2014 IEEE 17th International Conference on
Intelligent Transportation Systems (ITSC)
October 8-11, 2014. Qingdao, China

Stochastic Bus Traffic Modelling and Validation Using Smart Card Fare Collection Data

Jordan Ivanchev[1], Heiko Aydt[2], Alois Knoll[3]

Abstract— **This paper provides guidelines for designing bus transit simulations and justifies the need of a qualitatively and quantitatively correct model that is validated and can serve as a benchmark testing platform for the numerous optimisation strategies that researchers present. It describes and comments on some of the most important contributions regarding bus transportation control and optimisation from a modelling perspective. Furthermore, a case study concerning a bus line in Singapore is conducted where smart card data is used to extract the parameters needed in order to construct the model, while discussing the methods and challenges of this process. Consequently, the simulation is calibrated and validated in order to ensure maximum accuracy of its outputs.**

I. INTRODUCTION

Transportation networks play a vital part in every urban area. The design and control of such systems is particularly challenged by their temporal dynamics (rush hours, weekdays vs. weekends etc.). This which introduces extreme traffic conditions. Moreover, with ever growing cities the existing infrastructure is bound to accommodate an increasing number of agents (both vehicles and people). A well functioning public transportation system can reduce the number of private cars in the city's road network. This, however, leads to a severe load on the system itself, which turns its design and constant improvement into a both important and challenging problem.

Bus transit control presents numerous challenges which mainly stem from the high degree of traffic interaction that the system exhibits with the rest of the traffic participants. It has to cope with increased intensity of both cars and people during rush hours. This manifests in high deviations from the predefined schedules. A common issue that arises on such conditions is bus bunching. It accounts for the phenomenon that buses tend to bunch together even though dispatched with a predetermined headway. This results in them arriving at stops in the later stages of their trips in couples or even triples. This occurrence reduces their efficiency in terms of waiting time for the passengers. Moreover, it also leads to empty trips of the vehicles. There are a number of factors that are believed to contribute to this phenomenon such as non-static people arrival rates at the bus stops, fast varying traffic conditions, severely decreased ratio between travel time and dwelling time.

[1]TUM CREATE, 1 CREATE Way, 10-02 CREATE Tower, Singapore 138602 jordan.ivanchev@tum-create.edu.sg
[2]TUM CREATE, 1 CREATE Way, 10-02 CREATE Tower, Singapore 138602 heiko.aydt@tum-create.edu.sg
[3]Technische Universität München (TUM), Institute for Informatics, Robotics and Embedded Systems, Germany knoll@in.tum.de

Large numbers of researchers have suggested various techniques to avoid or at least diminish the problem. Very little or almost no attention is, however, given to the models that are used to evaluate and validate the results of the suggested methodologies. Every proposed technique utilises its own modelling approach, which makes it virtually impossible to compare it to other strategies. This clearly calls for a robust widely accepted model that should be used as a common evaluation platform in the future.
There are five main issues that model designers should have in mind:

1) Oversimplified models might not grasp the complexity of the problem: for example assuming uniform placement of bus stops, velocity of buses, and volume of people arriving at each station could altogether lead to no bus bunching at all just as assuming uniform working hours distributed along the day may lead to no rush hours and traffic jams.
2) Quantitatively mismatched models might solve different problems than the targeted ones: if the number of people that use the bus service is significantly different from the actual one the phenomenon might not be observed at all. In the same way if the velocities, boarding times and OD matrix do not represent reality, the optimal control strategies may not be realistic either.
3) Qualitatively inaccurate models can lead to intrinsically imprecise results: For example if the wrong types of distributions are assumed for bus velocities or people arriving at the stations, the distributions of the output parameters that characterise the system might also be qualitatively different from what is observed in reality. This would make the evaluation of a control strategy statistically invalid. Moreover, provided that transportation networks are complex systems, it is widely accepted that small differences in the initial conditions and parameters can lead to much larger changes in the outputs of the system.
4) Computationally intense models might be unnecessary: Even though micro and nanoscopic simulations can provide valuable insights into certain type of systems, sometimes a much simpler model might lead to virtually identical results. In the case of bus movement we are not looking for any emerging behaviours nor for multidimensional outputs that require high level of granularity. Therefore simplified models that take several orders of magnitude less computational time

978-1-4799-6078-1/14/$31.00 ©2014 IEEE

to produce the same results should be favoured.

5) Calibration and validation: 1), 2), and 3) combined point to the statement that in order to properly evaluate a control strategy the model should produce accurate results. Calibration and validation are standard steps in modelling that are rarely taken in the existing literature on bus movement modelling.

