Research in Transportation Economics 69 (2018) 615–620

Contents lists available at ScienceDirect

Research in Transportation Economics

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

# Workshop 8 report: Big data in the digital age and how it can benefit public transport users

Menno Yap[a], Marcela Munizaga[b,*]

[a] Delft University of Technology, Delft, The Netherlands
[b] Universidad de Chile, Santiago, Chile

ARTICLE INFO

*JEL classification:*
R40
R41
R42

*Keywords:*
Big data
Public transport
Smart card data
User perspective

ABSTRACT

This paper synthesizes evidence from Workshop 8 'Big data in the digital age and how it can benefit public transport users' of the 15th International Conference on Competition and Ownership in Land Passenger Transport. Big data in public transportation has increasingly attracted the attention from both scientists and practitioners, resulting in an increasing number of scientific studies and practical applications in this field. However, compared to the scientific developments, we see that practical big data applications are relatively limited, and that these are applied with a relatively low pace. This indicates that big data has not been used to its full potential in practice yet, meaning that public transport passengers currently do not fully benefit from the opportunities big data offers in terms of public transport quality and attractiveness. Based on literature study and input gained from a qualitative expert session with scientists, public transport authorities, public transport operators and transport consultants together during the conference workshop, we come to the conclusion that the challenges to stimulate further and faster use of big data in practice are institutional rather than technical. This complexity results from required coordination and cooperation among public and private entities that are not always aligned. A framework has been proposed with four components to stimulate a further and faster adoption of big data in practice, directing to different stakeholders or relations between stakeholders: align technical ambitions of big data applications with their institutional environment; enable/ease the use of big data by PT authorities by developing common definitions, data standards and consolidation; incorporate the use of big data by PT operators in the contract between authority and operator; quantify and visualize the business value of big data for PT operators. We illustrate our framework by successful case studies in Chile, the Netherlands and Sweden.

## 1. Introduction

In the last decade many developments have taken place in what has been called the era of big data. This has also been affecting the public transport sector. Big data in public transportation has increasingly attracted the attention from both scientists and practitioners, resulting in an increasing number of scientific studies and practical applications in this field. A variety of data sources are used in these studies and applications, such as smart card data resulting from Automated Fare Collection (AFC) systems, GPS and Automated Vehicle Location (AVL) data, information about occupancies based on Automated Passenger Count (APC) data, Wi-Fi information and mobile phone data. Over the recent years, scientific research has further evolved, leading to more advanced and complex big data studies, which potentially can benefit the quality and attractiveness of public transport for its users. Also in the public transport industry, big data is increasingly used. However, compared to the scientific developments, we see that practical big data applications are relatively limited, and that these are applied with a relatively low pace. This indicates that big data has not been used to its full potential in practice yet, meaning that public transport passengers currently do not fully benefit from the opportunities big data offers in terms of public transport quality and attractiveness.

Already in 2016, Sanchez-Martinez and Munizaga (2016) stressed the importance of the further development and dissemination of tools and use cases for big data, to stimulate the transport industry to adopt the application and development of big data. The main question we address in this research is therefore how big data in this digital age can further benefit public transport users, and how to stimulate further and faster big data applications in the public transport industry. We address this question based on scientific literature and based on input gained

* Corresponding author.
*E-mail addresses:* M.D.Yap@TUDelft.nl (M. Yap), mamuniza@ing.uchile.cl (M. Munizaga).

https://doi.org/10.1016/j.retrec.2018.08.008

Available online 03 September 2018

from a qualitative expert session with scientists, public transport authorities, public transport operators and transport consultants together during the 15th Thredbo International Conference Series on Competition and Ownership in Land Passenger Transport held in August 2017 in Stockholm, Sweden. To be able to address this question, a clear definition of big data is needed. Although big data can be considered a catch-all with a variety of definitions existing, in line with Sanchez-Martinez and Munizaga (2016) we consider big data as 'large amounts of data resulting from increased automation, improvements in sensing, storage and communication technologies, exceeding capabilities to use and understand with traditional tools and methods'. This means that what is being considered big data is dynamic, can change over time and is discipline-dependent, depending on the baseline dataset size which can be handled by tools and methods being used in a certain discipline at a certain time.

