Research in Transportation Economics 59 (2016) 242–249

Contents lists available at ScienceDirect

Research in Transportation Economics

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

Using smart card and GPS data for policy and planning: The case of Transantiago

Antonio Gschwender [a], Marcela Munizaga [b, *], Carolina Simonetti [a, c]

[a] Directorio de Transporte Público Metropolitano DTPM, Moneda 975, Santiago, Chile
[b] Departamento de Ingeniería Civil, Universidad de Chile, Blanco Encalada 2002, Santiago, Chile
[c] Independent consultant

ARTICLE INFO

Article history:
Received 1 November 2015
Received in revised form
30 April 2016
Accepted 1 May 2016
Available online 23 September 2016

JEL classification:
R4 Transportation Systems

Keywords:
Public transport
Passive data
Automatic vehicle location
Automatic fare collection

ABSTRACT

The introduction in 2007 of a new public transport system in Santiago, Chile, brought to us an unexpected gift: the availability of Big Data; massive amounts of passive data obtained from technological devices installed to control the operation of buses and to administer the fare collection process. Many other cities in the world have experienced the same, and sooner or later, this is likely to happen everywhere. Seeing this opportunity, many researchers have developed tools to obtain valuable information from the available data. However, the case of Transantiago is particularly advantageous because all buses have GPS devices and the smart card presents an overall 97% penetration rate.

We describe a successful experience of collaboration between academia and the public transport authority to develop tools based on passive data processing. We include a brief description of the Transantiago system and the agreements made to develop the aforementioned tools. We also describe the methods developed to obtain valuable information like public transport trips origin-destination matrices, speed profiles of buses and service quality indicators, among others. Several examples of specific uses of the information for public transport policy and planning in Santiago are presented. The paper concludes with a discussion of what else can be obtained from this data and why we believe that this can change the way we do transport planning.

1. Introduction

The big changes of the last half century did not come from technological advances such as flying cars and human-looking robots, as was envisioned by some science fiction authors; unexpectedly, the big revolution was in connectivity and availability of data. Never before has so much information about events, people and objects been so widely available (ITF, 2015). This scenario presents important challenges to transport authorities, including how to administer this massive and constant flow of data, and how to make good use of it. In the case of Santiago, Chile, the introduction in 2007 of a new public transport system suddenly began to produce massive amounts of passive data obtained from the technological devices installed to control the operation of buses and to administer the fare collection process. Many other cities in the world have been experiencing the same, and sooner or later, this is likely to happen everywhere. Some researchers have seen this as an opportunity and have developed tools to obtain valuable information from the available data. Pelletier, Trépanier, and Morency (2011) made a literature review that synthesizes development until 2010, identifying contributions that can help at different stages of transport planning: strategic, tactical and operational. However, the case of Transantiago is particularly advantageous because since 2007 all buses are equipped with GPS devices that generate a position record every 30 s and a smart card called "bip!" is the only payment option available at buses and by far the most popular in Metro, implying an overall 97% penetration rate.

In this paper, we describe a successful experience of collaboration between academia and the public transport authority to develop tools based on passive data processing that are useful for planning and operation management. We include a brief description of the Transantiago public transport system and the agreements made to develop the aforementioned tools, including data confidentiality guards and the data access for third parties policy.

* Corresponding author.
E-mail addresses: antonioe.gschwender@dtpm.gob.cl (A. Gschwender), mamuniza@ing.uchile.cl (M. Munizaga), carolina.simonetti@gmail.com (C. Simonetti).

http://dx.doi.org/10.1016/j.retrec.2016.05.004
0739-8859/

We also describe the methods that have been developed to obtain valuable information from the data available: public transport trips origin-destination matrices, speed profiles of buses, service quality indicators and time use patterns. We present some behavioural models that have been developed using the information generated and several practical applications of the information for public transport policy and planning in Santiago. The paper concludes with a discussion of what else can be obtained from this data and why we believe that this can change the way we do transport planning.

2. Context

2.1. Description of the Transantiago data

The public transport system Transantiago, available in Santiago de Chile since 2007, is a multimodal integrated system (bus and metro) that serves a population of 6.6 million inhabitants. Administratively, it is coordinated and supervised by the Metropolitan Public Transport Directory (DTPM), a state agency that depends on the Ministry of Transport and Telecommunication (MTT). The operation of the bus services is contracted out to seven private companies, each of whom owns between 400 and 1200 vehicles. Metro is operated by a publicly-owned company. Moreover, there are four companies that provide other "complementary services" to the system: the financial administration, the smart card management and its sales and charging network, the technological services for the bus companies, and the technological services for the Metro company.

