Using smartcard data for agent-based transport simulation

Using smartcard data for agent-based transport simulation

P.J. Fourie, Alex Erath, S.A. Ordonez, A. Chakirov and K.W. Axhausen, Future Cities Laboratory, Singapore ETH Centre, #06-01 1 CREATE Way, Singapore, 138602

Abstract

The disaggregate nature of transit smartcard data is congruent with the travel demand specification as used by agent-based approaches to transport modelling. Using a full day of public transport smart card transactions recorded in Singapore, we developed an approach to transform the smartcard data into both transport supply and demand, while simultaneously eliminating the need to simulate the interaction between cars and buses. In order to produce realistic travel times for buses, we estimated a regression model of bus speed between stops that is dependent both on the level of demand and network topology. We implemented a model of bus dwell time at stops that is dependent on the ridership of the bus and its configuration. As the need for simulating the dynamics of the bus between stops is eliminated by the speed model, it allows us to simplify the supply network dramatically with only one link between bus stop combinations, and another link at the stop for buses to queue in order to perform dwell operations. These modifications, along with a simplified mobility simulation dramatically improves simulation times ensuring useable results in under an hour. In addition, our modelling framework is highly adaptable and requires only limited efforts to be applied to other public transport systems in cities where similar data streams are available.

1. Introduction

It is widely agreed that the provision of attractive public transport services is of central importance for the sustainable development of cities as it outperforms individual motorized transport in terms of cost, environmental impact and social equity. To plan and design efficient urban public transport service provision, municipal planning organizations (MPOs) and service operators usually develop transport demand models. The models currently used in practice operate on the principle of modelling trip flows between geographical zones and hence are subject to aggregation over the horizons of time and space. However, current urban transportation problems such as congestion and service reliability are of an inherently dynamic nature. This is particularly the case for public transport as overcrowding, schedule reliability and bus bunching are inherently dynamic phenomena observed in many cities all over the world.

To address the shortcomings of aggregate methods, large-scale, agent-based transport demand simulation (LSABTS) models have been developed that preserve full temporal dynamics as well as disaggregate information on individuals through the entire modelling and simulation process. Software packages such as MATSim (Balmer et al., 2009)or TRANSIMS (Smith et al., 1995) are designed to dynamically simulate transport demand and supply for millions of agents over an entire day at the temporal resolution of a second. These models take an activity-based approach, acknowledging that travel demand is the result of the need to perform activities in different points in space and time. Entities in the simulation have a one-to-one correspondence with their real-world equivalents, therefore an agent in the simulation represents a single commuter in the physical system, and private and public transportation vehicles have equivalent entities in the simulation. Dynamic phenomena such as congestion and bus bunching emerge from the interaction of all participating agents in the simulation.

While many cities recognize the potential of agent-based and activity-based approaches, these methods have come under considerable criticism for being exceptionally data hungry, with finely-grained information needed at all stages of the travel demand modelling process, i.e. a detailed synthetic population describing travel demand as a function of various household and personal demographic attributes; as well as the modelling of the transport supply system, i.e. the transportation network, vehicle fleet, public transport schedule and activity facility capacities serving the demand. Dynamic assignment models are also notoriously difficult to calibrate, as observed traffic volumes and travel times are emergent phenomena resulting from the dynamics in the system. Because these systems work from the bottom up, they require that the range of individual behaviors have to be adequately represented in the travel demand description, and the interactions between entities have to be adequately described in order for the full range of dynamic phenomena to emerge as observed in reality. And because the full system of all participating transportation modes subject to the full range of commuter choice dimensions has to be simulated repeatedly in order for the system to reach a steady state, simulation times are notoriously long.

In the light of these difficulties, some authors have argued for the use of so-called direct demand models that do not attempt to capture the full gamut of cause and effect as is being attempted in the activity-based methods, but instead rely on inferring the reaction of the transportation system to dynamically changing demand based on observation.

The data produced by automatic fare collection systems (AFCS) represents a uniquely appropriate input to a direct demand model. The data records produced by such systems document public transport ridership patterns in great detail, at the level of individuals with precise spatio-temporal information. Since 2005 authors such as (Bagchi and White, 2004, 2005) have been studying the potential of using this type of data. (Pelletier et al., 2011) present a summary on how AFCS data has been used to analyze public transport systems worldwide. Different measures of quality indices have been proposed, and methods to find behavioral patterns. Such data can be directly translated into an agent-based travel demand description, because travel demand is recorded at the level of individual commuters that translate into individual agents in the simulation.

The aim of this paper is to assess the potential of using AFCS data for agent-based simulation for the case of Singapore. To this end, we set up a MATSim scenario using smart card transaction data as travel demand input and detailed public transport schedule information and a global positioning system (GPS) navigation network to describe supply. In order to eliminate the need for simulating the full transport system of public and private vehicles for realistic travel times to emerge from repeated network loading, we derive a model of the speed of buses between public transport stops as a function of public transport demand from the smartcard data and topographical information contained in the network description. We extend the existing MATSim framework by introducing stochastic terms to describe bus dwell time behavior at stops and travel time between stops, which are the main determinants of service reliability in public transport operations. We validate the resulting model against information contained and derived from the smartcard data.

The applicability of the approaches is presented using the splitting of a very long existing bus line as a study case. We conclude by evaluating the approach's applicability for practice and identifying future research directions.

2. User equilibrium and public transport in MATSim

MATSim is a platform to simulate transport demand and supply interactions allowing for large-scale scenarios where millions of agents represent people interacting. For each agent, a daily activity plan is assigned representing the sequence of activities it has to perform at different times and at different locations within a specific period of time (in general one day). MATSim utilizes an evolutionary algorithm to reach a steady state. That is, the same day is simulated many times, where a fraction of the agents modify their plans after each iteration. There are many ways to modify their plans; they can change the departure time, the travel mode of a sub-tour, or the location of a given type of activity, among others. This work is focused on the modification of the route, more specifically for public transport users. The utility of the day is measured for each agent in each iteration using a scoring function that rewards agents for performing activities, while penalizing them for travelling, transferring between transport modes, waiting at transit stops and arriving late for activities, etc. (Charypar and Nagel, 2005). Agents save a small number of plans, remembering those that scored well and forgetting the others. Thus, the general score of the population tends to grow until, after hundreds of iterations, the system reaches user equilibrium and the generalized utility can not be improved any more (Balmer et al., 2009).

MATSim includes a full implementation of public transport (Rieser, 2010). On the transportation supply side, the system is represented by stop facilities and transit lines. Several routes belong to each line. Each of these transit routes holds the information of the sequence of stop facilities with the expected arrival and departure offsets, the sequence of links in the road network a vehicle of this route has to follow, and the departure times of all the services of the route. As the links that public transport vehicles have to follow belong to the road network that private vehicles use in the simulation, public transport vehicle travel times are affected by congestion or modes that share the network with private transport, while modes with exclusive networks and precise signaling and control (i.e. rail, subway, monorail) tend to operate close to scheduled times.

Another source of deviation from schedule, especially for bus, is the time spent to take on boarding passengers and allowing passengers to alight. This dwelling process can be modelled in two ways in MATSim: the simple approach just calculates the time a vehicle has to be stopped according to the number of passengers, type of vehicles, number and configuration of vehicle entrances and exits, and vehicle occupancy, while a more fine-grained approach would simulate a queue agents use to enter the vehicle.

For bus stop facilities, the availability of a bus bay can be specified to account whether a bus is obstructing a link for cars during the dwelling process. MATSim also allows the same vehicle to be scheduled to perform several services; if it is late, the next service won't be able to start. Thus, the level of detail of the public transport module can simulate phenomena such as early or late services, crowded vehicles, bus or train bunching and long waiting times resulting from service denial to fully loaded vehicles.

3. CEPAS

1. Suitability of using CEPAS data to describe public transport demand

Contactless, stored value smart cards for fare collection have been introduced in Singapore in April 2002 under the name EZ-Link. In 2009, EZ-Link was superseded by a new standard for electronic payment smartcards called Contactless e-Purse Application, or CEPAS. CEPAS-compliant smart cards can be used island-wide for payment of all modes of public transport, regardless of operator, as well as for minor retail transactions, parking and road toll payment. Though cash payment of single fares at higher rates is still possible, e-payments with CEPAS cards account for 96% of all trips, which makes the data records from CEPAS highly comprehensive and the missing cash paying travelers negligible (Prakasam, 2008).

In Singapore, the fare system is distance-based and customers have to tap their CEPAS card on the reading device every time they enter and leave a train station or a bus, or they will be charged the maximum amount for that particular service. GPS devices on buses ensure that each transaction is associated with a unique transit stop identifier, as well as the vehicle identification number. Each transaction therefore contains information on timing and location, and generally most trips contain information on both boarding and alighting transactions; a notable exception is the case of concession cards for schoolchildren, students and senior citizens where the maximum charge is capped at 7.2 km. These users therefore sometimes do not tap out, especially when the bus is very full and users want to alight faster.

