies should welcome collaborations with academia for developing intellectual capital, and so academics focus their research on problems relevant to the industry. Agencies should also adopt standards for their data, so that tools can be used across agencies rather than being developed for specific agencies. Both academics and agencies should avoid the success bias, and share failures of data projects and research ideas so that others can learn and avoid repeating mistakes made by peers. The learning process of practitioners is as important as developing new tools. This includes educating agency staff on the uses of data and creating business strategies that emphasize objective data-driven analysis.

The *smart cities* paradigm is an opportunity to bring about positive change by emphasizing data-driven decision making and collaboration across stakeholders and sectors (e.g. healthcare, education, public safety, and transportation), with goals of becoming more efficient, competitive, and sustainable. It can take the shape of integration of transportation modes. For instance, Mexico City is integrating driver's licences and smart cards that can be used to pay for both transit and shared bicycles. It can also take the shape of integration across sectors. For example, technology can be used to monitor urban flows of drinking water, wastewater, energy, telecommunications, and transport, and all can be made accessible through the same interfaces. Big data could help build robust business cases for investing more in public transportation, by estimating its economic benefits and showing how these are distributed in more detail. The smart cities paradigm positions transportation as a means to a successful city rather than as an end in itself.

Most big data efforts have focused on analyzing the past to plan a better future, but there is growing interest in the development of predictive models. One of the challenges is real-time cleaning and validation of data feeds, since short-term predictions of, for example traffic states, are only practical when conducted automatically. Example applications include real-time mining of social media to identify events and problems as or before they occur, and coordinating transfers in low-frequency public transportation services when users alert the agency they wish to transfer through the use of a journey planner.

7. Conclusions

A group of researchers, consultants, software developers, and transit agencies convened in Santiago, Chile over 3 days as part of the Thredbo workshop titled “Harnessing Big Data”, to present their recent research and discuss the state of practice, state of the art, and future directions of big data in public transportation. Big data refers to the increasing size and complexity of data collected from systems, increasingly in an automated fashion, which presents challenges for analysis. In the context of public transportation, it includes vehicle locations, fare transactions, loads, traffic, and infrastructure, among others. Researchers and practitioners in the urban public transportation industry have made progress in utilizing big data to improve planning processes, with the ultimate goal of improving service quality and increasing the efficiency of operations. However, much remains to be done, both in terms of developing further analysis tools and use cases of big data, and of disseminating best practices so that they are adopted across the industry.

8. Policy recommendations

The most pressing needs include standardization of data formats, sharing of analysis tools and best practices across the industry (including smaller, less sophisticated transit agencies), documentation and dissemination business cases for big data projects in public transportation, collaboration between transit agencies and academics, utilization of data generated in transportation systems for applications outside of transportation (such as urban planning, healthcare, and public safety), application of existing tools to improve the management of operations contracted out, and promotion of smart cities. These issues should be addressed always recognizing that big data provides a detailed view of only part of the system, and safeguarding both user privacy and freedom from unreasonable legal and regulatory restrictions that prevent collection and utilization of data for the public good.

Public transportation providers and other institutions who collect transportation data should consider publishing their performance measures and, to the extent possible, their data, because this can lead to (1) increased transparency and public involvement, (2) the development of data access and visualization tools by enthusiasts at no cost to the agency, and (3) development of insights by researchers working with the data. Institutions should make human resource investments that increase the capacity to manage and utilize data, and they should be open to adapting their decision-making processes in order to make more objective and data-informed decisions, even when this implies organizational changes.

The gap between academics and transit agencies can be closed in a number of ways. Researchers can package their work and the tools they develop in a way that is more attractive to practitioners, and they can emphasis communicating with practitioners rather than only other academics (e.g. through journal publications and academic conferences). Tools that work when data is provided in standard format can encourage agencies to adopt data standards in order to benefit from available tools, and to open their data to the public so that researchers and developers can add value. Governments can contribute by encouraging managers to meet researchers and learn about new tools and methods through professional capacity development programs, and by funding graduate research programs in public transportation and then hiring the graduates, who will contribute to the agency the ability to work with big data to solve challenging problems and improve processes.