Overall, the system has over 6500 buses all equipped with GPS devices, operating daily in a network that contains 68 km of segregated busways, 150 km of reserved streets or exclusive bus lanes and over 11,000 bus stops. The integrated Metro network has 5 lines, 104 km of rails and 108 stations, and it is currently expanding (DTPM, 2015). The fare scheme is based on trips; a flat fare is applied to trips of up to three stages made within two hours. A small surcharge, larger in peak hours than in other periods, is applied to trips that use the Metro network. The payment system is based on a contactless smart card called "bip!", which is the only way to pay in buses and by far the most important payment method in the Metro, globally accounting for 97% of the payment transactions of the 4.5 million daily trips by public transport. Given the fare structure, tap-off validation is not required in buses or Metro.

The available data come from three different sources. The GPS devices send a position-time observation for each bus every 30s, thus generating 80–100 million observations per week. The bip! smart card transactions generate 35–40 million observations per week. Complimentary information includes route paths, route assignments, position of bus stops, position of Metro stations and position of bus stations. This raw information can be observed over time (transactions) and space (bus movements). Figs. 1 and 2 show the distribution of transactions over time and the distribution of buses in the space at an instant.

2.2. Collaboration between academia and the public transport authority

However, much more relevant information can be obtained from processing these data. In order to do this, after more than a year of exploratory joint work, in 2010, a cooperation agreement was signed between the University (Universidad de Chile, a public institution) and the Public Transport Authority (DTPM[1]). This agreement establishes the possibility of sharing data and methods for mutual benefit. For the University, processing these data was a methodological challenge; therefore, there were researchers willing to dedicate time to explore different aspects of it. For the Public Transport Authority, the possibility of obtaining detailed information (at low cost) for planning and monitoring was also attractive; therefore, they were willing to share the data and dedicate time to this joint work. In 2012, a new stage in this collaboration started, with a three-year project partially funded by FONDEF, a Chilean financial support program for the development of science and technology.[2] The Public Transport Authority co-financed the project, aimed at the consolidation of software with the methodologies developed. This software will allow the Public Transport Authority to autonomously obtain trip and speed information using the methodologies developed in these projects.

The benefits of this type of project for academia come from the availability of interesting data and problems for case studies. This is convenient for the three missions of the University: research, students' formation and extension. Researchers get the opportunity to develop new methods and publish those results in peer-reviewed journals (see for example Cortés, Gibson, Gschwender, Munizaga, & Zúñiga, 2011; Devillaine et al. 2012; Munizaga & Palma, 2012; Munizaga, Devillaine, Navarrete, & Silva, 2014; Gibson et al., 2016). Attractive opportunities for thesis works are generated; for example, as a result of this project, over 10 students graduated from different MSc programs (Transport, Operations Research, Computer Science). In terms of extension, the core of this project is actually to transfer the results to practice. This is shown in Section 4, where we describe in detail the benefits for transport planning and policy. We detail the benefits for the Public Transport Authority in that section, as well.

The cooperation agreements include data confidentiality guards in order to protect the privacy of the information associated to the data, mainly related with the ID of each smartcard. In addition, they consider the possibility of including third parties in the project, as long as this contributes to the objectives of the agreement.

This collaboration has needed at least two types of know-how. On the one hand, the development of the technologies relies on a clear understanding of problems in the field of transportation engineering. On the other hand, for the programming of the tools and the management of the large databases, specific knowledge in computer sciences was indispensable. In addition to the technical work, these types of collaborations are quite time-consuming in terms of their administrative management. As both parties are public agencies, they are subject to the natural state supervision and bureaucracy, which means that there is a constant interaction with lawyers and administrative personal. This issue implies an extra effort for the technicians because of the need to explain the nature and value of the work that was being developed to people who are not transport experts.

The success of this experience of collaboration between Academia and the Public Transport Authority is explained not only by the formal collaboration agreements signed, but also because of a long-term commitment based on mutual confidence among the people working at each side of the table. Interestingly, former students that worked on their degree theses at the University in

[1] Directorio de Transporte Público Metropolitano (www.dtpm.gob.cl).

[2] The Scientific and Technological Development Support Fund (FONDEF) is a program managed by the National Commission for Scientific and Technological Research (CONICYT). FONDEF works within the framework of the National Innovation Strategy of the Government of Chile and follows the long-term guidelines designed by the National Innovation Council. Its mission consists in promoting ties and partnerships among research institutions, corporations and other entities. Its goal is to develop applied research projects that can improve Chile's competitiveness and the quality of life of its population (FONDEF, 2015).