This paper is structured as follows. Section 2 provides a non-exhaustive overview of some state-of-the-art studies and applications of big data in public transport. Section 3 identifies four key components which can contribute to bridge the gap between scientific research and practical applications related to further and faster adoption of big data. Based on this, we introduce three case studies in section 4 of this paper which were presented in the workshop and are illustrative of how some of these four components can be addressed, resulting in public transport users further taking the benefit from the big data developments taking place. Conclusions and policy recommendations are formulated in section 5.

## 2. Literature

In this section we show some examples of state-of-the-art research to public transport systems in which big data is used. Given the large amount of studies existing around this topic, the aim is not to address all study areas and methods, but mainly to give a flavour of the range of big data applications which can currently be found.

Data from AFC systems is widely used in relation to origin-destination (OD) matrix estimations. Many studies can be found focusing on inference of passenger destinations, in case of AFC systems where passengers only have to tap in with their smart card, or in case passengers forget to tap out in case of entry-exit AFC systems (e.g. Munizaga & Palma, 2012; Nunes, Dias, & eCunha, 2016; Trépanier, Tranchant, & Chapleau, 2007; Zhao, Rahbee, & Wilson, 2007). Also transfer inference is a topic widely addressed in science, aiming to infer whether a passenger alighting is considered a transfer or final destination of the public transport journey. These algorithms range from using relatively simple criteria (e.g. a maximum time threshold between a passenger tap out and next tap in), to more complex behavioural criteria, and address generic travel patterns or specific scenarios such as behaviour during public transport disruptions (see for example Devillaine, Munizaga, & Trepanier, 2012; Gordon, Koutsopoulos, Wilson, & Attanucci, 2013; Munizaga, Devillaine, Navarrete, & Silva, 2014; Seaborn, Attanucci, & Wilson, 2009; Sánchez-Martinez, 2017; Wang, Attanucci, & Wilson, 2011; Yap, Cats, Van Oort, & Hoogendoorn, 2017a).

While the studies mentioned above mainly focusing on improving OD matrix estimation, other studies focus on data fusion, travel pattern identification or visualization. For example, Nijënstein and Bussink (2015) perform a fusion of AFC data from different public transport operators; De Regt, Cats, Van Oort, and Van Lint (2017) fuse smart card data and mobile phone data to identify latent public transport demand. Fusion and visualization of AFC, APC and GTFS data is for example applied by Giraud, Légaré, Trépanier, and Morency (2016). Also unsupervised learning methods such as k-means, hierarchical clustering and DBSCAN are applied to AFC data to identify, classify and visualize spatial and/or temporal journey patterns (e.g. Agard, Morency, & Trépanier, 2007; Briand, Come, Trépanier, & Oukhellou, 2017; Cats, Wang, & Zhao, 2015; El Mahrsi, Come, Oukhellou, & Verleysen, 2017; Luo, Cats, & Van Lint, 2017; Ma, Wu, Wang, Chen, & Liu, 2013).

Other big data applications relate to better prediction of public transport ridership based on AFC and AVL data. Idris, Habib, and Shalaby (2015) improve mode choice models based on revealed preference data obtained from AFC data. Van Oort, Brands, and De Romph (2015) developed an elasticity-based public transport ridership prediction model based on smart card data, to predict public transport mode and route choice after network changes. Yap, Nijenstein, and Van Oort (2018) further calibrate the parameters of this model to predict passenger behaviour specifically during planned track closures and disruptions, based on AFC data obtained during several previous track closures. Several studies adopt machine learning approaches for short-term passenger predictions. For example, Wei and Chen (2012) use neural networks for short-term metro ridership predictions. Ding, Wang, Ma, and Li (2016) use gradient boosting decision trees to predict metro ridership, and Li, Wang, Sun, and Ma (2017) apply a multiscale radial basis function (MSRBF) network for ridership prediction under special occasions.

## 3. Key components for big data adoption in practice

After addressing some state-of-the-art research in the field of big data and public transport, in this section we formulate four key components to stimulate a further and faster adoption of big data in public transport practice. These components are directed towards different stakeholders relevant in the industry and their institutional