The completeness of the Singaporean smart card data, both in terms of market penetration and recording of both boarding and alighting locations, distinguishes it from those collected by the majority of other automatic fare collection systems and allow more detailed assessment of travel behavior and mobility patterns. In many other countries users do not have to tap out of the bus or tram and the alighting location is therefore not recorded, although researchers recently proposed techniques to impute its value in the absence of such information (Munizaga and Palma, 2012). Furthermore, as the CEPAS cards are durable and easily rechargeable, people tend to continuously use one single CEPAS card with a unique card ID for all their public transport journeys for substantial periods of time. As the technical setup of the system doesn't allow more than one person to travel on a single CEPAS card, it can be assumed that each unique card ID represents a single person. This enables highly disaggregated analysis of individual itineraries and opens new ways for understanding people's travel behavior over the short as well as longer term scales.

Given the temporal and spatial resolution of the CEPAS data, it is perfectly suited to represent travel demand in a simulation of Singapore's public transport system using MATSim. By combining it with information on supply derived from published schedule information we can generate a simplified MATSim scenario. We then can use this scenario as a predictive system to evaluate changes in public transport service provision such as the type of buses being used, service frequency and service network.

2. Combining agent-based transport simulation and CEPAS data

CEPAS data only describes demand for public transport services. Therefore, we restrict the scope of the MATSim model presented in this study to public transport only and simulate its operation in isolation of other transport modes.

The scope of analysis is restricted to route choice effects as the simplified scenario covers the public transport system in isolation and no information is available concerning the trip purposes or socio-demographic background of travelers. Furthermore, the system cannot account for mode choice effects, i.e. passengers switching away from or switching to public transportation due to changes in system performance; neither can it account for so-called induced demand, where changes in the level of performance of the public transportation system result in people performing more or fewer activities because of more or less time opening up in their travel time budgets.

Information on individual travelers is restricted to a unique card identifier and fare type category, namely child/student, adult and senior. Furthermore, we do not directly know about the trip purpose and actual trip origin and destination at the level of buildings but only the public transport stop where the transaction took place. As we neither infer actual trip origin and destination nor trip purpose, the scope of the analysis cannot include destination choice effects.

To restrict the scope of the simulation to public transport we need to account for interaction effects with cars resulting in increased travel times. To this end, the observed travel times from the smart card data are used to develop a regression model of bus movement between transit stops. We use the error term from this regression model in order to arrive at a stochastic model of travel times between transit stops that eliminates the need to model a fully detailed network during the simulation, and allows us to predict the distribution of travel times during the course of the day for network links that are not currently in existence, making system-wide network re-design evaluation possible.

4. Methodology

In MATSim, public transport vehicles share roads with other vehicles and dwell operations are modelled in detail. As the objective of this work is to set-up a simulation only using AFCS data we simplified the MATSim mobility simulation and the transport network, restricting it to public transport vehicles only. Figure 1 shows the processes we designed and implemented and how these affect a standard MATSim simulation. In the next section we describe the reconstruction of bus trajectories, the generation of a public transit schedule, then the generation of public transport trips as MATSim plans (the demand), followed by the simplification of the road network, and finally the new mobility simulation model and its constituent sub-models in MATSim.

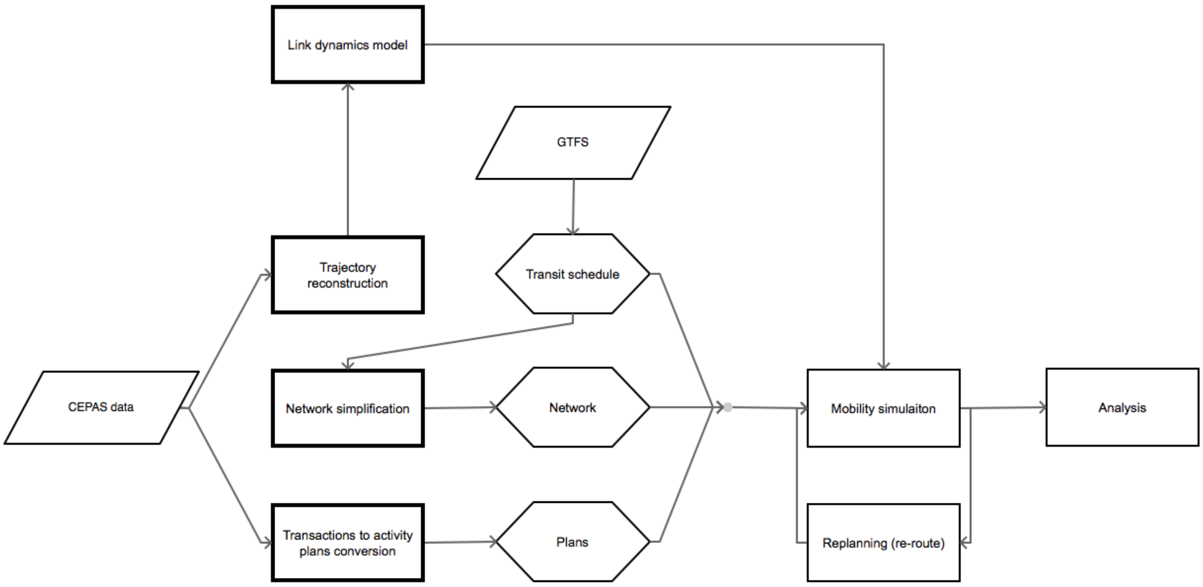


Figure 1. **Simplified public transport simulation overview**

1. Reconstruction of bus trajectories

Given boarding and alighting transactions of bus users it is possible to estimate the position in space and time of the corresponding buses (their trajectories). For each vehicle ID in the system, by grouping its transactions at each stop into sets that represent bus dwell operations, we can impute the time that it takes for the bus to travel between bus stop locations. From our electronic transit route profile, we know the exact route between bus stop locations, and therefore we can reconstruct the vehicle’s trajectory once all dwell operations have been identified.

There are a number of challenges in the trajectory reconstruction process:

Bus stops without transactions: As boarding and alighting actions might not occur at every stop, the bus can remain "invisible" to the system. We used interpolation techniques to estimate the time when the bus reached these stops. For stops that precede or follow the first and last “visible” stops we didn't apply extrapolation to find the bus arrival times at those stops.

Early tap-outs and late tap-ins: As the bus approaches the public transport stop, the GPS system automatically activates the reading device, making it possible to tap out before the bus doors have opened. Furthermore, sometimes passengers have entered the bus but are still fumbling to get their cards out for the reader, and the tap-in registers late. As these transactions don't happen when the bus is already at the bus stop, they have to be recognized and filtered to obtain a better estimation of the arrival and departure times of the public transport vehicles.

GPS errors: The way the system recognizes the stop where the transactions are occurring is to read the position of the buses from their GPS devices. If GPS readings are incorrect, especially when stops are very close to each other, during inclement weather or in high-rise urban environments, the stop identifier can be recorded incorrectly.

For a complete description of the trajectory reconstruction process, the reader is referred to (Fourie, 2014). We coded these estimations of when dwell operations occur and the trajectories between stops into MATSim 'events'; time-stamped, atomic units of information normally generated by the agent-based simulation that give a complete description of all vehicle and commuter agent actions during the course of the simulated day. The resulting XML file can therefore be visualized and analyzed using MATSim-compliant software, and direct comparisons against MATSim simulations are therefore greatly simplified.

2. Generation of a public transit schedule

We used the reconstructed bus trajectories to determine the number of services and the time when the services start for every bus line in Singapore. As the vehicle identifier of each bus is known in the CEPAS data, we assigned the corresponding type of bus in the simulation, accounting for carrying capacity, doors operation, single or double decker configuration (which affects bus dwell time). Figure 2 summarizes this process. We compared these results with the commonly used Google Transit Feed Specification (GTFS) of the public transport system in Singapore. We recognized a significantly smaller number of services in CEPAS data: 4 bus lines were not found, 33 bus routes (different sequences of stops within a bus line) were not found, and from the 91115 services specified in GTFS only 78515 services were recognized (86%). It is possible that a whole service is not visible due to lack of transactions or GPS errors as mentioned before. The difference in this comparison is still considerable, so the GTFS numbers can be overestimated. As the reconstruction of train trajectories presents even greater challenges than those for bus, because transactions are recorded at the station entrance and not when passengers enter and exit the vehicles, we have not implemented it at this point, but we are working on implementing the method developed by (Sun et al., 2012) in a future iteration. Consequently, the number of train services and their start times are directly obtained from the GTFS.

