2018 21st International Conference on Intelligent Transportation Systems (ITSC)
Maui, Hawaii, USA, November 4-7, 2018

Generating Synthetic Passenger Data through Joint Traffic-Passenger Modeling and Simulation

Rongye Shi[†], Peter Steenkiste[‡†], Manuela Veloso[‡†]
†Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA
‡School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
rongyeshi@cmu.edu, prs@cs.cmu.edu, mmv@cs.cmu.edu

***Abstract*—Real passenger data available to city planners are usually incomplete. The goal of our work is to generate synthetic passenger data using a novel methodology that leverages joint traffic-passenger modeling and simulation on a city scale. A demonstration of such an idea in generating synthetic bus passenger data was implemented. Specifically, we 1) learned a bus passenger demand model from indirect people-mobility data to generate bus passenger demand samples, and we 2) developed a bus passenger behavior model, which runs jointly with a traffic simulator (SUMO), to generate synthetic bus passenger data. We applied the proposed methodology for a case study of Porto city, Portugal. The synthetic bus passenger data presents significant similarity in terms of spatial-temporal distributions to the real-world bus passenger data collected by the bus automated fare collection (AFC) system in Porto.**

***Keywords*—public transportation, simulation, synthetic data, behavior modeling, Poisson process, KL divergence**

I. Introduction

It has long been known that research into estimating passenger behavior and corresponding mobility patterns requires access to large-scale and multi-source human mobility data. The availability of human mobility data is increasing. Fortunately, thanks to advances in sensing technologies and the widespread use of automated data collection (ADC) in public transportation, it is possible to collect large quantities of diverse data on urban spaces and city populations, for example, vehicle global positioning system (GPS) data and automated fare collection (AFC) data.

Unfortunately, when it comes to passenger-related research, the data available to researchers are usually insufficient, either because the data are *incomplete* with important features missing or because the data are only *indirectly related* to the topic of focus (see Table I for more information). Complete data are usually lacking due to the challenges that urban infrastructures face in integrating large-scale multi-source data in a timely and low-cost fashion. This lack-of-complete-data issue limits passenger-related research. For example, in bus transportation systems, the bus passenger data are usually collected by automated passenger count (APC) systems or AFC systems. Unfortunately, the data collected by those systems are often incomplete (no alighting feature is recorded), limiting the estimation of the overall demand profile. More seriously, the origin-destination (O-D) survey is expensive in terms of human effort and financial cost. As a result, the development of many state-of-the-art methods for bus passenger estimation and prediction (e.g., [1], [2]) are unable to validate themselves because some necessary features in the real data are not available.

To cope with the issue, utilizing indirectly related data could be a way out. In general, indirectly related data are from a different source, with some features correlating positively to those in the unknown complete data. Grounded on these correlated features, we attempted to develop a method for generating synthetic complete data that are most likely to be observed in reality. Inspired by this idea, and as a main contribution of this paper, we proposed a joint traffic-passenger modeling and simulation methodology to generate synthetic passenger data based on other indirectly related people mobility data. To be specific, we demonstrated and verified the proposed method in the setting of bus transportation systems:

- We learned a bus passenger demand model from other indirectly related people mobility data to generate bus passenger demand samples. This was motivated by the insight that the people mobility trend reflected by the data of different sources can correlate positively to the mobility trend of real bus passengers;
- We developed a passenger behavior model to jointly run with a mature traffic simulator (SUMO) to generate city-wide synthetic bus passenger data;
- We implemented the methodology for a case study of Porto city, Portugal. The simulation outcomes were validated by measuring the distribution difference between the synthetic passenger data and the real bus Automated Fare Collection (AFC) data of Porto;
- Our work is the first successful attempt to transfer indirect people mobility data to complete bus passenger data by means of joint traffic-passenger modeling and simulation.

II. Background and Related Work

Passenger behavior modeling and simulation have often been used in transportation system research. To evaluate the performance of vehicle scheduling and platform deploying, such as selecting bus stop sites, the behavior of passengers must be simulated and analyzed in detail. Although they are sophisticated enough to take into consideration individual preferences [3], the seat allocation process [4, 5], and even pressure from passengers behind a person [6], most studies are highly microscopic, confining their domains to limited

This work was supported by the FCT under the Carnegie Mellon – Portugal ERI S2MovingCities project. The authors would like to thank Teresa G. Dias, António A. Nunes, and João F. Cunha for providing the AFC data.