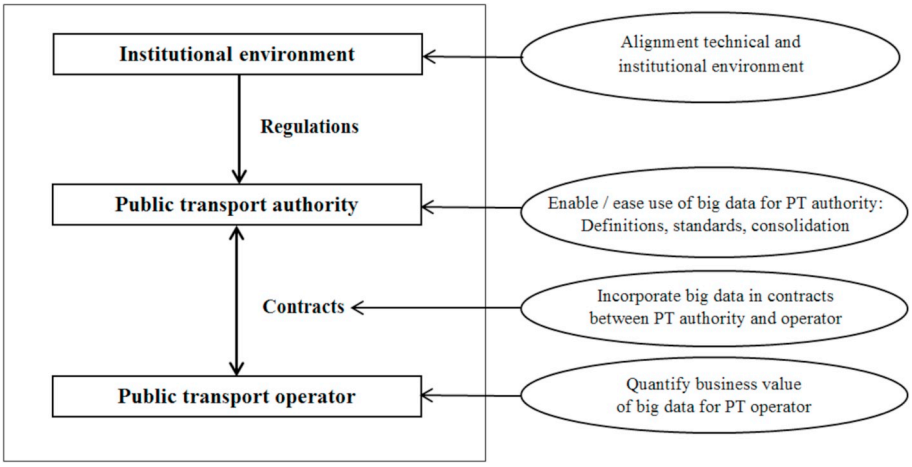


**Fig. 1.** Framework with key component to stimulate big data adoption in the public transport industry.

environment, and summarized in the framework we developed as shown in Fig. 1. In this framework, the left-hand side shows the public transport authority (PTA), operating within a certain institutional environment, and the public transport operator (PTO). From this institutional environment in a specific country certain regulations, policies and culture result, which give directions to the way the PTA works. One hierarchical level lower, there is the relation between the public transport authority and the public transport operator. Although not the case for all countries, it has become standard practice for many countries to separate the PTA and PTO as different stakeholders. This relation is formalized based on a contract between PTA and PTO, for example resulting from a tendering process, which specifies the PT supply the PTO must deliver against certain pre-defined KPIs. The four key components to stimulate further and faster big data adoption in practice are shown in the right-hand side of Fig. 1. These components affect the public transport system via different stakeholders or relations between stakeholders, operating on different hierarchical levels. The framework shows the importance of an integrated approach to really fasten big data adoption in practice, where attention should be given to all different stakeholders, their institutional environment, and the regulations and contracts relating stakeholders and their environment.

### 3.1. Governance: alignment between technical and institutional environment

Cities worldwide are keen on exploring ways to use mobility platforms using big data to better predict current and future vehicle and passenger flows, in order to improve accessibility, liveability and sustainability. Veeneman, Van der Voort, Hirschorn, Steenhuisen, and Klievink (2017) studied 10 mobility platforms, and pilot projects in Haifa, Rome and Venice, with the aim to identify key governance mechanisms that affect the success of mobility platforms using big data. They concluded that the more ambitions a mobility platform has, the more governance challenges arise due to several mechanisms. A higher technical ambition level leads to more misalignment with existing institutions, especially if this higher ambition level requires the use of more personal data in relation to privacy regulations. A higher ambition level of a mobility platform regarding the number of goals to realize and the number of stakeholder to get involved, makes it more difficult to define and realize consensus about the directions of the platform, and increases the need for a clear governance structure and coordination.

The study performed by Veeneman et al. (2017) shows that substantial differences exist in both technical ambition level and institutional environment for different mobility platforms over the world. For example, there can be substantial differences regarding privacy regulation for the use of big data between countries. They also state that the institutional context should mainly be considered an ‘as-is’ state. Technical ambitions should be aligned with the given institutional environment the mobility platform is positioned in, to successfully institutionalize the use of big data in mobility platforms. For researchers to have their state-of-the-art work adopted by practitioners, it is thus essential to be aware of the institutional environment a specific PTA or PTO operates in. Next to the traditional focus researchers tend to have on technical ambition and innovation, for a successful adoption in practice it is therefore necessary that researchers make sure the methodology and data used in a study can be institutionalized.

### 3.2. Enable/ease use of big data by public transport authorities: definitions, standards, consolidation

Once technical ambitions and the institutional environment regarding big data are aligned, the second component to further stimulate big data adoption in practice relates to enabling and easing the use of big data by public transport authorities. As illustrated in section 2.1, there are many scientific applications of big data. These applications are however relatively scattered, performed in different institutions, different countries and by different researchers applying their own definitions, data formats, methods and software. In order to ease the use of big data for PTAs, it is important for the scientific community to develop common definitions and standard data formats, together with the industry. We therefore urge to define and further standardize the data formats and variables resulting from AFC, AVL and APC systems.