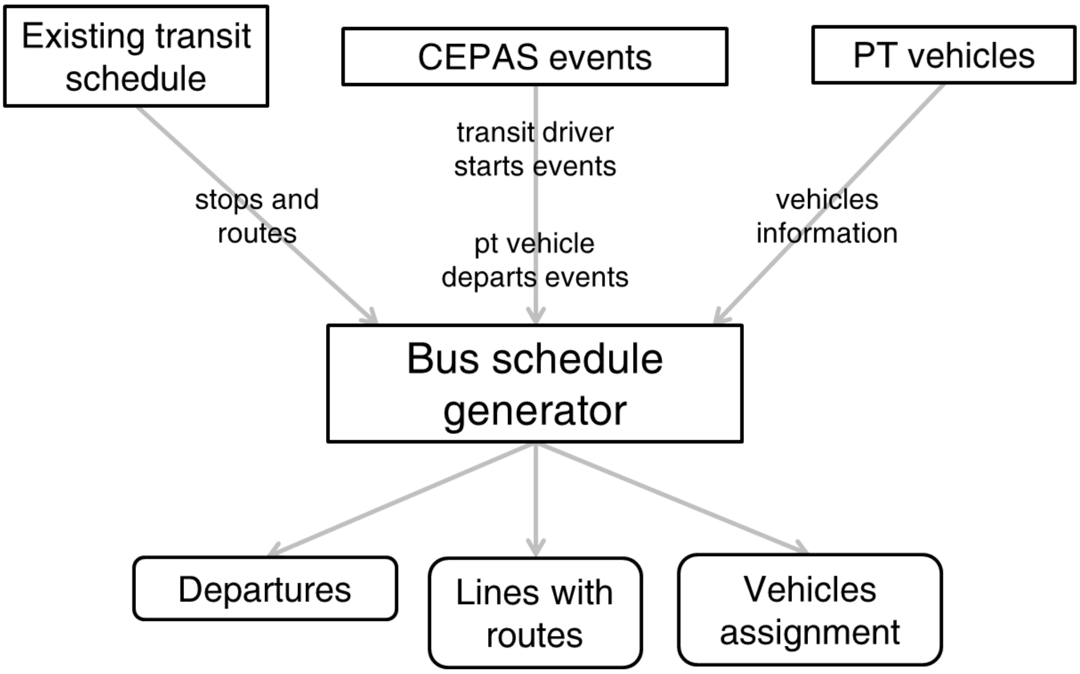


Figure 2. **Transit schedule generation using bus trajectory information from reconstruction process**

3. Generation of public transport trips

MATSim is an activity-based simulation framework, and its demand description is a timed sequence of activity locations and connecting trips for each agent in the study area. Generating an agent-based demand description from the smart card data is at first glance a straightforward task; each boarding and alighting location can be used as an activity location in an agent's activity schedule. However, this would mean that we identify each transfer in a public transport trip as a significant activity, and we would then also over-specify the demand description by determining transfer location. It is important that realistic transfer locations, and their associated walking and waiting times, rather emerge from the simulation than be specified in the demand description. To this end we want to identify the initial boarding and final alighting location of each multistage trip in the smart card data, and use these transactions as approximate activity locations and activity start/end times of the agents. A number of challenges have been encountered in this process.

Access waiting forms an important component in an individual's transit experience, however, in the case of buses, recorded times don't correspond to user arrivals and departures to the public transport system. As transactions correspond to boarding and alighting only, the time when users arrive at the bus stop are unknown (except in transfers). More realistic bus stop arriving times for passengers are important for waiting time calculations. On the other hand, bus-bus, bus-train and train-bus transfer times are known, and even exact bus lines can be assigned. That means, bus routes are fully reproducible from reality.

To assign bus users trip start times and identify individual multistage bus trips, we developed a two-step procedure. First, when a user alights from a bus and enters another vehicle, we established a threshold of 25 minutes to categorize those transactions as transfers or not transfers. If the time between alighting and boarding is more than 25 minutes, we assume that the user has left the system, therefore they accumulate newly recorded access waiting time upon re-entry. The second step assigns bus users start times. For this we used the reconstructed bus trajectories to extract headway times between consecutive services of the specified line. We had to assume (i) users wait exclusively for services of the line that they boarded in the transaction, ignoring other lines that serve the same stops (ii) they don't have external information on bus arrivals. This is not always true as users can be waiting for more than one bus number. They also can have more information about reliable bus arrivals from experience, or digital apps which estimate bus arrivals. Given these assumptions we assigned a uniformly distributed user arrival time to the bus stop within the corresponding headway.

Thus, we generated a MATSim activity plan for each CEPAS user, assigning dummy activities between given or estimated arrival and departure times to the public transport system.

4. Simplification of the network and mobility simulation

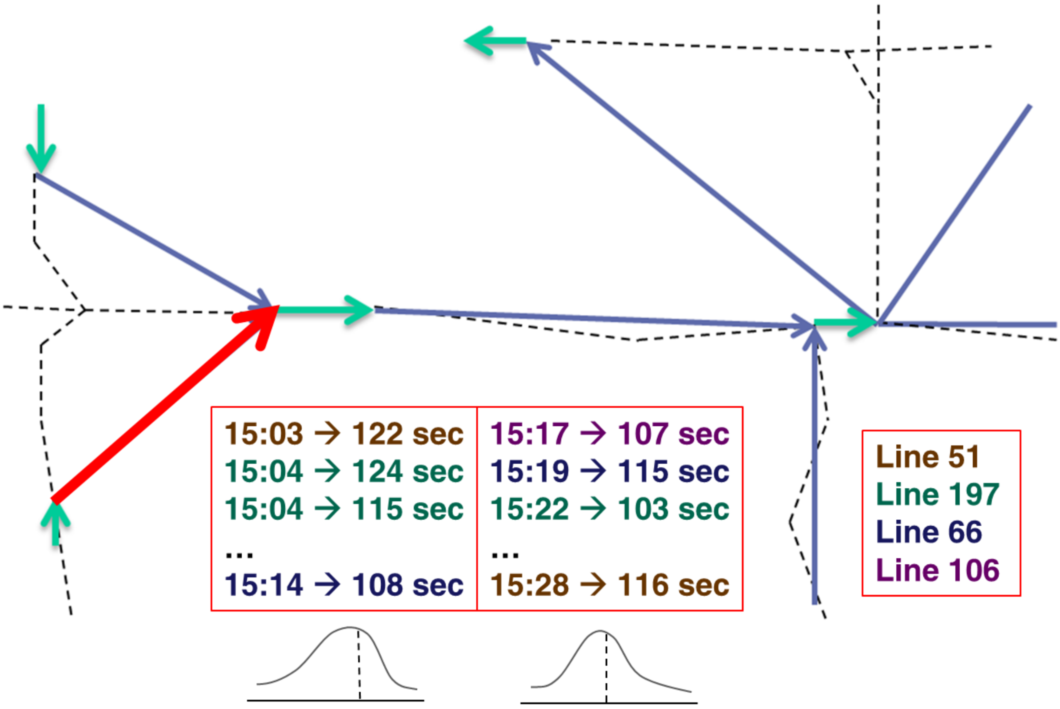


Figure 3. **Simplification of the MATSim network topology, showing stop to stop links and dwelling links before the stop.**

As only public transport vehicles are simulated, a detailed topology of the road network is not necessary. A reduction in the number of links and nodes of the road network represents a direct reduction in the MATSim mobility simulation computation time as its complexity is proportional to the network size and the number of agents. Thus, as Figure 3 shows, a single link was created for each public transport stop (dwelling link) and a single link connects each pair of consecutive stops. If two stops are consecutive in at least one line a link is created between them.

As mentioned before, the original MATSim mobility simulation is based on queues of vehicles at every link of the road network, depending on its corresponding capacity. That's how it accounts for the effect of car congestion on buses or vice versa. Without information about cars, but many observations of buses travelling, we dropped the queue model and introduced a stochastic travel time model, where the speed of buses on each link are drawn from a normal distribution; the parameters of which vary by time of day and are the result of a multinomial regression model that was estimated from the speeds observed between stops from the trajectory reconstruction step (Fourie, 2014). The parameters and results of the regression estimation are discussed in the following section.

With the modelled dynamic distributions, we modified the standard MATSim link dynamics (the queue model). Now, when a vehicle enters a link after a dwelling operation, a speed value of the link’s distribution for the corresponding time of day is sampled. During that time value the vehicle "goes to sleep", and after that, it appears at the next stop ready to start a new dwelling process. As we have not reconstructed train trajectories yet, the standard queue model for the rail mode is maintained (as trains in subway systems have less interaction with other vehicles this approach is not far from the real behavior).

5. Speed regression model

As shown by (Sarlas and Axhausen, 2015), the speed of vehicles in a network link are related not only to the level of demand on the link, but also to the network topology, the presence of signaling systems and surrounding urban density and activity level. While their study calculated average travel times for the entire Swiss road network of private and public transport, we focus our investigation on determining observed speeds at any given time of day as a function of network topology and indicators of the level of activity and demand that we can derive from the smart card data. The results of the estimation are shown in Table 1.

Table 1. **Coefficient estimates of a multinomial regression model predicting the natural logarithm of bus speeds between stops (m/s).**

	Estimate	**t-value**	**Relative importance**
Intercept	3.07E-01	6.48	
Intersections per km	5.07E-03	7.99	6.87%
Fraction of path with bus lane	3.54E-02	9.17	0.49%
Number of passengers tapped in	**-1.55E-06**	**-20.13**	**3.38%**
Avg. number of intersections per roving sq. km	-7.32E-04	-24.98	13.60%
Avg. degree of intersection nodes along path	-5.07E-02	-14.77	4.67%
Right turns made at intersections	-7.46E-02	-12.06	2.47%
Left turns made at intersections	-1.29E-02	-2.06	0.28%
Right turns passed at intersections	-3.47E-02	-14.47	2.67%
Left turns passed at intersections	-2.45E-02	-10.56	1.59%
Number of nodes within traffic control buffer	-3.12E-02	-30.25	8.99%
Path length (log)	4.81E-01	64.87	21.81%
Number of arrivals at destination stop per day (log)	-3.72E-02	-14.90	2.61%
Number of nodes in path (log)	-6.42E-02	-11.71	2.97%
Path length over Euclidean distance (log)	-3.83E-01	-24.19	4.33%
RMS radians turned (log)	-9.66E-03	-4.78	2.19%
Activities in progress per roving sq. km (log)	**-2.28E-02**	**-11.26**	**6.19%**
Smart card transaction rate per roving sq. km (log)	**-7.28E-02**	**-29.90**	**14.90%**

Multiple R-squared: 0.2054, Adjusted R-squared: 0.2053

The model predicts the **natural logarithm** of speed (m/s) as a function of the 15 variables listed in the table. Variable names in bold denote dynamic quantities that change on a per second basis. All the other variables are derived from the network topology. The table shows the estimated value of the parameter, followed by the t-value. The last column shows the relative importance of the variable in terms of its contribution to the multiple R squared value listed at the bottom, using the method of (Lindeman et al., 1980), implemented in the R statistical analysis platform (R Core Team, 2014) by (Grömping, 2006).