The main contributions of this paper are:

- A comprehensive quantitatively and qualitatively sound bus movement model that can be used by researchers or transportation authorities as a medium for comparing various optimisation techniques . A particularly new approach utilised in the described simulation framework is the modelling of the correlation between the transit times of adjacent buses that further decreases the discrepancies of simulation output and reality.
- A robust and simple technique for calibration of the model and its subsequent validation. Almost none of the models in the existing literature are calibrated and none are validated.
- Extensive description of the techniques for extracting model parameters from a standard smart card data set, which simplifies the calibration procedure and increases the precision and efficiency of the underlying model. Previous research on this topic is scarce and the field will potentially significantly expand with the increased utilisation of smart sensing systems.

II. LITERATURE REVIEW

In this section we examine the relevant literature on the topic of optimising bus transit. We, however, look at it mostly from a modelling perspective. As mentioned in the previous section, in most cases every author uses a self-defined and developed model for testing of various strategies. This review does not aim at evaluating the optimisation techniques but rather looks at the models described.

Research on vehicle control strategies and thus modelling of bus movement originated from the analytic model of [19], which is a simplistic model that allows convenient mathematical analysis. Further refinements and extensions were suggested in the following years by [17], [4], [5], and [6]. All of those early models assume fairly simple transit networks, consisting of a service loop, with just a small number of vehicles operating on it. Using this framework, control policies are analytically defined. Even though, the simplicity of such approaches allows for analytical work, the numerous assumptions turn into limitations that make the models inapplicable in real world situations.

Next, the stochastic models in [3] and [16], which calculate the first and second moments of vehicle travel times in order to analyse the robustness of bus movement, were introduced. The modelling of transit networks and route performance have been further addressed in analytic models in [22] and [16]. A correlation between transit times is introduced by [2]. In [8], an extensive analytical study on a schedule-based transit system is performed by deriving a set of integral equations to describe the arrival and departure time distributions. An objective function is designed to describe the trade off between the service cost and the deviation from the schedule. In [12], a comprehensive analytical model is presented that uses most of the previously developed analytical modelling efforts that came before it as a foundation.

With the increased amount of data available to services operators from *Automated Vehicle Location* (AVL) systems, *Automated Passenger Counting* (APC) and *Computer Aided Dispatching* (CAD) systems, intelligent transportation has gained popularity. Furthermore, with the increase of available computational resources the simulation approach is preferable in order to evaluate real time control strategies. In [13], the author describes in great detail how to extract important parameters from AVL data. He, after that, uses those parameters to define a model for the movement of the buses and the boarding and alighting procedures.

In [10], the authors use AVL, AVI and APC data to correct errors in their deterministic bus movement model. The whole day is split into different time regions with deterministic behavior. The parameters needed to describe the behavior are extracted from the available data. This technique may be slightly prone to lag during the day, since it corrects the behaviour for the next time period using the data from the previous. Depending on the length of the regions, this could have a negative effect on the accuracy and predictive power of the model. Moreover, it might be more desirable to simulate the bus movement using a stochastic approach. In the case of the deterministic one, especially when it is corrected in short intervals, it might become over fitted. This means that even though the error is very small, the generalization and predictive properties of the model are compromised.

In [25], a Monte Carlo simulation approach is taken in order to evaluate a time control point strategy for robust bus schedule design. It uses AVL data to calibrate a real bus route in China. After that a Monte Carlo method is applied in order to evaluate the expectation of the deviation from the scheduled arrival time. It considers only evening rush hour time-zone, which provided the fact that bus movement in big cities is considered a complex system (a small change in initial conditions can lead to much greater discrepancies in the final results) might not produce representative results. It should also be pointed out that AVL data provides only a few measuring points along the bus route, resulting in the need for interpolation of the missing data.

Efforts in the utilisation of smart card data include [15] where linear interpolation of the buses' positions is used in order to gain information about the buses' spatio-temporal movement. Uniform distribution of passenger arrival is assumed. In addition to that the work described in [20] uses Korea smart card analysis to extract useful data about the traffic in the city in general. They have GPS data as well which significantly easens the process.

Another important parameter that needs to be modelled is the passenger arrivals at stops. Stochastic arrival and

boarding processes of passengers are described in the models of [14] and [1]. For large headways, some studies show that schedule-based control may be better than headway-based control, because passengers usually become aware of the schedules. As stated in [18], a headway of 12 to 13 minutes usually is the treshold, where the majority of people become aware of the schedule.

In [7], the authors developed a model representing passenger arrival times that are determined by the agent's choice with respect to both service frequency and reliability. The model was constructed based on utility functions and was calibrated by empirical data. In [11], the OD matrix is used to generate the passengers and deal with the boardings and alightings. Uniform distribution of people arriving at the station is assumed rather than Poisson, which is the widely accepted method. In [23], the authors use Poisson distribution for arrival and binomial distribution for alighting of passengers. Following are some efforts that utilise active control of the buses included for completeness. We look at them from the perspective of bus movement modelling.