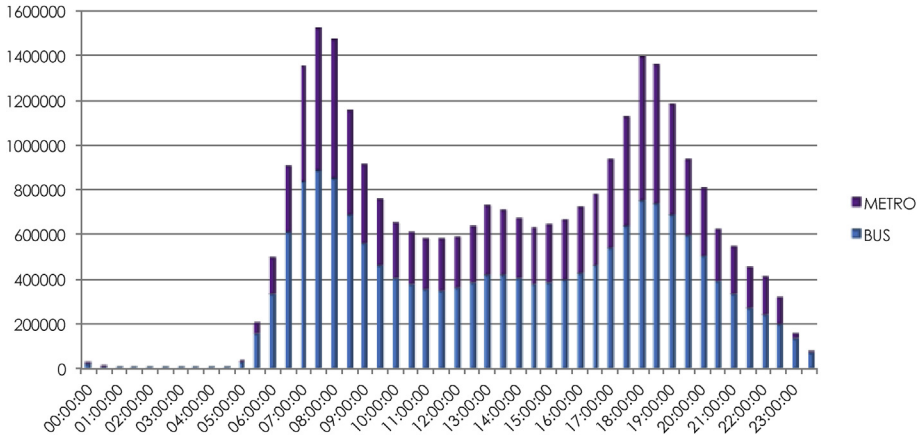

Fig. 1. Distribution of smart card transactions on a Weekday.

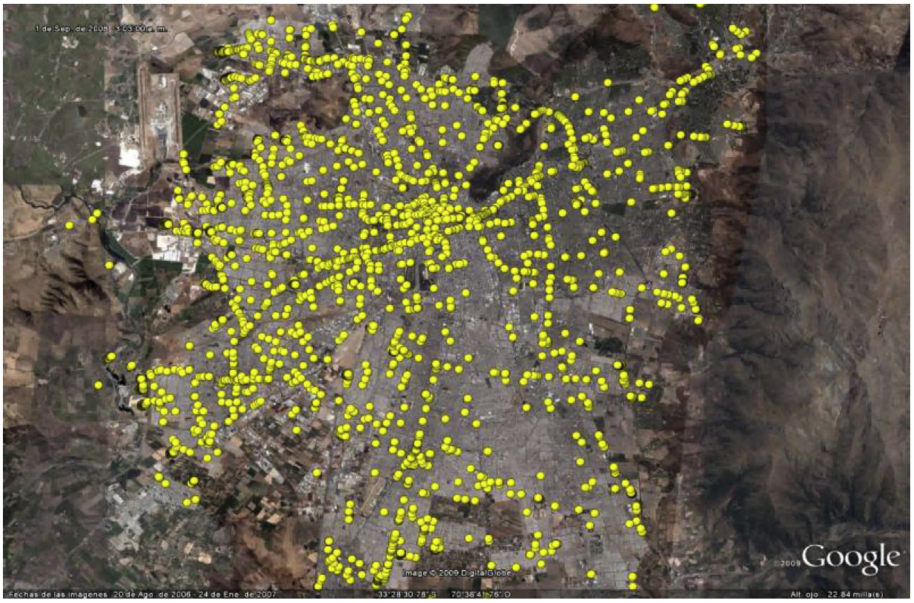

Fig. 2. Distribution of buses (GPS emissions) at an instant.

topics related with the development of these methodologies later have begun working at the Public Transport Authority using the tools and information developed. This has enriched and facilitated the collaboration among institutions.

3. Most important data processing methods

In this section, we recall the methods that have been developed to obtain valuable information from the data available, i.e., GPS, smart card and GIS information.

The first important processes required to perform more interesting analyses are to allocate bus GPS points to bus routes (Fig. 3) and to match the smart card transactions and positions databases, using the bus plates or station codes and times. Using this information, several other developments can be made. Complementing GPS data with the bip! database, the number of passengers boarding each bus route at every bus stop could be identified for different time intervals. Moreover, as only boarding information is stored in the databases, a procedure to estimate the alighting bus stop or Metro station for each trip-stage (Fig. 4) was developed (Munizaga & Palma, 2012), yielding trip stages matrices at bus stop level. Also, using the trip-stages linking procedure proposed by Devillaine et al. (2012), origin-destination trip matrices were generated.

The estimation of alighting estimation, trip linking and trip purpose have been validated by Munizaga et al. (2014), using exogenous data from measurements, surveys and personal interviews. The validation is very positive, showing correct estimation for over 80% of the cases.

On another stream, a time–space diagram for all buses operating in all routes is built from the GPS information (Fig. 5), and bus speed profiles for every bus route disaggregated by segments and time intervals are obtained using the methodology proposed by Cortés et al. (2011). Fig. 6 shows an example of a speed profile for a particular route. Time is in the X-axis, and the route segment is in the Y-axis. The speed is shown in a scale of colours ranging from red (very bad—below 15 km/h) to blue (excellent—over 30 km/h). This type of visualization allows detecting problems such as bus bunching (when the lines in Fig. 5 begin to bunch together) or possible infrastructure problems, when observed bus speeds are too low, or their variability is too high. DTPM regularly uses this data to identify locations where an intervention to the infrastructure or the operation of the system is required.