Besides, valuable information can be extracted from big data once different data sources are fused. Fusion of AFC and AVL data can for example result in vehicle occupancies, without the need of having a separate APC data source (see for example Yap, Cats, Yu, & Van Arem, 2017b). As another example, fusion of AFC and mobile phone data can provide information about the share of public transport in the modal split between specific parts of the city (see for example De Regt et al., 2017). While fusion of data sources is commonly applied by several researchers, consolidation is not always common practice in the public transport industry yet. Agreeing on common data formats also eases consolidation of data sources and can support PTAs to get more value from big data. We therefore recommend the development of standard definitions and data formats for commonly consolidated datasets, such as AFC-APC, AVL-APC and AFC-AVL data fusions.

### 3.3. Incorporate big data in contracts between public transport authority and public transport operator

The third component to stimulate big data adoption in practice relates to incorporating the use of, and access to, big data in contracts between public transport authority and operator. Given the different institutional environments in different countries, access to different big data sources differs between countries as well. Most often, AVL and GTFS data containing information about public transport schedules, locations and realized arrival and departure times, is publicly available. Data availability and data ownership regarding AFC data however differs more between countries and also depends on agreements specified in contracts between PTA and PTO. For example, in the Netherlands AFC data is owned by the PTO. Whether PTAs have access to this data, merely depends if the contract specifies that the PTO needs to share AFC data with the PTA. Such sharing agreements can stimulate the use of big data by PTAs, so more of its potential can be used. Given the often relatively long duration of a tender, it is therefore recommended that the requirement to share AFC data with the PTA is explicitly added in tendering documents as well.

Besides, big data use can be further stimulated if contracts between PTA and PTO require usage and innovations of big data. For example, in these contracts KPIs could be specified which could or should be obtained from big data. Continuous innovation using big data is another aspect which is recommended to incorporate in contracts. The fast pace with which technological developments take place can be contradicting with the usually relatively long duration of contracts between PTA and PTO. It can be challenging for a public transport authority to incorporate KPIs and requirements in a contract that is valid for 10–15 years, which are robust and flexible enough to keep stimulating big data innovation throughout the whole duration of the contract. We therefore recommend designing contracts that incentivize big data innovation during the execution of a concession, in order to align practical operations with technological developments and to keep stimulating big data innovations.

### 3.4. Quantify the business value of big data for public transport operators

The fourth component aims to stimulate big data use by public transport operators, by quantification and visualization of the business value of big data. Especially in more deregulated environments, public transport operators generally align their performance with the KPIs as specified in the contract with the public transport authority. Therefore, it is important to show how big data can contribute to achieving or improving these KPIs. Next to studies using big data to show direct

passenger or societal benefits, there can be potential for big data studies that contribute to operator efficiency or increase operator revenues. For example, if big data applications can better align PT services with PT demand, this might save the operator redundant capacity, or can improve the crowding level and perceived quality of public transport on busy tracks, which in turn can increase public transport ridership and operator revenues.

In many contracts between PTA and PTO, KPIs related to customer satisfaction are incorporated. From the expert session, a clear demand arose to better quantify and predict customer satisfaction, since this is currently considered a difficult KPI to predict due to the discrepancy between objective performance (such as punctuality) and subjective perception by passengers. There is an opportunity to show the business value of big data, if big data studies can contribute to better predict customer satisfaction.

At last, a clear need for visualization of the value of big data outputs can be observed. Next to a mathematical quantification of the value of big data outputs, there is a need to show this business value in clear visualizations to, for example, managers, authorities or policy makers. Investments in systems to handle big data can be expensive and decisions for this are often not made by the public transport planners themselves. This means there is a necessity to clearly communicate and illustrate the business value of big data, as contrast to required costs for systems to handle these amounts of data.

## 4. Case studies

In this section we introduce three case studies of big data applications in Chile, the Netherlands and Sweden, being illustrative how can be contributed to the four components addressed in our framework to stimulate that big data is being used more and at faster pace in public transport practice.

### 4.1. Case study Santiago, Chile

Over the last years, several methods have been developed and applied to data from Transantiago, the public transport system in Santiago de Chile. In the previous Thredbo conference, Gschwender, Munizaga, and Simonetti (2016) presented a synthesis of the developments and big data applications made up to that moment, that included bus speeds estimations, OD matrix estimations, and quantification of level of service indicators. Most importantly however, they describe “a successful experience of collaboration between academia and the public transport authority to develop tools based on passive data processing”. In the expert session held during Thredbo 15, four applications of passive data processing are presented that use the Transantiago data.