A mixed model approach taking the bus capacity as well as holding and stop-skipping strategies into account was proposed in [9]. In [26], messages are sent between the buses to optimise the waiting time of the passengers both on and off the bus. The model assumes identically distributed stop to stop times. In big cities where buses also travel on highway roads, this assumption might not hold. Similar to previous works a fixed portion of the passengers on the bus is considered to be the number of alighting passengers. This makes the OD matrix uniform in space, while real world data clearly shows that some stops are significantly favoured to others by the passengers, which is one of the main reasons for bus bunching.

In [21], a bus bunching reduction strategy is thoroughly studied. The model assumes homogeneous route (equal stop to stop distances) and homogeneous demand of the passengers. They also assume that the buses are identically send out throughout the day and that the they stop at every station regardless of demand, which in reality only holds for about 50% of the stops. A continuous approximation model is proposed. It includes the dwelling time of the buses into the average cruising speed. This is a simplification that speeds up the simulation while still providing adequate results for the traffic flow. It, however, reduces the amount of information available for analysis coming from the model and its granularity. For example it will be impossible to get statistics about the waiting times of the agents and the effects of the bus transportation on traffic conditions.

In [24], dynamic bus holding strategies are proposed to achieve schedule reliability. A comparison study of different control methods is also conducted. Two assumption that are made are not fully reasonable especially regarding mega-cities: time of day independent transit times and time of day independent passenger arrival rates. Those practically remove rush hours and result in a traffic intensity, which is homogeneous throughout the day. This might barely be the case only during weekends. Passenger alighting times are discarded from the dwelling time of the bus and the buses stop at every station. That assumption might on its own have a reduction of the bus bunching problem since it is considered holding.

III. MODEL DESCRIPTION

We start by introducing the notation that will be used throughout the paper. Refer to Figure 1 for a basic outline of the model:

- μ_b - boarding time mean for a passenger
- σ_b - boarding time standard deviation for a passenger
- μ_i^j - mean of travel time between stop i and $i-1$ in time region j
- σ_i^j - standard deviation of travel time between stop i and $i-1$ in time region j
- H_i^m - headway of m-th bus at stop i to the $m-1$-th bus
- T_i^m - time for bus m from the $i-1$-st station to i-th station
- pb_i^m - number of passengers boarding bus m at stop i
- pa_i^m - number of passengers alighting bus m at stop i
- B_i^m - boarding time for the m-th bus at stop i
- A_i^m - alighting time for the m-th bus at stop i
- L - average loss time of bus due to deceleration and acceleration at a stop
- D_i^m - dwelling time for the m-th bus at stop i
- c^j - correlation of headway in the j-th time period
- M - number of buses for the whole day
- S - number of stops on the route
- R - number of time regions the day is split in
- R_l - length of a time region
- f^j - frequency of buses starting during time period j
- w_n^j - waiting time of the n-th person in the j-th time region
- λ_i^j - Poisson distribution parameter for people generation at stop i in time period j
- C_i^j - number of passengers that boarded a bus on station i in time region j
- OD^j - OD matrix for period j
- $tp_{i,j}^k$ - the arrival time of passenger k at stop i in time period j

A. Generation of people at the stations

Passengers at the stations are generated using a Poisson process. This is a good approximation in cases where the schedule is not well-maintained and/or people do not have access to it. More than that, it represents well situations in which the frequency of buses is quite high (in the order of 8 to 10 buses per hour), since in these cases the agents do not plan their arrival at the bus stop. Given that the aforementioned conditions hold, in mega-cities the Poisson assumption can be made. In case of low bus frequency and availability of a schedule, a mixture of two distributions can be used that is described in [7]. The arrival intensity parameter is extracted from the collected smart card data. This is explained in more detail in the next section. The time

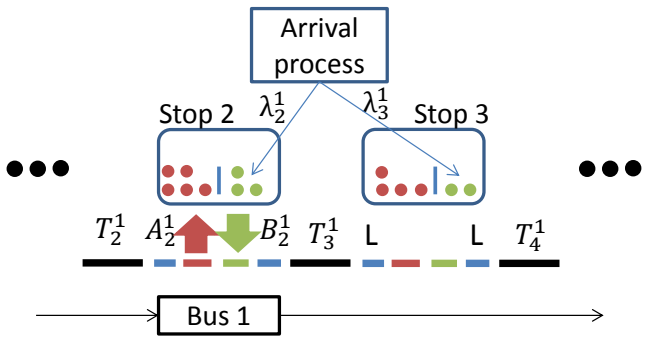

Fig. 1. A sketch representing the basic functioning of the whole model. People are arriving at the station as explained in section A governed by the parameter λ_i^j. The bus arrives at the station, losing time L for deceleration and for the return to traffic after the stop. The time spent at the station is then calculated by either adding up the alighting time A_i^m and the boarding time B_i^m or taking the maximum of the two. The travel time T_i^m is then sampled from the respective distribution.

of arrival of the next passenger on the bus stop is calculated using the following process:

$$tp_{i,j}^k = tp_{i,j}^{k-1} + \frac{\log(U[0,1])}{R_l \lambda_i^j} \tag{1}$$

When an agent is generated at the bus stop, he is also given a destination stop which is sampled from the OD matrix, extracted from the smart card data. This is a vital step since it provides the model with information about people alighting the bus thus improving the precision of modelling the dwelling time at the stations. In Figure 2 the arrival of people at the stations throughout the day extracted from the simulation output is presented.