A Java class was created to calculate the topological variables, as well as to associate smart card transactions with network locations and use them to derive indicators of the level of activity and traffic that a bus might encounter between two stops at any given time of day. The variable names as listed are relatively self-explanatory; in many cases the natural logarithm of variables were used instead of their original values, in order for these variables to appear more normally distributed. The values of the less self-explanatory variables were calculated as follows:

Intersections per km: the number of nodes along the path of the bus between two stops that have more than one ingoing and one outgoing link or two pairs of parallel ingoing/outgoing links (nodes denoting changes in direction for one-way or bi-directional roads, respectively, therefore not intersections), divided by the length of the path in kilometers.

Fraction of path with bus lane: a number of road segments in Singapore have bus-only lanes in the leftmost lane, that are either exclusively for bus during the entire day or during peak hours. This variable denotes the fraction of the path that has such a bus lane. We do not take account of exclusivity by time of day, so this is a static variable.

Total number of passengers tapped in system wide: this variable denotes the general accumulation of passengers in the entire public transport system, and is therefore an indication of overall system load by time of day.

Avg. number of intersections per roving sq. km: for each node in the path between two stops we draw a circle with a 1 km² area and count the number of intersections within that area according to the definition stated previously, then divide the sum by the number of nodes in the path.

Avg. degree of intersection nodes along path: for each intersection along the path we take number of links that meet in the node as an indication of its relative complexity, as the more links that meet at an intersection affects the signaling times.

Right turns made at intersections, etc.: when a bus has to make a right turn at an intersection it generally takes longer than making left turns, as the estimates of these variables clearly reflect. In fact, making a left turn does not seem to have an appreciable effect on the model as is reflected by its low t-value, despite the large sample size of more than a hundred thousand stop-stop combinations used in the estimation.

Number of nodes within traffic control buffer: the locations of traffic control signals were supplied to us as a geographically encoded shape file. This variable records the number of nodes in the path of the bus that fall within a buffer of 30 meters from a traffic control signal.

Number of bus arrivals at destination stop per day (log): this variable accounts for the traffic at the destination stop, with the expectation that the more services that are offered at the stop, the longer a bus is likely to wait in a queue before it can perform dwell operations.

Number of nodes in path, path length over Euclidean distance, RMS radians turned: these variables attempt to capture the degree of 'friction' between two consecutive stops, that prevent the bus from reaching top speed.

Activities in progress per roving sq. km (log): for each node in the path between two stops, we draw a circle with a 1 km² area around the node and retrieve all smart card transactions that have been recorded at public transport stops within the circle. We assign each boarding transaction a value of -1, and each alighting transaction a value of +1, and find the running sum of the values by time of day. We subtract the minimum of the running sum from all its values, and use the resulting set of values as an indication of the number of activities that take place within the circle. For a given time of day the value of the running sum at each node is read from a table, and the average of these values across all nodes in the path is used in the regression.

Smart card transaction rate per roving sq. km (log): for each node in the path between two stops at a given time of day, we calculate the 15 minute moving average of the total number of transactions taking place per second within a 1 km² circle around the node, and use the average of these values across all nodes in the path. This value is used as an indication of the general level of traffic that the bus encounters along its path between stops.

A correlation analysis of the variables used in the estimation shows that some variables have high degrees of correlation for obvious reasons; for instance, the number of left and right turns passed or taken at intersections are obviously dependent on the number of intersections encountered. The variables describing the degree of ‘friction’ encountered along the path are also positively correlated, while path length and number of nodes within traffic control generally increase with the number of nodes in the path. Exclusive bus lanes also appear to be associated with stops with a large number of services arriving per day.

The model might therefore suffer from a significant degree of multicollinearity; however, estimated variable coefficients are stable for different sample sizes, and the addition of variables to the model do not lead to erratic changes in estimated values. But correlated variables could, arguably, be replaced by their first principal components to remove collinearity effects while retaining predictive power, as is done in principal component regression.

The current estimation of the model does not take account of spatial autocorrelation. An initial investigation of the residual (using a k-nearest neighbor approach on the destination stop in order to define the neighborhood matrix used in spatial autocorrelation models) does reveal a significant degree of spatial autocorrelation that varies by time of day, with a maximum value of Moran's I of 0.12. Effects of spatial autocorrelation will be further investigated in future studies, however as will be seen in the validation section to follow, the current ordinary least squares estimation already gives very reasonable predictions of speed and resulting passenger travel times.

The MATSim link dynamics model uses the predicted logarithm value of speed from the regression model for the given time of day as the mean for the normal distribution to sample the final speed value from, and the standard deviation of the distribution used for sampling is that of the residual for predicted speeds within 0.5 km/h of the mean.

6. Dwell time model

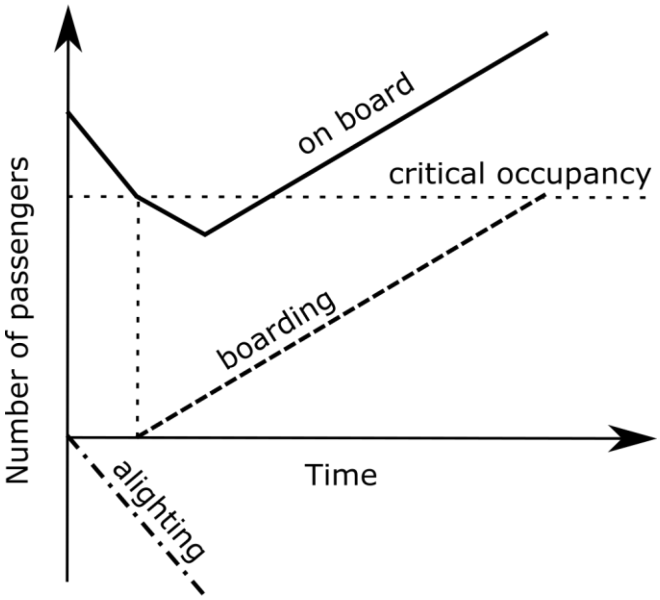


Figure 4. **Dynamics of passengers boarding and alighting from a crowded bus, revealing how boarding can only proceed once a critical occupancy has been reached.**

In (Sun et al., 2013) the authors show how different bus configurations translate into different rates of boarding and alighting. Furthermore, from a study of the Singaporean smartcard data and knowledge of the bus type associated with each vehicle identifier in the data, the authors have derived a model of dwell time variability as a result of boarding and alighting transactions and the number of passengers on board the bus. The most significant effects can be observed when the bus is very full and it is impossible for passengers to board until enough passengers have alighted, as can be seen in Figure 4.

We have incorporated this variable dwell-time model in our MATSim simulation. As will be seen later in the validation section, simulated results from the model fits very well with observed values.

5. Validation and performance

We ran our modified MATSim simulation model for 50 iterations, and compared various measures of system performance against the original smartcard data and the trajectory reconstruction-related data.

1. Speed

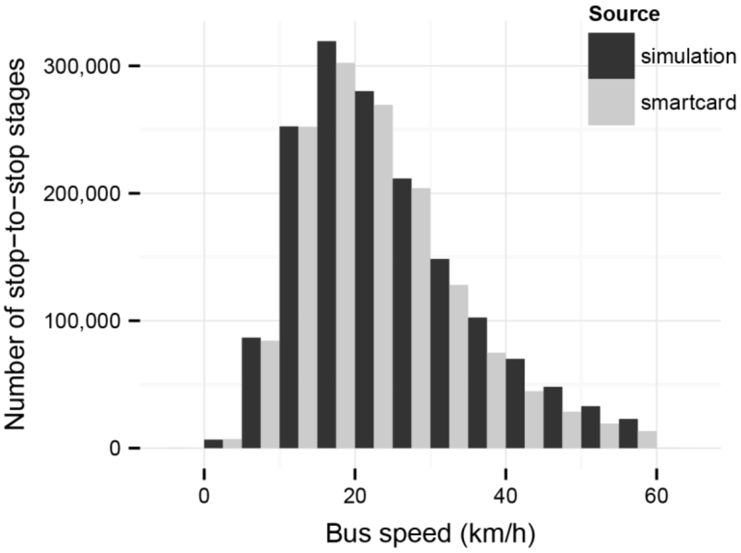


Figure 5. **Distribution of bus speeds from the smart card trajectory reconstruction process and the simulation**

Figure 5 compares the distribution of bus speeds between stops in the simulation against the speeds that were derived from the trajectory reconstruction process. Both the shape of the distributions and absolute numbers correspond very well.