In Santiago many passenger complaints are related to buses skipping a scheduled bus stop, while passengers were waiting for that bus. To make public transport services more attractive for users, it is therefore important for the PTA to determine how often this occurs. In the contract between PTA and PTO, it is stated that the PTO is fined if a bus incorrectly skips a scheduled bus stop in such cases. While the PTA performed several manual inspections resulting in fines, in a city with such a large bus network as Santiago - having 11,000 bus stops - it is impossible to perform a structural, systematic check on this manually. However, with the availability of GPS data of each bus every 30 s, big data has potential to contribute to the quantification of this KPI as specified in the contract between PTA and PTO.

Garcia and Herrera (2017) study models which identify buses which incorrectly skipped formal bus stops based on GPS data. Next to existing models, they develop a machine learning model based on Support Vector Machines to improve the prediction accuracy of these models. In this study, GPS data of 371 buses around a case study stop is gathered. Based on this data, a kinematic-based model and two machine learning models – Support Vector Machines and Linear Discriminant Analysis – are developed to predict if a bus incorrectly skipped a stop based on time- and position information from GPS data. The study results during the model validation showed that the prediction accuracy of the kinematic-based model was about 86%, whereas this was about 89–90% for both machine learning methods. Both machine learning methods also slightly outperform the kinematic-based model in terms of the percentage buses incorrectly identified as ‘incorrect stop skipping’ while they were actually stopping, and incorrectly identified as ‘stopping’ while in reality they incorrectly skipped the stop.

This approach can provide the PTA insights how often stops are skipped in an automated way using big data, which can be used to quantify the pre-defined KPI in the contract between authority and operator. This case study is therefore exemplary for how big data could be incorporated in contracts to quantify a certain KPI, which would be very difficult to fully quantify by inspections only.

Bus speeds prediction is valuable for the PT operator. It can support prioritizing mitigation measures at locations where the largest bottlenecks for PT occur, or operators can anticipate their supply, which possibly results in higher punctuality, customer satisfaction, and an increase in PT ridership as consequence.

Schmidt, Moya, Cruz, and Munoz (2017) use AVL data to identify and visualize operational bottlenecks for public transport lines. Based on previous research, they build on a method proposed by Bucknell, Schmidt, Cruz, and Munoz (2017), using more detailed and more frequent GPS data. They incorporate queue length for identified bottlenecks, and incorporate bus occupancies in prioritizing the bottlenecks. Berczely and Giesen (2017) develop different models to better predict bus travel speeds, to better align scheduled and realized travel times and thus to improve public transport reliability. They compare three machine learning based models, Multiple Linear Regression, Support Vector Machines and Neural Networks, with traditional benchmark models. Results showed that all machine learning models reduce the prediction error up to 10–25% compared to the traditional model, whereby the use of Neural Networks outperforms the other two machine learning models. Cubillos and Munizaga (2017) develop a model to evaluate the impact of different bus priority schemes on bus travel speed and time using big data. Based on AVL data from buses of Santiago, Chile, a bus travel time prediction model is developed, among others incorporating dummy variables reflecting if there is mixed traffic, a median busway or a separated one or two lane bus segment. Their results show that median busways are predicted to realize the largest reduction in bus travel times.

These applications show how the collaboration between the academia and the public transport authority can contribute to the alignment between technical and institutional environment. It also shows how the business value of big data for a PT operator can be quantified and visualized. The studies contribute to better regularity and reliability of bus services, if travel times can be predicted more accurate, bottlenecks can be better identified, and if the effectiveness of measures aiming to improve bus speed and reliability can be compared. This improves the public transport service provided to passengers and can result in increased operator revenues from ridership increases, but also improve the operator performance on pre-defined KPIs between authority and operator in relation to punctuality, regularity or customer satisfaction. The applications presented here facilitate the use of big data by the public transport authority and by the operators.