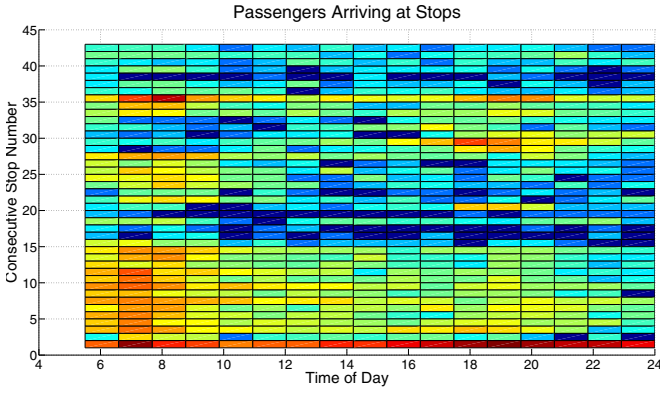

Fig. 2. People arriving at stops for a full day. Red colours indicate high intensity of people arrival and blue indicate low intensities. It can be noted that people board this particular line mostly on its first stop. More than that, the morning and evening rush hours can be observed with higher passenger arrival intensities.

B. Modelling dwelling time of buses at stations

The dwelling time of a bus can be divided into three portions; time loss due to deceleration before the stop and acceleration (entrance of the road) after the stopping as shown in Figure 1. These are the time losses compared to the case when the bus skips the stop. The second part is the alighting time for passengers and the third is the boarding time of new passengers. The latter might be combined into a single term depending on the system that is being modelled. In the case of Singapore, alighting and boarding happen through different doors and we might use the following expression for the dwelling time:

$$D_i^m = L + \max(B_i^m, A_i^m) \tag{2}$$

where

$$A_i^m = \sum_{k=1}^{pb_i^m} \log \mathcal{N}(\mu_b, \sigma_b) \tag{3}$$

$$B_i^m = \sum_{k=1}^{pa_i^m} \log \mathcal{N}(\mu_b, \sigma_b) \tag{4}$$

If the passengers alight before boarding begins one might prefer to use the sum of the two terms. The alighting and boarding times for every passenger are sampled from a lognormal distribution with mean and deviation extracted from the data. The loss time L cannot be extracted straight from the data therefore the simulation has to be calibrated against it.

C. Simulation of bus movement from station to station

This is the main part of the model. It deals with modelling the transit times of the bus from leaving a stop until it reaches the next one. For each of the R time periods the day is divided into, a distribution for the transit times is extracted from the data. Every time a bus travels to the next station the transit time is sampled from the respective distribution. This aims at modelling rush hour conditions slowing down the buses and thus achieving a realistic congestion temporal profile. It is, however, important to note that an additional factor is added. This is the correlation of the transit time of the bus to the time of the previous bus, which is computed as follows:

$$T_{i,j}^m = T_{i,j}^{m-1} + \mathcal{N}(0, \sigma_i^j \sqrt{2(1-c_j)}) \tag{5}$$

Here the time of the previous bus is taken and a sample from a 0 mean normal distribution is added. In the case that the bus is the first one for the respective time period the transit time is just sampled from the original distribution:

$$T_{i,j}^1 = \log \mathcal{N}(\mu_i^j, \sigma_i^j) \tag{6}$$

The assumption that the transit time distribution is lognormal is verified by performing the *Anderson-Darling* goodness of fit test on the collected samples for every distribution for every stop to stop segment for the respective time zones. In 95 % of the cases the null hypothesis that the distribution is not lognormal is rejected with an average p value of 0.72. The results for Normal, Gamma, Exponential, and Pareto distributions show worse goodness of fit measures.

The deviation of the distribution is a scaled version of the deviation of the original distribution. It decreases if the buses are correlated and increases in the opposite case. For practical purposes it has been demonstrated that the correlation coefficient c_j is in the region $[0, 1]$. It is dependent on the headway of the previous bus. Meaning that if the buses are travelling through the road network in high proximity to each other (if they are bunched together), they would experience the same traffic conditions and their transit times will be more correlated. Surely the travel times of buses might also be higly correlated in extreme cases of congestion or in the other extreme, if the roads are emtpy. This phenomenon is modelled by the small deviation of the travel times distribution for the periods of the day when such conditions might apply. The headway is defined as follows:

$$H_i^m = H_{i-1}^m + D_i^m - D_{i-1}^{m-1} + T_i^m - T_i^{m-1} \tag{7}$$

$$H_1^1 = 0 \tag{8}$$

The correlation coefficient dependent on the headway is computed using an exponential envelope:

$$c_{i,j}^m = exp(-\beta_j H_{i-1}^m) \tag{9}$$

The factor β_j is calibrated against the data.