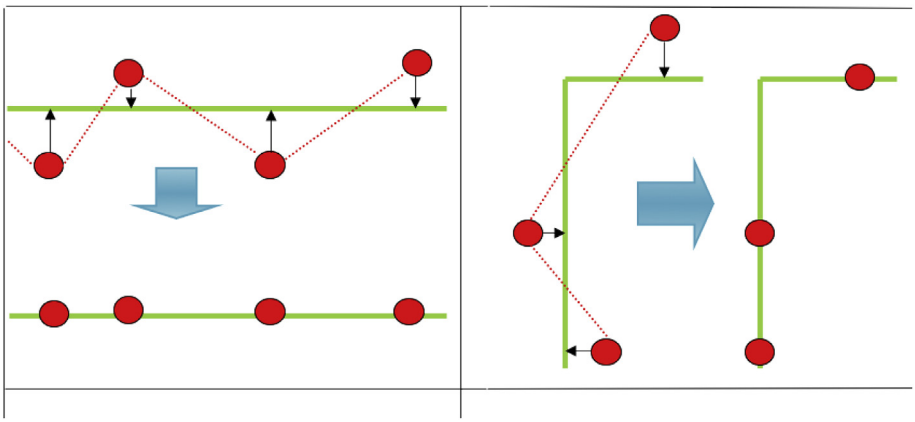

Fig. 3. Allocation of GPS pulses to bus routes. Source: Cortés et al. (2011).

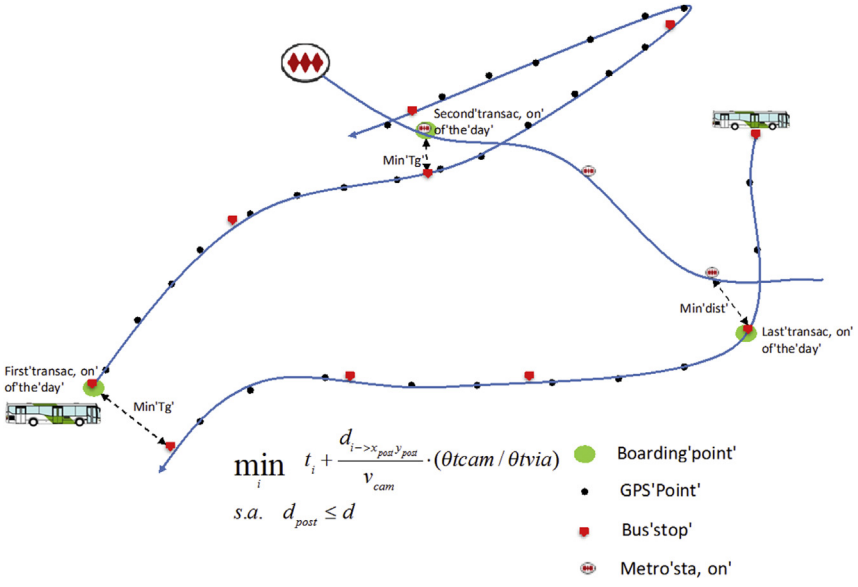

Fig. 4. Alighting point estimation. Source: Munizaga and Palma (2012).

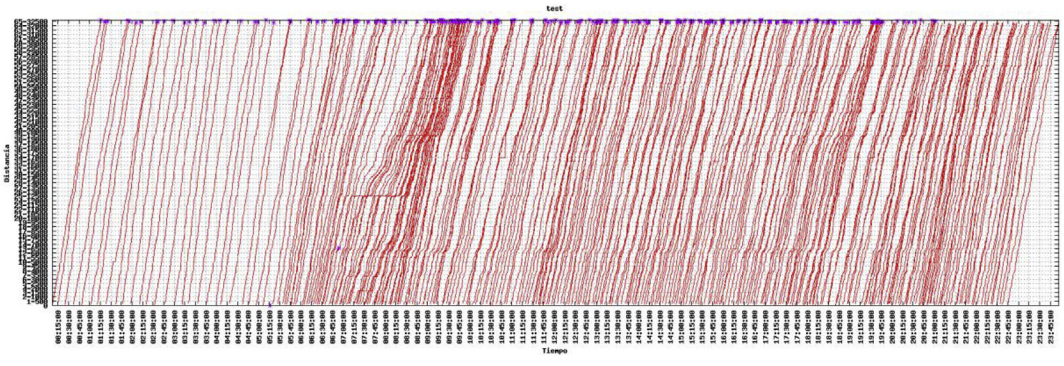

Fig. 5. Time-space diagram for one particular bus route.