### 4.2. Case study The Hague, the Netherlands

This case study, presented in Van Oort, Brands, De Romph, and Yap (2016) and Yap et al. (2017b), considers a possible frequency increase on the urban public transport network of the city of The Hague, the Netherlands. Specifically, it relates to a busy tram line in this city, currently operating 6 trams per hour per direction, of which the PTO considers a frequency increase in morning and evening peak to 8 trams per hour. From the operator perspective, it is important to have a business case with expected additional operating costs and revenues to

support this decision. Traditional ridership prediction models, such as elasticity-based approaches, do incorporate the impact of reduced waiting times of such frequency increase on ridership. However, ridership increase is also expected from a reduced crowding level in case of a frequency increase, resulting in a less negative perception of in-vehicle time. In this case study, the aim is to use big data to predict the impact of this reduced crowding level on public transport ridership. This contributes to the fourth component as mentioned in chapter 2, by quantifying and visualizing the business value of big data to the public transport operator.

In Yap et al. (2017b) revealed preference data obtained from AFC and AVL systems is used to estimate a Mixed Logit model with panel effects to infer how passengers value crowding in urban tram and bus transportation. Contrary to traditional stated preference (SP) experiments which tend to overestimate estimated coefficients, this study uses revealed preference (RP) data obtained from passive data source to estimate crowding valuation. Only OD pairs with more than one observed route alternative chosen are incorporated in the model estimation. The attribute levels in relation to travel time are obtained from AVL data, whereas attribute levels related to crowding are obtained by fusion of AFC and AVL data. Based on observed route choices resulting from AFC data, crowding valuation has been estimated using a maximum likelihood estimation. Study results show that crowding valuation obtained from SP studies tend to overestimate crowding valuation in public transport up to 40%, compared to our RP-based values.

The results of the case studies revealed the predicted ridership increase of this frequency increase without and with incorporation of benefits from reduced crowding levels. As Fig. 2 visualizes, ridership benefits would be underestimated substantially if crowding was not incorporated. For the AM and PM peak period, the ridership and operator benefits would be underestimated by 30% and 20% respectively, given the more concentrated morning peak compared to the evening peak in the case study area. This case study is illustrative how research to crowding valuation based on big data rather than SP experiment is translated to quantified and visualized ridership increases, which allows calculation of the expected revenue increases for the PT operator.

### 4.3. Case study Stockholm, Sweden

Another study illustrating the business value of big data for public transport operators is undertaken by Jiang and Brundell-Freij (2017) in Stockholm, Sweden. In this study Jiang and Brundell-Freij (2017) use big data to estimate price sensitivities of different public transport passenger segments in Stockholm, showing to be highest for youngster and senior segments. Based on this, they estimate a MNL model for card-type choice for these different segments. Compared to traditionally used stated preference experiments, using revealed preference data in the model estimation does not entail the risk of bias between stated and realized behaviour. Therefore, revealed preference data from AFC systems provides an opportunity to determine passenger preferences and sensitivities more accurately. Results of this study have direct value for PT operators by creating a better understanding of price sensitivities and ticket preferences of different passenger segments.

## 5. Conclusions and recommendations

Based on literature study and a three day expert session with scientists, public transport authorities, public transport operators and transport consultants together, we come to the conclusion that the challenges to stimulate further and faster use of big data in practice - to improve the quality of public transport for the user - are institutional rather than technical. In a time of fast technological developments, the technical challenges can be solved easier, whereas solving institutional challenges tend to be much more complex. This complexity results from required coordination and cooperation among public and private entities that are not always aligned. A framework has been proposed with four components to stimulate a further and faster adoption of big data in practice, directing to different stakeholders or relations between stakeholders:

- Align technical ambitions of mobility platforms using big data with their institutional environment;
- Enable and ease the use of big data for PT authorities by developing definitions, common standards and formats for big data sources and commonly consolidated datasets;
- Require PT operators in contracts between PT authority and operator to share big data with the PT authority, and require usage and innovation of big data by the operator in the specification of KPIs;
- Show and visualize the business value of big data for the PT operator.

We can formulate several policy recommendations when aiming to further adopt big data in the public transport industry. First, public agencies such as public transport authorities and operators should actively identify priorities and opportunities for big data applications, to give direction to the topics of scientific research. Second, as mentioned previously, we urge the development of standards, common definitions and data formats for separate and consolidated, integrated data sources in close cooperation between scientific researchers and practitioners. We also recommend an incremental, step-by-step approach when further adopting big data in the public transport industry. This gives public entities time to learn from their experiences in a controlled

**Fig. 2.** Comparison between predicted ridership effects of a frequency increase without crowding (left) and with crowding (right) effects incorporated (derived from Van Oort et al., 2016).

environment, and time to attract sufficient qualified people to their organizations to transfer information from big data to knowledge. Reporting and sharing these experiences among the industry is essential, so that stakeholders can look at successful cases and learn from unsuccessful cases. Building trustworthy relations between stakeholders is important, but at the same time it is relevant to secure data sharing between stakeholders, for example between PT operator and authority, to make sure the passenger and societal perspective prevails. At last, we recommend looking at experiences with big data applications in other sectors than public transport to learn from (un)successful cases and experiences.