There are three sets of parameters affiliated to the model: parameters extracted from smart card data, parameters that are calibrated and validated against the data, and output parameters computed by the model as shown in Table I. It

TABLE I
DIFFERENT TYPES OF PARAMETERS ASSOCIATED WITH THE MODEL

Extracted Parameters	μ_b, σ_b, μ_i^j, σ_i^j, λ_i^j
Calibrated Parameters	L, β_j
Input Parameters	f^j, R_l, R
Output Parameters	All other parameters

is important to note that, provided all the output parameters described above are available, a detailed analysis of both the passengers' and buses' trips is available: waiting time of the passenger, boarding and alighting time, total trip duration (from arriving at the station until alighting the bus), time spent on a moving bus etc. In Figure 3 a sample of the bus trajectories produced by the simulation can be seen.

IV. DATA EXTRACTION

The extraction of the parameters needed by the model is a challenging task requiring a robust and methodological approach. Therefore it will be thoroughly discussed in this section. Moreover with the increase of smart card integrated transportation systems, researchers are nowadays increasingly involved in performing statistical analysis using such kind of data.

The format of the samples is passenger oriented, opposed to AVL data for example where buses are tracked. This makes it suitable for OD matrix estimation and for analysis of the transportation demand in the city. However, in order to

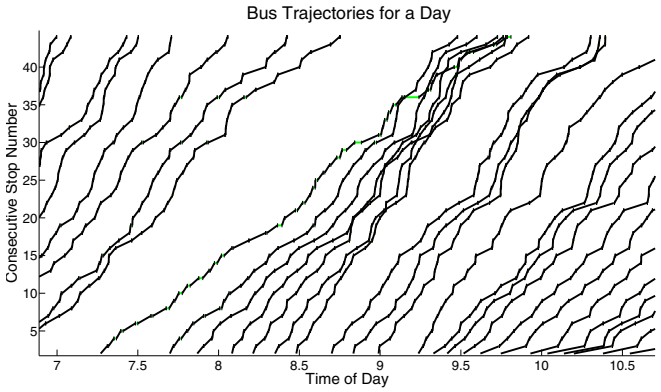

Fig. 3. Bus trajectories produced by the simulation during morning rush hour. The green horizontal lines indicate boarding of passengers on the stops. The phenomenon of bus bunching can be clearly observed resulting in larger periods of time where no buses are available for the passengers.

have a good model, information about the buses' movement must be extracted as well.

There are several major challenges that imply the need for a robust data analysis pre-step prior to modelling. Since there is no bus id recorded in the sample format, it is troublesome to extract the path of a specific bus in an error-free manner. Information can be extracted about a bus being at a certain stop, those points of "knowledge" can, however, not be connected in a bus movement trajectory. Another issue comes from the human factor involved and from the infrastructure of the smart card recording system; in case of underground transportation network people tap in and out when they enter the station which makes it almost impossible to extract information about the arrivals and departures of trains or at least severely increases the error of such type of estimation.

The data provided to us comprises of the recorded usage of a bus line in Singapore for a week in 2011 and is of the following form: [Used Line, Start Station, End Station, Boarding Time, Trip Distance, Trip Duration, Trip Date].

A. Estimation of passenger arrival rates at stops

As described in the previous section, the arrival of passengers at stops is modelled by a Poisson process. The only parameter needed to be specified is the intensity λ_i^j of people coming to the bus stop. The parameter, however, varies significantly both in time (rush hours) and in space (stops exhibit non-uniform demand). Therefore, a different value is extracted for every stop and for every predefined time period of the day. The procedure consists of collecting all passengers that board in the respective time-zone on the respective stop and divide their count by the duration of the time period.

$$\lambda_i^j = \frac{C_i^j}{R_l} \tag{10}$$

Moreover, the OD matrix is also estimated at this step which involves simply counting the number of people taking

the trip for every possible origin-destination pair. Once again, a separate OD matrix OD^j is extracted for every time period.

B. Estimation of bus dwelling times

To estimate the dwelling time of buses at stations we need to extract the boarding/alighting time distribution parameters in order to describe the log normal distribution. To get the boarding time of every passenger we take the difference of the time stamp to the previous one. The means μ_b^j and deviations σ_b^j of this parameter do not demonstrate any correlation to the time of the day, which allows us to use just one distribution: $\ln\mathcal{N}(\mu_b, \sigma_b)$, that is valid for the whole timespan of the simulation rather than extracting the parameters for every separate period of the day. The distribution for alighting time of passengers is equivalent to the distribution for boarding time.

The second parameter that needs to be adjusted is the time loss L of the bus due to deceleration for the stop, opening doors, returning onto the road. This parameter is calibrated against the simulation rather than extracted from the data. More details are provided in the calibration section.