2. Headways, dwell times and bus bunching

Figure 6 shows the distribution of headways in the simulation versus those derived from the trajectory reconstruction process. In its current state the simulation appears produce too many short headways; this is due to somewhat excessive bus bunching that occurs during the simulation, reducing the headway between consecutive buses to nearly zero.

In terms of headway variability, we see that the simulation produces increasing headway variability with increasing number of stops along the route, however the effect is much more pronounced in the simulation.

It can be seen from the joint distribution of headway versus number of stops along the route that the simulation produces many headways in the 0-1 minute bin, which indicates bus bunching. This behavior in the simulation is probably largely due to the fact that the first-in-first out queue dynamics of the simulation prevent buses of the same service from passing each other. We will therefore investigate passing behavior in future iterations, as buses of the same service can pass each other in reality when the first bus is already engaged in a dwell operation at a stop.

As our trajectory reconstruction process does not extrapolate the trajectories beyond the last recorded transaction for a circuit run, headways for the stops towards the end of a route might be inaccurate, which accounts for the lighter shading of the joint distribution of headway versus stop number in the smart card data. However, the distribution of the headways does appear considerably narrower for the smart card data than what the simulation produces. We are not aware of any bus bunching control measure in operation for the buses, whether it be centralized control from the operations center, or by intelligent actions of the bus drivers themselves. Such measures would naturally account for the increased reliability of services. However, it would also be worth investigating if allowing buses of the same service to pass each other when one bus is already occupied at a bus stop, serves as a bunching control measure in itself.

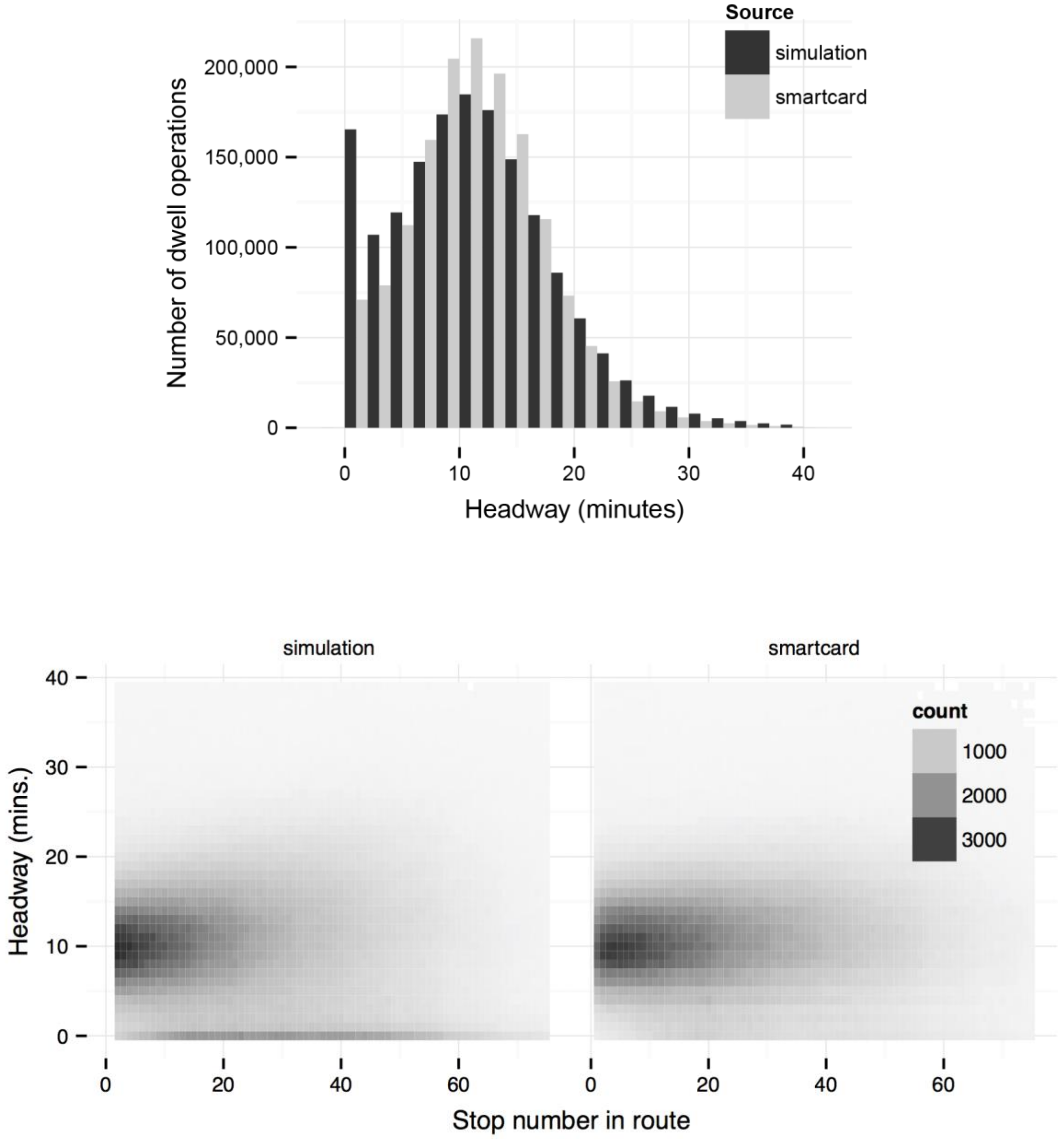


Figure 6. **Comparison of bus headway distributions and headway variability with increasing stop number along the route**

In Figure 7 the dwell time of buses in the simulation is compared against those derived from the trajectory reconstruction process. In terms of absolute numbers, nearly 1 million dwell operations with zero length occur in the simulation; these are cases where no boarding or alighting transactions take place. In the trajectory reconstruction, dwell operations that have been interpolated are assigned a zero dwell time. Dwell operations where only a very small number of transactions are recorded within a time span of less than six seconds, are assigned an arbitrary minimum dwell time of that value, which is responsible for the second spike in dwell times that we can see in the histogram. In terms of absolute numbers, the sum of these trivial cases for the smartcard data corresponds reasonably well with the number of dwell operations in the simulation where no passengers board or alight.

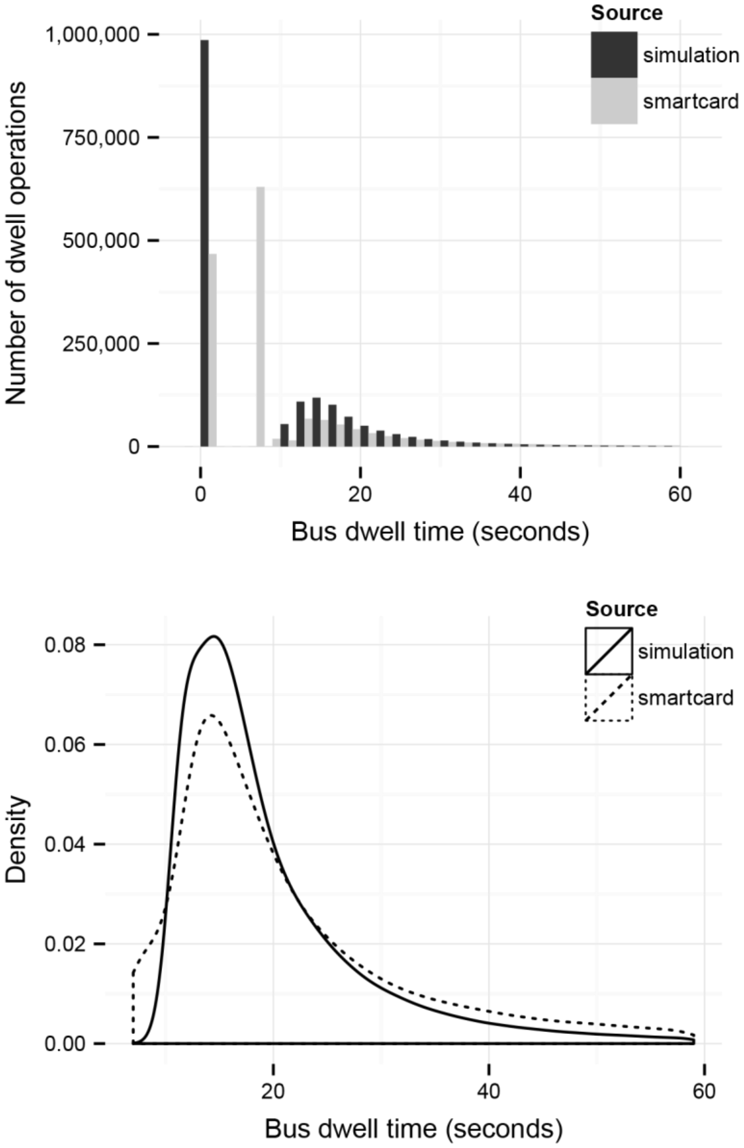


Figure 7. **Bus dwell time histogram and density comparison of non-trivial cases.**