Following our study we also formulate contracting recommendations to stimulate big data adoption in practice via contracts between PT authority and PT operator. First, it is recommended that PT authorities guarantee access to (integrated) big data sources in these contracts between operator and authority. Second, incorporating the use of big data in the definitions of KPIs can be valuable to stimulate and guarantee the use of big data by the operator. Third, it is recommended to design contracts between PT authority and operator in such manner that incentives are provided to the operator to perform big data innovations during the execution of a contract. This can reduce the tension between the institutional environment with usually a relatively long duration of PT contracts on the one hand, and the fast pace by which technological developments take place on the other hand.

## Acknowledgments

We are indebted to all participants who contributed actively to the workshop discussions. Workshop participants included Clas Artvin, Helena Björn, Ricardo Cubillos, Ricardo Giesen, Mattias Haroldsson, Juan Carlos Herrera, Sida Jiang, Jonathan Leishman, Jaime Moya, Juan Carlos Munoz, Alejandro Schmidt, Wijnand Veeneman, and Björn Westerberg. This research was funded by NWO grant agreement 438.15.404/298 as part of JPI Urban Europe ERA-NET CoFound Smart Cities and Communities initiative (TRANS-FORM project), by Fondecyt Chile (Grant 1161589) and by the Complex Engineering Systems Institute Chile (CONICYT – PIA – FB0816).

## References

### Workshop papers

Berczely, J., & Giesen, R. (2017). *Machine learning methods to predict bus travel speeds and analysis of the impact of different predictive variables*.

Cubillos, R., & Munizaga, M. (2017). *Bus travel time model for different bus priority schemes using massive data for the city of Santiago*.

Garcia, N., & Herrera, J. C. (2017). *Assessment of models based on GPS data to identify buses skipping formal stops*.

Jiang, S., & Brundell-Freij, K. (2017). *Development of smart card data (SCD) approach for travel demand characterization – a Stockholm case study*.

Schmidt, A., Moya, J., Cruz, D., & Munoz, J. C. (2017). *Identifying and visualising operational bottlenecks for public transport considering queue length, bus load profiles and a sharper data source*.

Veeneman, W., Van der Voort, H., Hirschorn, F., Steenhuisen, B., & Klievink, B. (2017). *PETRA: Governance as a key success factor for big data solutions in mobility*.

Yap, M. D., Cats, O., Yu, S., & Van Arem, B. (2017b). *Crowding valuation in urban tram and bus transportation based on smart card data*.

### *Other*

Agard, B., Morency, C., & Trépanier, M. (2007). *Mining public transport user behaviour from smart card data.* CIRRELT-2007-42, Canada.

Briand, S. A., Come, E., Trépanier, M., & Oukhellou, L. (2017). Smart card clustering to extract typical temporal passenger habits in transit networks. Two case studies: Rennes in France and Gatineau in Canada. *3rd international workshop and symposium: 'Research and applications on the use of passive data from public transport', may 2017, Santiago, Chile*.

Bucknell, C., Schmidt, A., Cruz, D., & Munoz, J. C. (2017). Identifying and visualizing congestion bottlenecks with automated vehicle location systems: An application in Transantiago. *Transportation Research Record: Journal of the Transportation Research Board, 2649*, 61–70.

Cats, O., Wang, Q., & Zhao, Y. (2015). Identification and classification of public transport activity centers in Stockholm using passenger flow data. *Journal of Transport Geography, 48*, 10–22.

De Regt, K., Cats, O., Van Oort, N., & Van Lint, H. (2017). Investigating potential transit ridership by fusing smartcard data and GSM data. *Transportation Research Record: Journal of the Transportation Research Board, 2652*, 50–58.

Devillaine, F., Munizaga, M. A., & Trepanier, M. (2012). Detection of activities of public transport users by analyzing smart card data. *Transportation Research Record, 2276*, 48–55.