C. Weighted average time extraction technique for obtaining stop to stop times

This subsection describes the main and most challenging part of the data processing step. The distributions of times between consecutive stops in different time regions are extracted. The procedure is as follows:

1) Extract the arrival and departure times of each bus are in the following way. First combine the departures and arrival times of the passengers. Every trip contributes with four entries ([boarding station, time-stamp, alighting station, time-stamp] or [alighting station, time-stamp, boarding station, time-stamp]). After that this structure is sorted according to the first entry and its time-stamp.
2) Define the maximum time t_{tr} between boarding/alighting of passengers . If the difference between the timestamps of passengers is smaller than the threshold time, the respective passenger is added to the current bus, otherwise the previous passenger is declared the last one to board the bus and the departure time is set to be the time-stamp of this passenger. The arrival time of the next bus is set to be the time-stamp of the next passenger.
3) Combine the people boarding and alighting at stations into buses. Every bus has a stop id, arrival time, and departure time: $B_k = \{i, t_a^k, t_d^k\}$. This is a list of buses at stations where K is the number of unique stoppings of buses ($k \in \mathbb{N}[1, K]$)
4) Combine the separate stoppings of the buses into bus trajectories.
   a) Define the number of buses as the number of separate bus stoppings at the first stop. Create a tree of B_ks: T_B The head node of every such tree is the B_k at the first stop on the route. Also create a set that contains all the B_ks: S_B. Remove the head node of the tree from the set.
   b) Expand every node on the tree in the following way: Find all passengers for which the boarding or alighting time is between t_a^k and t_d^k and either alighting or boarding station equal to i for the currently expanded node.
   c) For every such passenger take the second part of the trip (the one that is not in the examined node) and find the stopping where its time-stamp falls. Check if the stopping is in the list S_B. If yes, add this stopping as a child node to the node that is being expanded and remove the stopping from S_B.
   d) Continue expanding the child nodes of the tree while it is possible. The order of expansion does not matter.
   e) After a tree is completely expanded, order its nodes by the stop ids i_k and the resulting list is the trajectory of the k-th bus $TR_k =< B_{k,1}, B_{k,2}..., B_{k,N} >$ where N is the number of stoppings of this bus. Figure 4 depicts the tree expansion process.

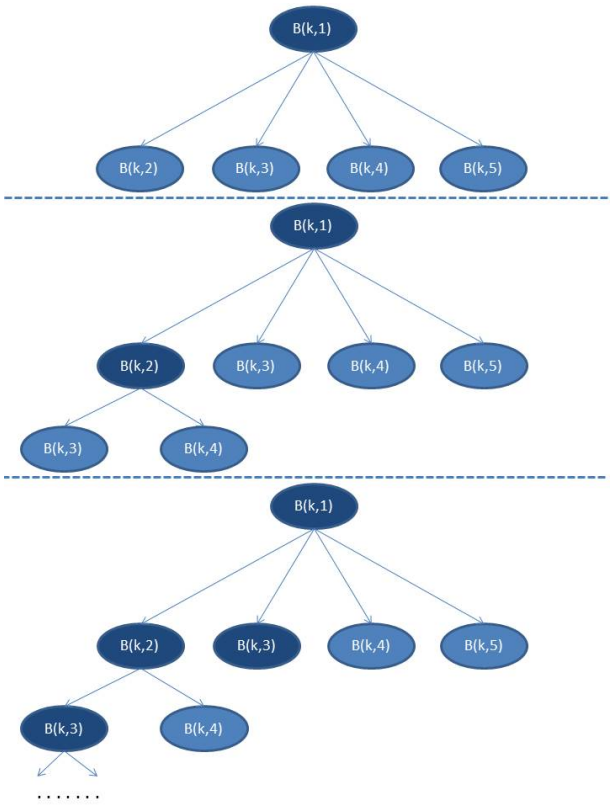

Fig. 4. Tree expansion process. Dark coloured nodes depict already expanded nodes, while light coloured nodes will be expanded in the future. Three expansion steps are shown.

5) Every passenger belongs to exactly one of the buses. Next, go trough all the trips and find the bus entry that the passenger belongs to at his departure and

arrival stop. Then the trip time of the bus (time on the road, excluding dwelling time at stops) is calculated by subtracting the departure time of the bus from the initial stop from the arrival time of the bus at the final stop and also subtracting the time the bus lost at the intermediate stations. If we assume that the departure station for the passenger with id p has an id s, the ending id e and that the bus id is k:

$$T_{road}^{p,s,e} = t_{d,s}^{k} - t_{a,e}^{k} - \sum_{i=s+1}^{e-1} t_{d,i}^{k} - t_{a,i}^{k} \tag{11}$$

6) After we have the pure trip times of the buses (excluding boarding times) we use them to estimate the time from every stop to the next one. This is done in the following way: Every extracted trip has a duration T_{road}^{p}, start stop s, end stop e, time-zone r, calculated using the average between the trip start and end time. Every trip contributes to every stop to stop time distribution that it contains with a weight proportional to the partition of the distance that this stop takes. The contribution to the time between stops s and $s+1$ coming from $T_{road}^{p,s,e}$ is:

$$T_{s,s+1}^{p} = T_{road}^{p,s,e} \frac{d_{s,s+1}}{d_{s,e}} \tag{12}$$

and the weight of this contribution is:

$$w_{s,s+1}^{p} = \frac{d_{s,s+1}}{d_{s,e}} \tag{13}$$

In this way we get greater weights when the trip is only one stop and smaller weights when the trip consists of many stops.