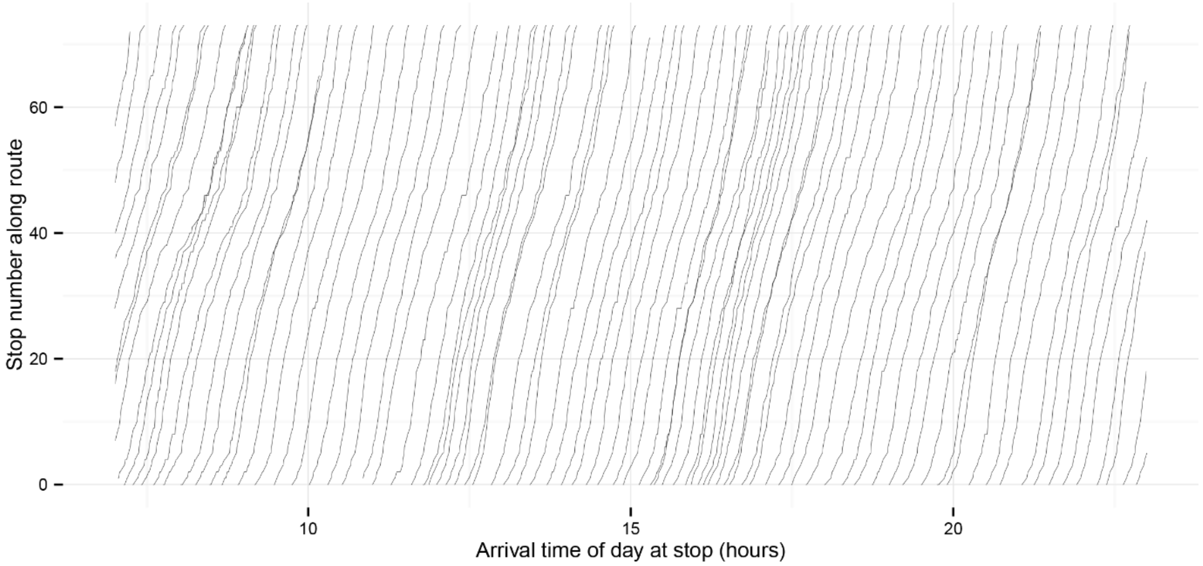
Because the absolute number of dwell operations for the nontrivial cases are different for the simulation and the smartcard data, we compared the distributions in terms of their density in the second part of the figure, which reveals reasonably good correspondence in terms of distribution between the simulation and the dwell times from the trajectory reconstruction process.

The trajectory reconstruction process produces 1.58 million dwell operations compared with the 1.7 million dwell operations recorded in the simulation; the difference between these numbers is due to the fact that the trajectory reconstruction process does not extrapolate the trajectories of buses beyond the first and last recorded transactions. So, if the first recorded transaction for a bus occurs at a stop after the first in its route profile, or the last recorded transaction is before the end of the line, then no dwell operations are created for the stops before the first transaction, or after the last transaction.

Figure 7 also shows that, of the 1.7 million dwell operations in the simulation, nearly one million have a zero duration, meaning that no passengers were picked up or dropped off. This means that buses in the simulation only pick up or drop off passengers approximately 40% of the time. Consequently, the simulation also produces more dwell operations of longer duration, as fewer dwell operations have to serve the same number of passengers. This might be a contributing factor to the higher incidence of bus bunching observed in the simulation.

If we assume that the actual total number of dwell operations also comes to 1.7 million, then the number of cases where buses don’t take on any passengers at stops for that particular day in the actual transport system comes to approximately 580,000, which accounts for approximately 34% of all dwell operations, meaning that buses in reality pick up or drop off passengers 66% of the time, in comparison to the 40% observed in the simulation. This difference might be due to the best response routing in the simulation resulting in increased coordination between agents and buses, with agents selecting services that get them to their destination with less access waiting time on average than the service that they picked in reality. Agents might also not be as averse to crowding as people in reality, causing them to opt for the next empty vehicle less frequently; a hypothesis that will require further investigation into the ridership of vehicles in the simulation versus those in reality.

The space-time diagram shown in Figure 8 compares the trajectory reconstruction results against the simulation for a bus line with 74 stops along its route. While the shapes of the trajectories compare reasonably well, we can see that the simulation produces more bus bunching than what this bus line experienced in reality, confirming what we saw in the histogram in Figure 6.



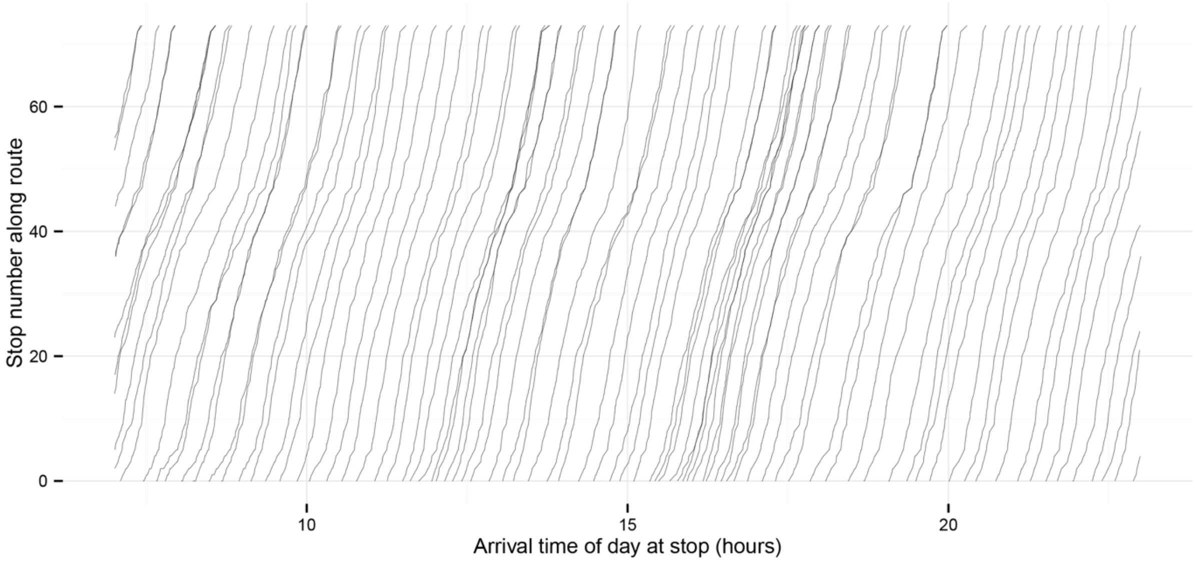


Figure 8 **Comparison of space-time trajectories of CEPAS (top) versus simulation (bottom)**

3. Passenger travel time measures

Figure 9 compares the trip travel time from the simulation with that of the smartcard data, where access and egress walking and waiting times have been extracted from the times recorded in the simulation. The histogram therefore compares only the sum of in-vehicle travel times, and transfer walking and waiting times.

Figure 10 similarly shows very good agreement between the bus stage in-vehicle times for the simulated versus actual values, although smart card values appear slightly skewed to longer times. While the simulated speeds are stochastic, in order to display the same range of values as those observed in reality, it is possible that not all dynamic effects have been captured adequately for perfect agreement, or that agents are routed more optimally than passengers are in reality. As we do not know when passengers board or alight from trains, we cannot construct a similar graph for rail modes. However, the good agreement that we see for trip travel time across all modes gives us confidence that the simulation of the rail mode is reasonably accurate, as passengers would have switched away or switched to using the subway during the simulation if this transport mode performed markedly different from reality.

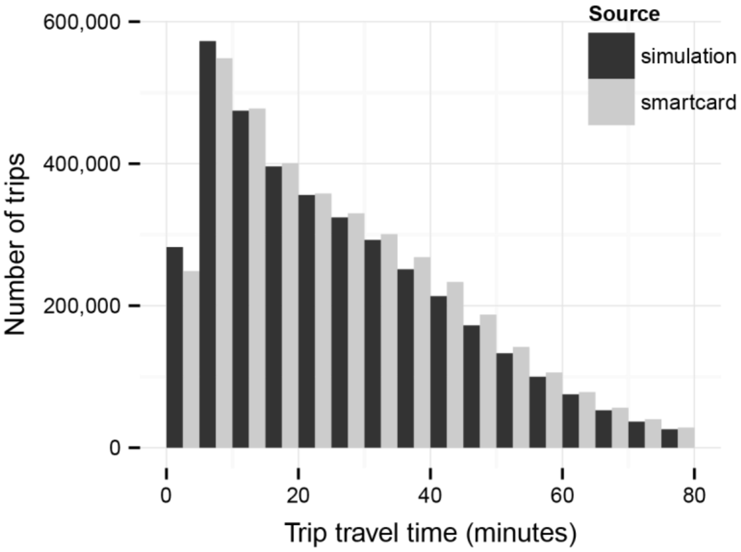


Figure 9. **Comparison of simulated versus actual trip travel times across all modes of public transport (excluding access and egress times).**