After these processes, a very detailed database of trips is obtained, which contains boarding stop, alighting stop, sequence of routes taken, travel time, transfer time and waiting time at transfer. This information can be used to generate detailed level of service indicators (Munizaga, Núñez, & Gschwender, 2016), that allow monitoring of system performance. Fig. 7 shows an example of global trip indicators over time for travel time, number of stages per trip, relation between real distance (RD) and Euclidean distance (ED) of the trips, speed of the trip and its Euclidean distance. This information can also be disaggregated by zone of origin, zone of

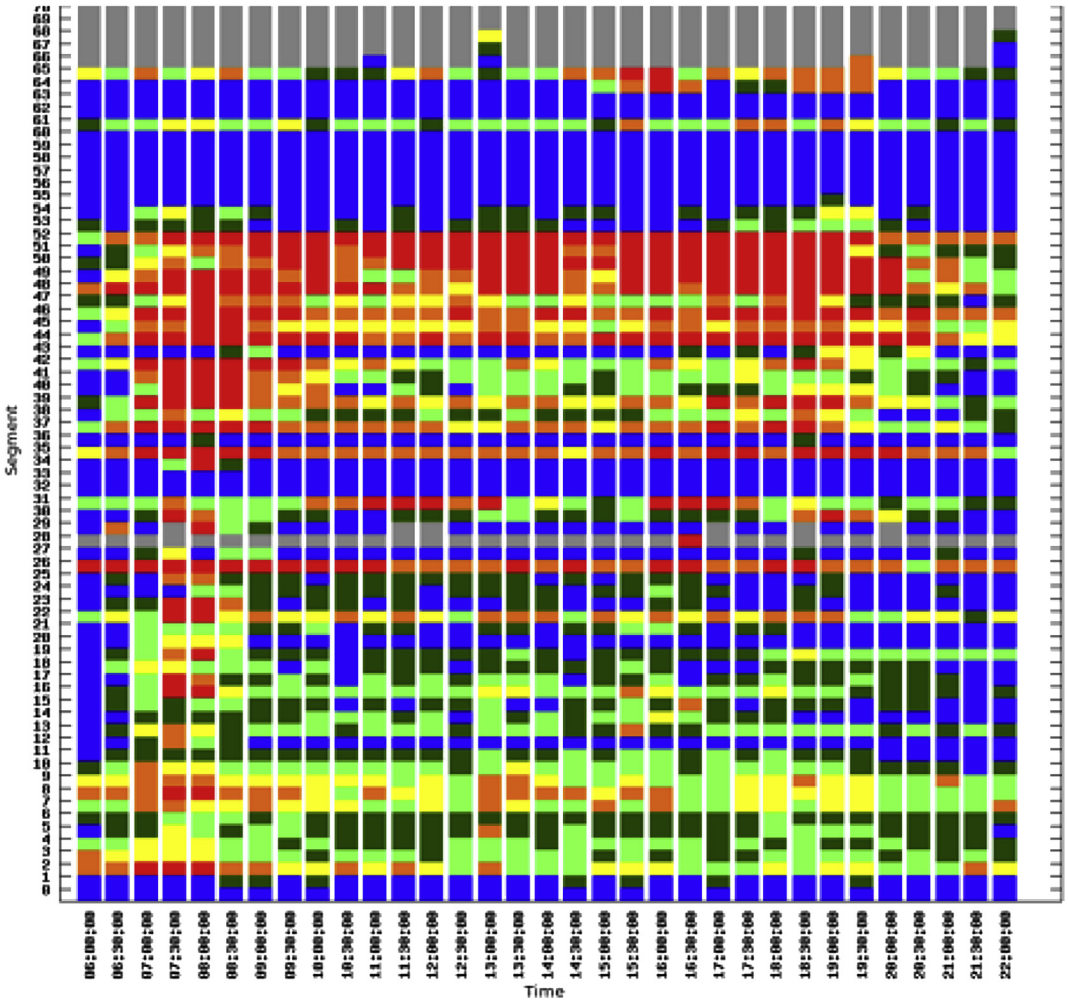

Fig. 6. Speed profile for one particular bus route.