7) After that we have measurements for every stop to stop distance with their respective weights and time-zones. We extract the distributions for every time-zone using weighting average and weighted standard deviation.

Following this procedure ensures that we use information from every single trip in the data, which naturally maximises the efficiency of our data extraction step.

D. Calibration of correlation between bus times and time loss at bus stops

Having extracted all the desired distributions and parameters from the data, only two parameters are left that need to be calibrated against the simulation. The smart card data at our disposal is for the duration of one week. We discard the data for the weekend days and extract the input parameters of the model using the data for the first three days of the week.

The starting times of the buses are averaged over the starting times observed in the data which can then be used to calculate the frequencies of dispatching vehicles in every time region or just use the extracted starting times in the model. The two parameters to be calibrated are also specified before execution of the simulation. The simulation is then run and all the data is collected.

We define the error between simulation and the smart card data (that should be minimized: $argmin_{L,\beta_j} err(SIM, EZLINK)$) in the following way: Iterate through the trips and try to fit every trip into the results of the simulation. Look for the closest bus (in time) that starts at the trip's starting station. Assume bus id k, start stop s, end stop e, and passenger p. Define a range of travel times as:

$$t_{range}^{p,s,e} = [t_{d,e}^{k} - t_{a,s}^{k}, t_{a,e}^{k} - t_{d,s}] \tag{14}$$

Then compute the error (in case the trip time does not fall in the region) for every trip. We use the normalized mean squared error to combine all the results into one number that is the error $err(SIM, EZLINK)$ of the simulation compared to the data. We make use of a genetic algorithm in order to find the optimal values of the parameters. The error for the fixed parameters is averaged over 100 runs of the simulation with those parameters and has a mean value of 0.01329 and deviation 0.00182.

E. MODEL VALIDATION

For the model validation we use the same procedure as for calibration. The data used to average the data for the calibration consists of 60% of the available data, we use the rest of the data (the two days that are left from the weekdays) that the model has not yet "seen" to validate our calibration and data extraction procedures.

This, however, still allows for the utilisation of the same methodology, the vital point of validation is that the model is tested on an unknown data set, so that we avoid over fitting of the parameters. It is crucial that the model is properly tested since this provides a proof of its quantitative and qualitative validity, which has not been rigorously presented in the literature until now. The error is averaged over 100 runs of the simulation and has a mean value of 0.01451 and a deviation of 0.00154.

V. CONCLUSION

This work examines bus transportation in mega-cities from a modelling perspective. The proposed simulation aims at providing the maximum amount of functionalities while keeping the level of assumptions to a minimum, thus ensuring precision of the acquired results. It is designed in order to create a valid, fast, and accurate simulation that can be used as a benchmark testing platform for comparing various techniques for improving bus transportation. Such a benchmark can allow for more structured and fair assessment of new ideas or at least show the need of a concept such as a unified model that can be used in the future.

The proposed simulation was described, its parameters were extracted from raw smart card data and the secondary parameters were calibrated against this data. Consequently, it was validated. Moreover, a number of techniques used for the extraction of the model parameters from smart card entries were described, which we believe will be helpful in the future due to the fact that such type of data will soon if not already be a primary source for calibration of transportation models.

The data structure used in the case study for Singapore is in its large part standard for smart card data collection. It might be even considered that it is a worst case scenario for such type of data due to the absence of bus id entries, which significantly complicates the parameter extraction step.

Future work might include extending the model so that researchers become able to easily adapt it to their experiments. Such extensions might include: communication between buses, integration of real time data streams that continuously update the simulation parameters so that the model can be used to model advanced dynamicallly controlled and sensor rich smart transportation networks. Another extension that would require some effort would be to simulate more than one bus line. In such a case a synchronisation mechanism between the buses that share a common road segment should be implemented. This would make buses aware of each others presence. Then the dwell time can be modelled in a more comprehensive way. Moveover, people will be provided with a wider choice of buses to board which will further increase the precision of the bus bunching modelling.

The problem of extracting valuable information with minimal error from smart card data is quite novel and largely unexplored until now. We believe that even more sophisticated algorithms could be designed to further improve the quality of the models.

Finally, the model and the extraction techniques described are robust enough to be used in any megacity that posseses a smart card transportation system. In case of an absence of such a system given that the parameters needed can be extracted in other forms the model itself can naturally be used on its own.