Ding, C., Wang, D., Ma, X., & Li, H. (2016). Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. *Sustainability, 8*, 1–16.

El Mahrsi, M. K., Come, E., Oukhellou, L., & Verleysen, M. (2017). Clustering smart card data for urban mobility analysis. *IEEE Transactions on Intelligent Transportation Systems, 18*, 712–728.

Giraud, A., Légaré, F., Trépanier, M., & Morency, C. (2016). *Data Fusion of APC, smart Card and GTFS to visualize public transport use: CIRRELT-2016-54, Canada*.

Gordon, J. B., Koutsopoulos, H. N., Wilson, N. H. M., & Attanucci, J. P. (2013). Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. *Transportation Research Record: Journal of the Transportation Research Board, 2343*, 17–24.

Gschwender, A., Munizaga, M. A., & Simonetti, C. (2016). Using smartcard and GPS data for policy and planning: The case of Transantiago. *Research in Transportation Economics, 59*, 242–249.

Idris, A., Habib, K., & Shalaby, A. (2015). An investigation on the performances of mode shift models in transit ridership forecasting. *Transportation Research Part A, 78*, 551–565.

Li, Y., Wang, X., Sun, S., & Ma, X. (2017). Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks. *Transportation Research Part C, 77*, 306–328.

Luo, D., Cats, O., & Van Lint, H. (2017). Constructing transit origin-destination matrices using spatial clustering. *Transportation Research Record, 2652*, 39–49.

Ma, X., Wu, Y., Wang, Y., Chen, F., & Liu, J. (2013). Mining smart card data for transit riders' travel patterns. *Transportation Research Part C, 36*, 1–12.

Munizaga, M. A., Devillaine, F., Navarrete, C., & Silva, D. (2014). Validating travel behaviour estimated from smart card data. *Transportation Research Part C, 44*, 70–79.

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

Nijënstein, S., & Bussink, B. (2015). Combining multimodal smart card data: Exploring quality improvements between multiple public transport systems. *European transport conference, Germany*.

Nunes, A. A., Dias, T. G., & eCunha, J. F. (2016). Passenger journey destination estimation from automated fare collection system data using spatial validation. *IEEE Transactions on Intelligent Transportation Systems, 17*, 133–142.

Sánchez-Martinez, G. E. (2017). Inference of public transportation trip destinations by using fare transaction and vehicle location data. Dynamic programming approach. *Transportation Research Record, 26*.

Sanchez-Martinez, G., & Munizaga, M. (2016). Workshop 5 report: Harnessing big data. *Research in Transportation Economics, 59*, 236–241.

Seaborn, C., Attanucci, J., & Wilson, N. H. M. (2009). Analyzing multimodal public transport journeys in London with smart card fare payment data. *Transportation Research Record: Journal of the Transportation Research Board, 2121*, 55–62.

Trépanier, M., Tranchant, N., & Chapleau, R. (2007). Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems, 11*, 1–14.

Van Oort, N., Brands, T., & De Romph, E. (2015). Short-term prediction of ridership on public transport with smart card data. *Transportation Research Record: Journal of the Transportation Research Board, 2535*, 105–111.

Van Oort, N., Brands, T., De Romph, E., & Yap, M. D. (2016). In F. Kurauchi, & J. D. Schmöcker (Eds.). *Ridership evaluation and prediction in public transport by processing smart card data:A Dutch approach and example, chapter 11, public transport planning with smart card data*. CRC Press.

Wang, W., Attanucci, J. P., & Wilson, N. H. M. (2011). Bus passenger origin-destination estimation and related analyses using automated data collection systems. *Journal of Public Transportation, 14*, 131–150.

Wei, Y., & Chen, M. (2012). Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. *Transportation Research Part C, 21*, 148–162.

Yap, M. D., Cats, O., Van Oort, N., & Hoogendoorn, S. P. (2017a). Robust transfer inference: A transfer inference algorithm for public transport journeys during disruptions. *20th EURO working group on transportation meeting, EWGT 2017, transport research procedia* (pp. 8). Elsevier.

Yap, M. D., Nijenstein, S., & Van Oort, N. (2018). Improving predictions of public transport usage during disturbances based on smart card data. *Transport Policy, 61*, 84–95.

Zhao, J., Rahbee, A., & Wilson, N. H. M. (2007). Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering, 24*, 376–387.