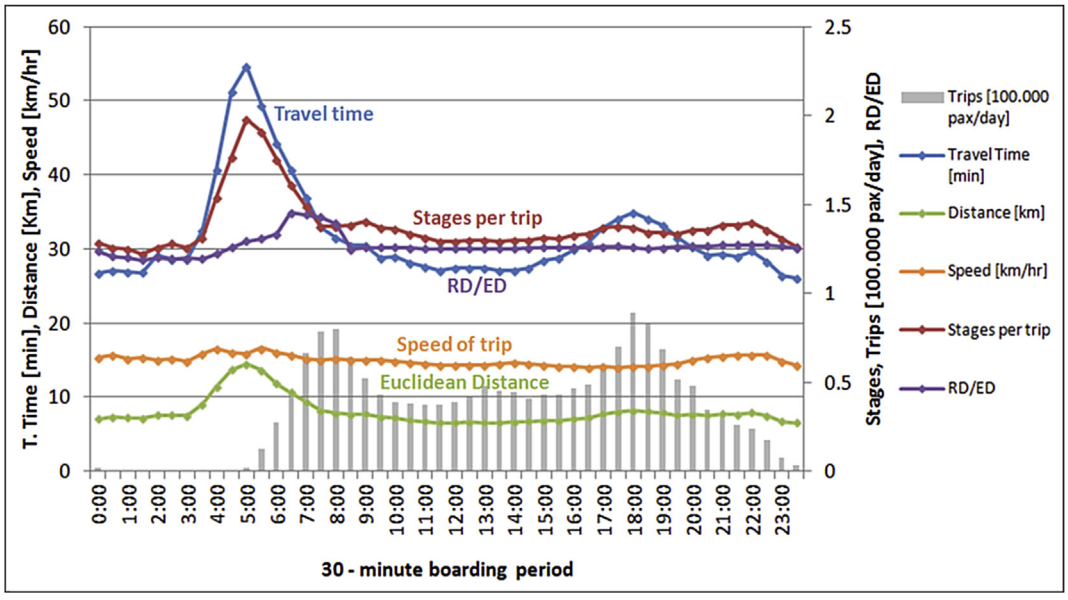

Fig. 7. Global level of service indicators. Source: Munizaga et al. (2016).

destination and operator. For example, in Fig. 8, the travel time in the morning peak from each zone of Santiago to the CBD is represented (red dots are Metro stations).

The software that produces these results has been protected, registered as intellectual property of the University, and licenced to DTPM. The information obtained from processing the data is regularly used by the transit authorities to make decisions about infrastructure and bus priority schemes; to define operational improvements to the system; to define modifications to the service network; to implement focused information campaigns; and to design new elements of the system such as bus stations, among other things. Also, outside DTPM, it is used by the transport planning agency SECTRA and by Metro and the bus operators.

3.1. Metro station

One of the limitations of smart card data in general, and of our data in particular, is the lack of socioeconomic information. In Santiago, most cards are not personalized, and in those cases where cards are personalized (e.g., student cards), that information is hidden for research analysis due to confidentiality considerations. This imposes a limitation on the types of analysis that can be performed. To overcome this difficulty, we have developed a method to estimate the zone of residence of frequent users, which allows imputation of socioeconomic variables (Amaya & Munizaga, 2015). The method is based on the observation of morning transactions of users observed at least four times in a week. It is applied to a sample of nearly 1 million frequent users (cards) with 60% of them also showing spatial regularity.

Also for frequent or regular users, we applied discrete choice theories to model customer loyalty. Using the modelling framework proposed by Bass, Donoso, and Munizaga (2011) and the estimated zone of residence as a proxy of income, we develop stability models for four income segments, using different levels of service indicators as explanatory variables.

4. Policy and planning applications

The information obtained from these methodologies and databases has recently seen an important increase in use. Starting in 2008, the database of one week per year has been processed. At the beginning, the data started being used by DTPM to analyse the speed of buses in specific streets where some intervention was already planned. This allowed confirming the speed problems in those cases. Later on, lists with the worse cases (bus line, 30min period and 500 m section) in terms of speed were constructed. Given that Santiago has some 3000 km of streets with buses, this automatic identification of the worse cases was very helpful. After a validation of each of the worse cases with field experts, they could be faced in order to find a solution to each problem. Recently, more elaborated indexes have been developed in DTPM to identify worse cases, using not only this disaggregated bus speed information, but also bus frequencies in links, to prioritize those cases where bad speeds are suffered by more buses and passengers. Moreover, cases where very low speeds sharply increase in the following link are being identified, as these represent bottlenecks with large potential of speed increase if they are solved.

On the demand side, after several exploratory experiences, in 2012, the detailed trip database started being widely utilized for practical uses at DTPM. The evaluation of changes in the routes layout benefitted from this detailed information, which allowed identification of the number of (positively or negatively) affected users in an easier way than the traditional and limited manual measurements. The same happened with the evaluation of the creation of new lines or the removal of some of them. In addition to the number of affected users, their alternatives can be analysed, given the origin-destination information available. The information of transfers between bus lines was the key input for a plan to reduce transfers by merging short lines into longer ones, implemented between 2012 and 2013.

The evaluation of new direct bus lines to provide alternatives to the most congested Metro links has also been improved by extracting those trips that use specific links of the network. The analysis of the origins and destinations of those trips permits the identification of potential layouts for these alternative bus lines. Moreover, the expected demand of future intermodal stations between Metro and buses has been estimated by means of this trip information. In several cases, load profiles have been constructed and used to assess the impact of the modification, creation or removal of bus lines.