9. Research recommendations

The workshop's participants, coming from Canada, Chile, Japan, Singapore, United Kingdom, Uruguay, and United States of America, found the workshop productive and useful. They recommended further opportunities for collaborative thinking, perhaps benefitting from the participation of stakeholders absent in the discussion, such as private bus operators, city planners, politicians, shared bicycle and car businesses, data scientists, information technology managers in public transport agencies, and traffic controllers, among others, whose views could only be considered by reference and anecdotes. The conversation started in this workshop should be continued, and perhaps it should be brought to associations such as SIBRT and UITP to encourage more participation from transit agencies.

The final conclusion of our workshop is that transportation researchers and practitioners should do more on big data: more research, more applications, and more discussion. In addition to continued work in models to extract knowledge from historical data, the area of predictive transportation models that consume big data is largely unexplored and merits more attention from the research community. Given the complexities discussed above, it is clear that this is not an easy task; however, it can be a rewarding one, because our understanding of traveller behavior and the operation of our systems can be greatly improved through the good use of available data and the information contained within.

References

Goulet-Langlois, G. (2015). *Exploring regularity and structure in travel behavior using smart card data*. Master thesis. Massachusetts Institute of Technology.

Machlab, F. (2014). *A methodology for identifying potential locations for bus priority treatments in the london network*. Master thesis. Massachusetts Institute of Technology.

Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport origin-destination matrix from passive smart card data from Santiago, Chile. *Transportation Research C: Emerging Technologies, 24*, 9–18.

Ortega-Tong, M. A. (2013). *Classification of London's public transport users using smart card data*. Master thesis. Massachusetts Institute of Technology.

Sánchez-Martínez, G. E. (2015). *Real-time operations planning and control of high-frequency transit*. Ph.D. dissertation. Massachusetts Institute of Technology.

Viggiano, C., Koutsopoulos, H. N., & Attanucci, J. (2014). User behavior in multiroute bus corridors: analysis by a web-based survey. *Transportation Research Record, 2418*, 92–99.

Workshop papers

Arriagada, J., Gschwender, A., & Munizaga, M. (2015). *Modelling bus bunching using massive GPS and AFC data*.

Bull, O., & Herrera, J. (2015). *Using GPS data to identify buses skipping formal stops*.

Danes, C., & Munoz, J. C. (2015). *Public transport's reliability: The case of Santiago, Chile*.

Gschwender, A., Munizaga, M., & Simonetti, C. (2015). *Using smart card and GPS data for policy and planning: The case of Transantiago*.

Giesen, R., Julio, N., & Lizana, P. (2015). *Real-time prediction of bus travel speeds using traffic shockwaves and machine learning algorithms*.

Hemily, B. (2015). *Big data vs. little data? Perspectives and challenges for public transportation agencies to use its data for planning and management*.

Nakamura, T., Uno, N., Nakamura, N., Schmöcker, J.-D., & Iwamoto, T. (2015). *Urban public transport mileage cards: Analysis of their potential with smart card data and SP survey*.

Pan, D., Sun, G., & Looi, T.-S. (2015). *Impact of integrated developments on transfer*.

Pineda, C., Schwarz, D., & Godoy, E. (2015). *Comparison of passengers' behavior and aggregate demand levels on a subway system using origin-destination surveys and smartcard data*.

Sánchez-Martínez, G. E., Koutsopoulos, H. M., & Wilson, N. H. M. (2015). *Optimal allocation of vehicles to bus routes using automatically collected data and simulation modeling*.

Shimamoto, H., & Kondo, A. (2015). *Analysis of the relationship between the public transit fare structure and passenger behaviour using a smart card data*.

Tamblay, S., Galilea, P., & Muñoz, J.-C. (2015). *Estimation of a zonal origin-destination matrix from observed public transport trips for Santiago de Chile*.

Valdez, A. M., Potter, S., Cook, M., & Miles, J. (2015). *Exploring crowdsourcing approaches to big data in Milton Keynes*.

Viggiano, C., Koutsopoulos, H. N., Attanucci, J. P., & Wilson, N. H. M. (2015). *Identifying opportunities for bus service expansion using automatically collected data*.