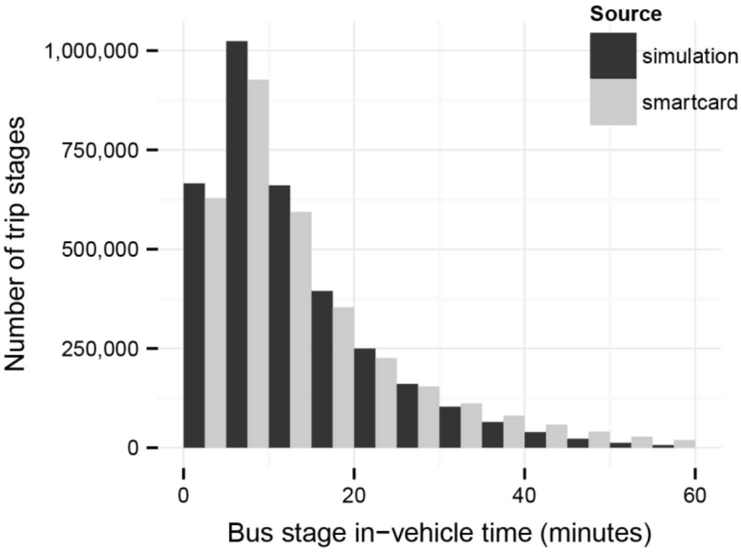


Figure 10. **Comparison of simulated versus actual bus stage-in vehicle time.**

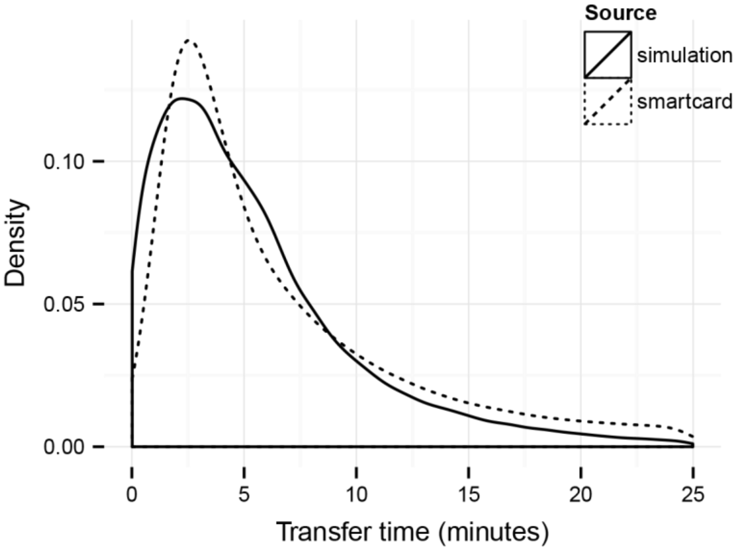


Figure 11. **Comparison of transfer time density in the simulation against that derived from the smartcard data**

Figure 11 compares the density of transfer times in the simulation against that which was derived from the smartcard data. In this case we do not display the histogram of transfer times, as the absolute numbers inferred from the smartcard data are inaccurate; especially for the train mode we do not know exactly which routes passengers have taken, nor exactly how long they have spent in transfer. The absolute numbers suggest that times in MATSim might be somewhat shorter than those experienced in reality, possibly due to the coordination that occurs due to best response re-routing during the simulation, as was alluded to earlier. We will need to revisit this part of the validation at a later stage, once we have reconstructed train trajectories from the smartcard data.

4. Computation times of simplified simulation

Using only best response re-routing, the simulation reaches a relaxed state in very few iterations. After only five iterations very little change in the average score of agent plans can be observed with increasing iterations. From experience we found that we only need to run a

25% sample of all agents in order to get realistic results; all counts recorded in the validation section have therefore been scaled up by multiplying them by four.

We ran our experiments on a latest generation 24 core Intel Xeon computer, with 64 GB of RAM. The initial routing of all agent plans takes approximately seven minutes, while a single iteration takes approximately four minutes. It is therefore possible to have usable results in under an hour. In the case of a standard MATSim simulation where we simulated both public and private transport, many more iterations are required for the system to reach a relaxed state, and a full simulation take up to two days to complete. The simplified simulation therefore represents a big step forward in terms of computation time performance.

6. Application

To show the potential of the simplified public transport simulation we designed a fictitious study case. In the proposed scenario we split one of the longest bus line in Singapore, which has more than 90 stops. The line was split according to the method used in (Lee et al., 2012) in order to minimize the number of transfers resulting from the split; in this case the optimal split point happens to be close to the center of the route. Agents are prepared to re-route their public transport routes within the MATSim co-evolutionary algorithm until they reach equilibrium (100 iterations). That means the agents who were taking the long line or any other line in Singapore can decide to take the new split line or switch to another transit line. As in the case of the validation study, we simulated a 25% sample of the population, with vehicle carrying capacities reduced to a quarter of their real-world values. In the following section we compare the performance of the line split against the baseline case.

1. Impact on bus bunching

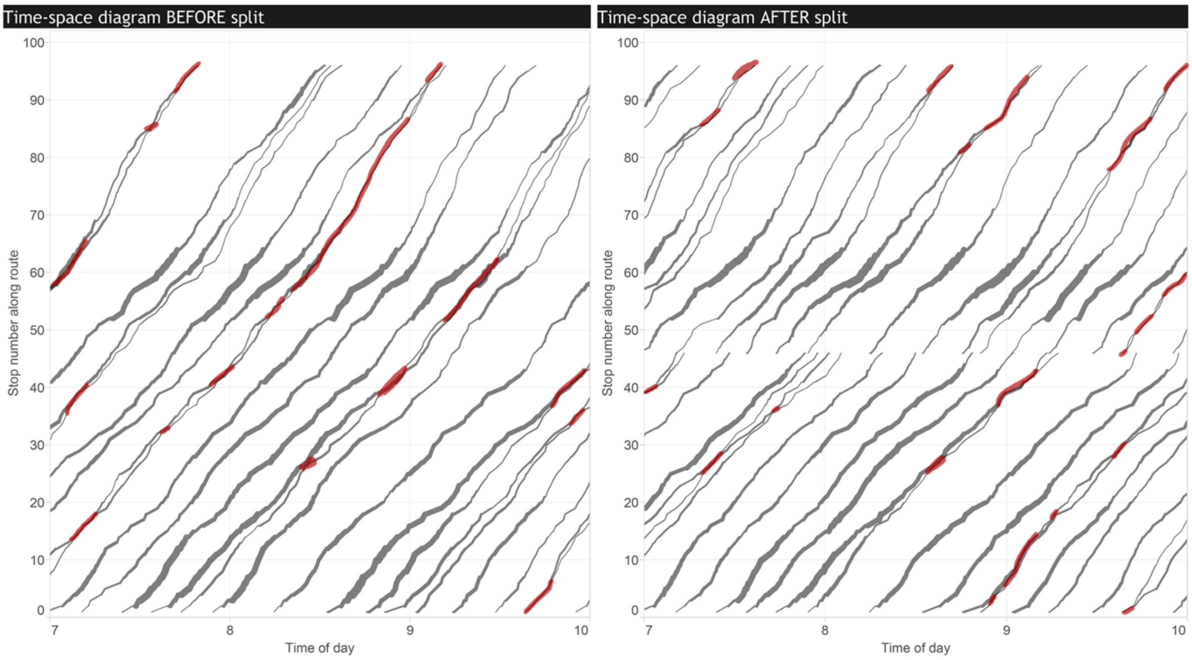


Figure 12 **Space-time diagram of a bus line before and after the line has been split in two.**

Figure 12 shows the space-time diagram of the bus service before and after the split, with cases of bus-bunching highlighted in red, and line thickness increasing with bus ridership. The plot confirms that incidences of bus bunching is significantly reduced during the morning peak hour, and that headway reliability is improved considerably, especially towards the end of the bus route. Note that we replicated departure times from the start of the service for buses departing on the second part of the line split, which means that but these services start with an inherent lack of reliability. Furthermore, even though we have reduced the number of stops in the two resulting routes, it is clear in the base case that bus bunching can result relatively early and that 45+ stops might still be too many bus stops for a reliable bus service.

2. Excess waiting times

Excess waiting time (EWT) is one of the most common reliability indicators for high frequency public transport services (e.g. a service frequency of five or more buses per hour). Using the definitions used by the London transport authorities, EWT assessment includes calculation of the following two elements:

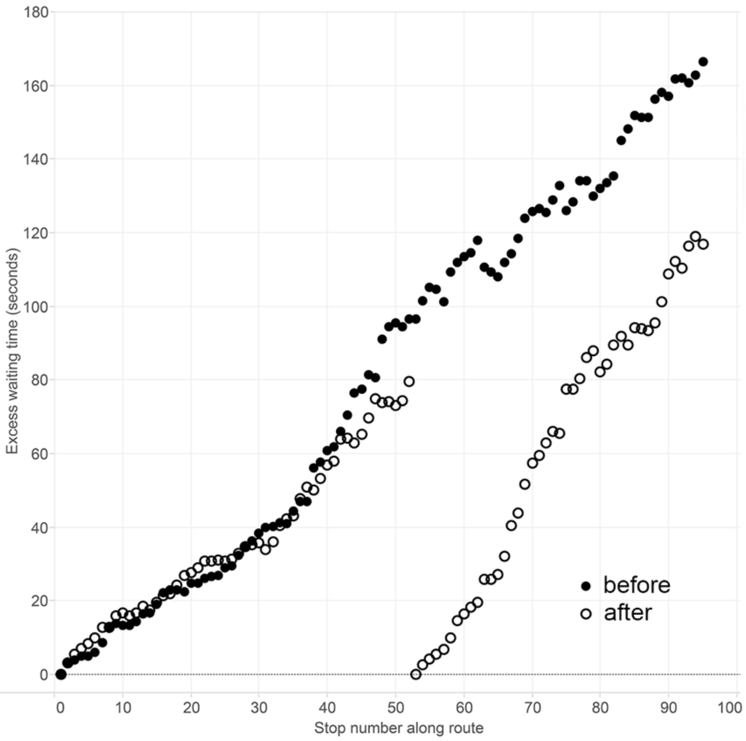


Figure 13. **Comparison of the excess waiting time before and after a long bus line has been split into two separate routes.**

Average scheduled waiting time (SWT): the time passengers would wait, on average, if the service ran exactly as schedule, assuming that waiting time is, on average, half of headway time:

$$SWT = \frac{\sum_{s \in S} {H_s}^2}{2\sum_{s \in S} H_s}$$

Average actual waiting time (AWT): the average time that passengers actually waited,

$$AWT = \frac{\sum_{s \in S} {H_a}^2}{2 \sum_{s \in S} H_a}$$

Where s represents each service of a bus line (excluding the first one), H_s is the scheduled headway of the service s and the previous service, and H_a is the actual headway of the service s and the previous service. EWT is simply the difference between AWT and SWT and represents the additional waiting time experienced by passengers.