VI. ACKNOWLEDGMENTS

This work was financially supported by the Singapore National Research Foundation under its Campus for Research Excellence And Technological Enterprise (CREATE) program.

REFERENCES

[1] Andrzej Adamski. Probabilistic models of passengers service processes at bus stops. *Transportation Research Part B: Methodological*, 26(4):253–259, 1992.

[2] O Adebisi. A mathematical model for headway variance of fixed-route buses. *Transportation Research Part B: Methodological*, 20(1):59–70, 1986.

[3] Perke Andersson and Gian-Paolo Scalia-Tomba. A mathematical model of an urban bus route. *Transportation Research Part B: Methodological*, 15(4):249–266, 1981.

[4] Arnold Barnett. On controlling randomness in transit operations. *Transportation Science*, 8(2):102–116, 1974.

[5] Arnold Barnett and Daniel J Kleitman. On two-terminal control of a shuttle service. *SIAM Journal on Applied Mathematics*, 35(2):229–234, 1978.

[6] Arnold I Barnett. Control strategies for transport systems with nonlinear waiting costs. *Transportation Science*, 12(2):119–136, 1978.

[7] Larry A Bowman and Mark A Turnquist. Service frequency, schedule reliability and passenger wait times at transit stops. *Transportation Research Part A: General*, 15(6):465–471, 1981.

[8] Malachy Carey. Extending a train pathing model from one-way to two-way track. *Transportation Research Part B: Methodological*, 28(5):395–400, 1994.

[9] Felipe Delgado, Juan Carlos Munoz, Ricardo Giesen, and Aldo Cipriano. Real-time control of buses in a transit corridor based on vehicle holding and boarding limits. *Transportation Research Record: Journal of the Transportation Research Board*, 2090(1):59–67, 2009.

[10] Xu Jun Eberlein, Nigel HM Wilson, and David Bernstein. The holding problem with real–time information available. *Transportation science*, 35(1):1–18, 2001.

[11] Liping Fu, Qing Liu, and Paul Calamai. Real-time optimization model for dynamic scheduling of transit operations. *Transportation Research Record: Journal of the Transportation Research Board*, 1857(1):48–55, 2003.

[12] Mark D Hickman. An analytic stochastic model for the transit vehicle holding problem. *Transportation Science*, 35(3):215–237, 2001.

[13] Antoneta X Horbury. Using non-real-time automatic vehicle location data to improve bus services. *Transportation Research Part B: Methodological*, 33(8):559–579, 1999.

[14] JK Jolliffe and TP Hutchinson. A behavioural explanation of the association between bus and passenger arrivals at a bus stop. *Transportation Science*, 9(3):248–282, 1975.

[15] D Lee, Lijun Sun, and Alex Erath. Study of bus service reliability in singapore using fare card data. In *12th Asia-Pacific Intelligent Transpotation Forum*, 2012.

[16] PHJ Marguier. *Bus route performance evaluation under stochastic conditions*. PhD thesis, Ph. D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1985.

[17] Gordon Frank Newell. Control of pairing of vehicles on a public transportation route, two vehicles, one control point. *Transportation Science*, 8(3):248–264, 1974.

[18] Maurice M Okrent. *Effects of transit service characteristics on passenger waiting time*. PhD thesis, Northwestern University, 1974.

[19] EE Osuna and GF Newell. Control strategies for an idealized public transportation system. *Transportation Science*, 6(1):52–72, 1972.

[20] Jin Young Park, Dong-Jun Kim, and Yongtaek Lim. Use of smart card data to define public transit use in seoul, south korea. *Transportation Research Record: Journal of the Transportation Research Board*, 2063(1):3–9, 2008.

[21] Joshua Michael Pilachowski. An approach to reducing bus bunching. 2009.

[22] Warren B Powell and Yosef Sheffi. The load planning problem of motor carriers: Problem description and a proposed solution approach. *Transportation Research Part A: General*, 17(6):471–480, 1983.

[23] Aichong Sun and Mark Hickman. The real–time stop–skipping problem. *Journal of Intelligent Transportation Systems*, 9(2):91–109, 2005.

[24] Yiguang Xuan, Juan Argote, and Carlos F Daganzo. Dynamic bus holding strategies for schedule reliability: Optimal linear control and performance analysis. *Transportation Research Part B: Methodological*, 45(10):1831–1845, 2011.

[25] Yadan Yan, Qiang Meng, Shuaian Wang, and Xiucheng Guo. Robust optimization model of schedule design for a fixed bus route. *Transportation Research Part C: Emerging Technologies*, 25:113–121, 2012.

[26] Jiamin Zhao, Satish Bukkapatnam, and Maged M Dessouky. Distributed architecture for real-time coordination of bus holding in transit networks. *Intelligent Transportation Systems, IEEE Transactions on*, 4(1):43–51, 2003.