Other useful information provided is the number of passengers boarding at each stop. This is used by DTPM to determine the infrastructure needed at every stop, in terms for example of the need of a shelter and its size. In addition, the implementation of solar energy illumination systems at stops was guided using this information, which indicates not only the aggregated number of

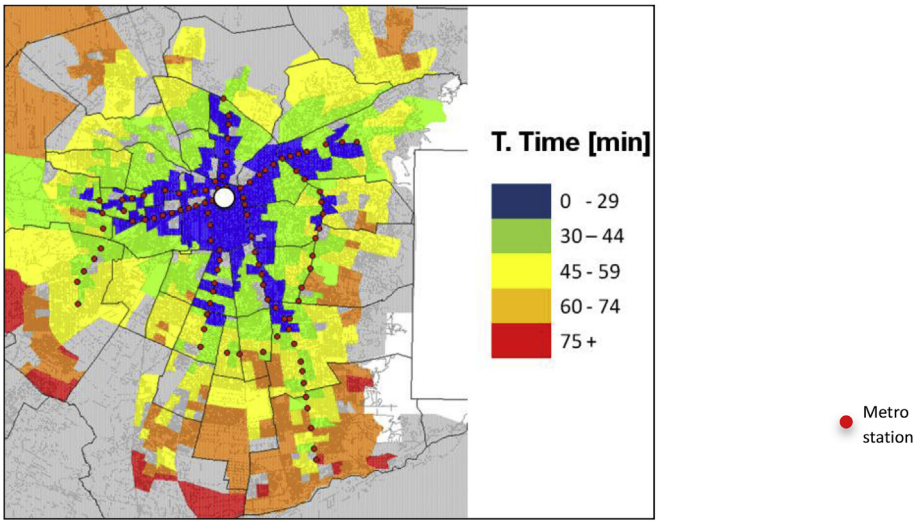

Fig. 8. Travel time to the CBD in the morning peak. Source: Munizaga et al. (2016).

passengers, but also their distribution throughout the day, thus allowing the quantification of the number of passengers boarding at night at every stop. The detailed design of express services, in terms of the specific stops that they skip, is facilitated with this boarding information. Changes of the system have to be shared with the users. To size up these information campaigns and determine where and how much printed information to distribute to users, DTPM uses the boarding per stop and period data.

The long-term monitoring of key quality variables of the public transport system (e.g., waiting and travel times) is radically facilitated and improved by this information. Formerly, since 2007, DTPM measured waiting and travel times for several periods in a small sample of bus stops and trips, at a quite large cost. Now (since 2012) these manual measures have been completely replaced by the information obtained from the trips database. This information covers all bus stops and a much larger sample of trips. Thus, not only representative averages are obtained, but also the variability of waiting and travel times can be followed. For instance, the percentage of trip stages with large waiting times can be obtained. Public transport users are very sensitive to large waiting times, even if they are not high on average. Analysing the distribution rather than the mean is crucial to understand quality of service, and this type of data makes it possible.

Not only DTPM, the public transport authority of Santiago, is currently using this data. The private bus operators also use it to analyse modifications to their services, which they can propose to the authority. The publicly owned Metro company has been comparing this origin-destination data with the results of their traditional yearly surveys (Pineda, Schwarz, & Godoy, 2015). Given the good performance of the matrices generated with the passive data, they are considering changing the focus of their surveys in order to complement the origin-destination information of the passive data matrices with that kind of information that they cannot provide (mainly socioeconomic characteristics of the travellers). Both for the authority and most operators, the trips databases have become the main demand input, which is complemented by specific manual measures (e.g., occupancy levels in maximum load points).

Beyond the public transport authority and operators, the information is also being used by other actors (universities, other public agencies, consultants), to complement other inputs and to avoid costly manual measures. Several public transport assignment models have been calibrated using this data—demand and bus speeds—for general or specific purposes (e.g., evaluation of suburban train or tram lines).

In the future, the traditional origin-destination surveys will probably be adapted to include this source of information in order to reduce costs and improve the quality of the data obtained. Although in Santiago only bus and Metro data is regularly being processed at the moment, additional sources of passive data exist and could be incorporated; data from the mobile phones system and the electronic toll collection system in urban highways can potentially widen the spectrum of this information, including other modes like pedestrians, bikes and cars. In this direction, the traditional transportation equilibrium models will probably need to be adapted as well, to take advantage of the new data sources.

5. Conclusions

In this paper, we describe a successful collaboration experience between academia and the public transport authority of Santiago to develop tools that are currently providing useful information from automatically collected GPS and smart card data. The success of this collaboration can be recognised by the academic achievements in research, teaching and extension, and also by the wide use that this information has by different actors, the considerable savings in manual measures that this has implied and the better quality of information obtained, improving public policies related to urban mobility.

The information obtained from the databases and methodologies described in this paper have proven to be highly valuable and useful. For example, the information is being used—both by the operators and the public transport authority—as the main demand input to assess the impact of the creation, removal or modification of specific bus routes, in terms of the positively or negatively affected users. The disaggregated demand information can be used to evaluate and prioritize new specialized infrastructure (e.g., transfer stations, bus stops or bus corridors), and to define their size and characteristics. It is even being used to dimension—both in terms of size and localization—users' information campaigns. Moreover, with this extensive source of data, it is possible to build and periodically bring up to date quality of service indicators both at the aggregate and the disaggregated level, including variability of the indexes (travel time, waiting time, number of transfers, etc.). Thus, the level of service can be monitored over time and space, identifying routes, zones and hours of the day that operate under specified quality standards, allowing focusing the improvement efforts in those cases.