The formulas have this form because AWT and SWT are weighted averages of all the service headways of a line, and the weight is the actual headway. So, if the line is designed to have a constant headway, the calculation of SWT can be simplified to $SWT = 0.5H_s$.

Figure 13 compares the calculation of the EWT of the base case against the split line scenario, in one direction of travel. The plot looks very similar in the opposite direction; EWT reverts to zero at the point with the line split, and consequently passengers experience much better reliability towards the end of the route.

7. Conclusion and future work

From the section on validation, our results so far appear to agree well for most part with actual observation. Most importantly, the simplified simulation manages to capture dynamic bus bunching effects; in fact, the effect might be slightly exaggerated in the simulation, and the possibility of mitigating this effect through the implementation of passing behavior in the queue simulation will be investigated. The simple fictitious case study also illustrates that the simplified simulation can be used to evaluate proposed changes to the public transport system.

The reconstruction of train trajectories is a very interesting problem as train-to-train transfers are not explicit in the CEPAS data. Furthermore, we need to locate public transport passengers to buildings that are close to public transport stops, in order to better simulate access walking and waiting times.

Our subway stops are also easily accessible in the simulation, and do not take account of the time that it takes for passengers travel all the way down to station entrances. Consequently, there is a slightly increased preference for the rail modes in the simulation compared to reality. We will attempt to address the shortcomings in upcoming versions of the simulation. Furthermore, we intend to perform case studies to predict the influence of newly introduced bus and train services and compare our results with smartcard data where these newly introduced services have been in operation for a number of months.

It has been our experience that the first-order analyses and operations on the smart card data, such as the conversion to trajectories, the speed regression and the models of boarding and alighting, are relatively straightforward to implement. The task of integrating results into a simulation model capable of providing insight into future transport system performance was a radically more involved task. The bugbears of systems engineering, namely unanticipated interactions and emergent phenomena, come into play even in this highly simplified integrated model, because of the dynamic and disaggregate interacting nature of the agent-based simulation. In design iterations leading to the current state of the system, we have spent many hours tinkering with its components in order to isolate cause and effect, and a number of challenges still remain, as we have highlighted throughout the section on validation of results.

Whether the integrated modelling approach ultimately proves to be worthwhile in predicting future transport system performance or not, the value of smart card data during all stages of the design and evaluation is undeniable. Whenever confronted with unexpected behavior in the simulation, we find ourselves constantly turning to the data for answers, trying to infer what actually happens in reality. We expect that this will become even more the case as data can be potentially enriched with bus GPS traces, and the mystery of where passengers are in the train system during the time between transactions at station entrances is revealed through rigorous statistical analyses, and the promise of coordinated data from underground cell phone transceivers.

References

Bagchi, M., White, P., 2004. What role for smart-card data from bus systems? Municipal Engineer 157, 39–46.

Bagchi, M., White, P.R., 2005. The potential of public transport smart card data. Transport Policy 12, 464–474.

Balmer, M., Rieser, M., Meister, K., Charypar, D., Lefebvre, N., Nagel, K., 2009. MATSim-T: Architecture and simulation times. Multi-Agent Systems for Traffic and Transportation Engineering 57–78.

Charypar, D., Nagel, K., 2005. Generating complete all-day activity plans with genetic algorithms. Transportation 32, 369–397. doi:10.1007/s11116-004-8287-y

Fourie, P.J., 2014. Reconstructing bus vehicle trajectories from transit smart-card data. Working paper 986, ETH Zurich, Institute for Transport Planning and Systems .

Grömping, U., 2006. Relative importance for linear regression in R: A vignette for relaimpo.

Lee, D.-H., Sun, L., Erath, A., 2012. Determining Optimal Control Stop to Improve Bus Services Reliability. Presented at the 1st European Symposium on Quantitative Methods in Transportation Systems, Lausanne, Switzerland.

Lindeman, R.H., Merenda, P.F., Gold, R.Z., 1980. Introduction to bivariate and multivariate analysis. Scott, Foresman Glenview, IL.

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: Emerging Technologies 24, 9–18. doi:10.1016/j.trc.2012.01.007

Pelletier, M.-P., Trépanier, M., Morency, C., 2011. Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies 19, 557–568. doi:10.1016/j.trc.2010.12.003

Prakasam, S., 2008. The Evolution of e-payments in Public Transport - Singapore’s Experience. Japan Railway & Transport Review 50, 36–39.

R Core Team, 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

Rieser, M., 2010. Adding transit to an agent-based transportation simulation. Ph. D. Thesis, Technical University Berlin, Berlin.

Sarlas, G., Axhausen, K.W., 2015. Localized speed prediction with the use of spatial simultaneous autoregressive models. Presented at the 94th Annual Meeting of the Transportation Research Board.

Smith, L., Beckman, R., Baggerly, K., 1995. TRANSIMS: Transportation analysis and simulation system. Los Alamos National Lab., NM (United States).

Sun, L., Lee, D.-H., Erath, A., Huang, X., 2012. Using Smart Card Data to Extract Passenger's Spatio-temporal Density and Train's Trajectory of MRT System, in: Proceedings of the ACM SIGKDD International Workshop on Urban Computing, UrbComp '12. ACM, New York, NY, USA, pp. 142–148. doi:10.1145/2346496.2346519

Sun, L., Tirachini, A., Axhausen, K.W., Erath, A., Lee, D.-H., 2013. Models of Bus Boarding/Alighting Dynamics and Dwell Time Variability. Transportation Research Part A: Policy and Practice 69:447-460.

Author biographies

Pieter Fourie is a simulation modeller at the Future Cities Laboratory in Singapore, specializing in transportation. Previously he developed integrated transport/land-use models for decision support to South African metropolitan planning authorities at the South African Council of Scientific and industrial Research (CSIR). His interests are focused on improving the performance of agent-based transport simulation, and simplifying its implementation through the use of alternative data sources, such as transit smart card and mobile phone data. He has used bus smart card data from Singapore to reconstruct bus trajectories and developed models of bus speeds, as well as in the implementation of this work to produce the first implementation of the smart-card driven agent-based simulation described in this chapter.

Sergio A. Ordonez M. is a Colombian computer scientist and mechanical engineer specialized in modelling and simulation. He worked in projects of national defense and traffic simulation in Bogotá before joining the Future Cities Laboratory in Singapore under the supervision of Kay Axhausen from ETH Zurich. In his dissertation he addresses the problem of simulating large scale urban transport scenarios during long periods of time, to study work-leisure human cycles. Multi-day public transport smartcard records and common travel surveys are the main sources of his activity-based model. His interests are focused on modelling and simulation of complex systems, big-data mining and visualization.

Artem Chakirov is an associate researcher in the areas of Mobility and Transport Planning at FCL. His current work focuses on mobility pricing in urban areas. Previously Artem was also involved in demand generation for Singapore transportation model and analysis of public transport smart card data.

Dr. Alexander Erath leads the Engaging Mobility group at the Future Cities Laboratory (FCL) of the Singapore-ETH Centre. In this role, he led the first implementation and further development of the large-scale, agent-based transport demand model MATSim Singapore an initiated the idea of using Smart Card Data for agent-based simulation. His main research interests are surveying and modelling of travel behaviour such as quantifying the impact of the built environment on mobility and transport demand modelling. He obtained his PhD from ETH Zurich (Swiss Federal Institute of Technology) where he studied the vulnerability of transport infrastructure.

Dr. K.W. Axhausen is Professor of Transport Planning at the Eidgenössische Technische Hochschule (ETH) Zürich. Before he worked at the Leopold-Franzens Universität, Innsbruck, Imperial College London and the University of Oxford. He has been involved in the measurement and modelling of travel behaviour for the past 30 years contributing especially to the literature on stated preferences, micro-simulation of travel behaviour, valuation of travel time and its components, parking behaviour, activity scheduling and travel diary data collection. Current work focuses on the agent-based micro-simulation toolkit MATSim (see www.matsim.org).

Keywords for Index

MATSim, agent-based, simulation, transport planning, transport modeling