There are several directions in which the methodologies can be enriched to improve the information obtained. Some of these directions will need additional information that will have to be collected manually (surveys, counts) to complement the passive data. Fare evasion, which is a relevant issue in Santiago (especially in case of the buses), is a good example of this. Some explorations have been made to correct the estimated trip information, based on the assumption that there are two types of fare evasion. First, there are complete bus trips that are not validated in any of their stages and therefore cannot be found in the databases, and second, there are combined bus-metro trips in which only the bus stage is not validated, implying that the estimated trip is distorted in its origin or destination. Another dimension of further work is the development of visualization tools. As the information generated allows disaggregated analyses of large data, ad-hoc tools to navigate in that information can be very useful. In their absence, all analyses have to begin with consulting the information databases, something that is time consuming and requires specific knowledge.

This alliance between academia and the public transport authority is strongly based on mutual confidence, which has been reinforced as both parties benefit from the joint work. The management of the formal collaboration agreements is time consuming, and the constant interaction with administrative procedures is sometimes frustrating. In spite of this, we—researchers at the University and professionals at the public transport authority—believe that the methodological challenges and the utility of the results are worth the effort. It has also been important for the success of this joint work that the participation of the public transport authority has not been limited to sharing data and declaring wishes. During all these years, its representatives have been actively participating in weekly work meetings at the University, looking in detail at the results to give feedback to the methodologies and also participating in the strategic decision-making meetings. At the same time, the commitment of the researchers at the University has been crucial. After all, the vision of the possibilities of this collaboration and the main idea of the project came from them in the first place.

It is worth noting that, thus far, all of this information is obtained from passive data (GPS, smart card, GIS information) only. This means that, after the development of the methodologies and tools, the cost to obtain all of this valuable information is practically negligible, in comparison to traditional field measurements. This

has allowed the public transport authority to avoid expensive field measurements, for instance OD surveys and level of service measurements. The latter were formerly made on a regular basis until 2012, when they were replaced by the indicators obtained from the passive data methodologies. The accumulated cost of those measurements between 2007 and 2012 exceeds by far the development cost of the passive data methodologies.

The level of detail of the new information allows examination at disaggregated levels of both time and space, and the amount of data (automatically collected) permits the analysis not only of average values, but also of their variability, which is a key aspect in public transport quality of service. This is why we believe that this sort of data, tools and information are changing the way transport planning is being done.

Acknowledgements

Funding: Fondef D10E-1002, ISCI (ICM-FIC: P-05-004-F, CONICYT FBO816), Fondecyt 1161589.

References

Amaya, M., & Munizaga, M. A. (2015). *Estimating the residence zone of frequent public transport users to make travel pattern and time use analysis*. Working paper. Universidad de Chile.

Bass, P., Donoso, P., & Munizaga, M. A. (2011). A model to assess public transport demand stability. *Transportation Research Part A, 45*, 755–764.

Cortés, C., Gibson, J., Gschwender, A., Munizaga, M., & Zúñiga, M. (2011). Commercial bus speed diagnosis based on GPS-monitored data. *Transportation Research Part C, 19*, 695–707.

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

DTPM. (2015). *Informes de Gestión del Directorio de Transporte Público Metropolitano*. retrieved from http://www.dtpm.gob.cl/index.php/2013-04-29-20-33-57/informes-de-gestion accessed 24.07.15.

FONDEF. (2015). *FONDEF program and its mission*. retrieved from http://www.conicyt.cl/fondef/fondef-program/ accessed 23.07.15.

Gibson, J., Munizaga, M. A., Tirachini, A., & Schneider, C. (2016). Estimating the bus user time benefits of implementing a median busway: Methodology and case study. *Transportation Research Part A, 84*, 72–83.

ITF. (2015). *Big data and transport: Understanding and assessing options. International transport forum* (Corporate Partnership Board Report).

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

Munizaga, M. A., Núñez, C., & Gschwender, A. (2016). Smart card data for wider transport system evaluation. In J. D. Schmöcker, & F. Karachi (Eds.), *Transport planning with smart card data*. CRC Press.

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 Part C, 24*, 9–18.

Pelletier, M.-P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C, 19*, 557–568.

Pineda, C., Schwarz, D., & Godoy, E. (2015). Comparación y validación de matrices origen-destino de viajes en Metro de Santiago obtenidas a partir de encuestas y de transacciones de pago. *Ingeniería de Transporte, 19*(1), 55–72.