TABLE I
TERMINOLOGIES

Term	Definition	Explanation and example
Data Domain	The features each data point presents (i.e., the feature space of the data) and the distribution of the data on these features (i.e., the distribution on the feature space)	The cat image data set has 100 pixels for each image, and the feature space is a 100-dimensional space. The feature space and the distribution of the cat images in this feature space determine the domain of the cat image data.
Complete Data	The data that are in a domain that is sufficient for solving a task	The set of cat images can solve the task to train a classifier to distinguish cat images from non-cat images.
Indirect Data	The data that are in a domain that is too different from the domain of complete data to solve the task	The dog image set is indirect data: it has 100 pixels for each image, but the distribution in the space is different, and it is insufficient for fully training a cat/non-cat classifier.
Indirectly Related Data	The data that are indirect data, and whose domain overlaps with or is similar to the domain of complete data (i.e., compared to complete data, some of the features are the same and the distributions on these features are similar)	The dog image set is indirectly related to the cat: its distribution on the feature space is more similar to the cat distribution than to that of other images, such as a vehicle, house, etc. Thus, the dog image data can help with partially solving the cat classification task by distinguishing a cat image from a vehicle image.
Trip Demand	A tuple (origin, destination, trip starting time)	A trip demand is (**O**, **D**, **t**).
Travel Plan	A set of midway O-D pairs without time information	A travel plan is {($\mathbf{O}$, D_1), (O_2, D_2),…,(O_n, **D**)}.
Travel Demand	A tuple (origin, destination, trip starting time, travel plan)	A travel demand is (**O**, **D**, **t**, {($\mathbf{O}$, D_1), (O_2, D_2),…,(O_n, **D**)}).
Passenger Trip Demand Model	The description of the distribution from which a passenger trip demand is generated	The distribution model specifies the probability of the occurrence of each trip demand in the demand space.
Experience	The sequence of circumstances and events that the passenger encounters during a trip, and their occurrence in time	During the trip from stop **O** to stop **D** starting at time **t**, the passenger may take several transits to get to stop D. Then, experience can be a set of tuples {($\mathbf{O}$, D_1, **t**), (O_2, D_2, t_2),…,(O_n, **D**, t_n)}. Here, t_n is the time of arriving at O_n.

numbers of buses and not scaling well to provide insights into the macroscopic flow of passengers in the entire city.

Compared with passenger modeling, traffic simulation, such as bus transportation simulation, has experienced rapid and significant developments. Many road traffic simulators, for example, VISSIM, AIMSUN, Matsim, SUMO, etc., are developed with good performance. One commonly used open-source traffic simulator is SUMO (Simulation of Urban MObility), which provides a platform for explicitly simulating vehicles, including cars, buses, and urban trains on a city scale. However, most traffic simulators are currently unable to provide information about passenger-vehicle interactions, which is of great interest in bus passenger behavior and prediction studies.

To fill the gap between passenger and traffic simulation, we propose a methodology for simulating bus passenger behavior in conjunction with the mature traffic simulator (SUMO) on a city scale. To the best of our knowledge, this is the first attempt to generate synthetic passenger data through city-wide traffic-passenger joint simulation.

III. METHODOLOGY

The problem of focus in this paper is as follows: *How can we make use of the knowledge learned from the indirectly related people mobility data to generate complete passenger data that are most likely to be observed in reality?*

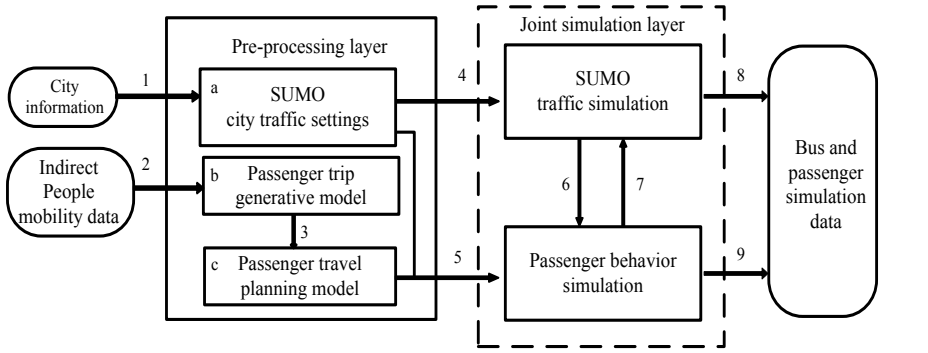
A. Importance of Combining Passenger Modeling and Traffic Simulation

Neither modeling passenger behavior nor simulating traffic can solve the aforementioned problem independently, leading to the idea of combining passenger modeling and traffic simulation. First of all, modeling passenger behavior specifies how people's travels are demanded and planned (e.g., the O_i and D_i in TABLE I), but it does not detail what passengers actually experience during travel in an urban traffic environment (e.g., the t_i in TABLE I). This missing experience can be supplemented with traffic simulation. Second, most traffic simulation provides representations of transportation systems and vehicle behaviors, especially how public transits operate in urban road networks. However, passenger-level travel demands/behaviors and the corresponding impact on the public transportation systems (e.g., bus dwell time affected by passengers) are unavailable. This can be supplemented with passenger modeling. Combining passenger modeling and traffic simulation is an effective way in which to employ the strong points of both approaches and overcome the shortcomings of either.

B. Overview of the Method

We provide a high-level overview of the proposed methodology in the setting of bus transportations system settings. Note that technical details may vary according to different cities and available data sources. The joint traffic-passenger modeling and simulation methodology is designed to thoroughly capture the interactions among passengers, buses, and traffic. Specifically, it simulates the behavior of bus passengers moving through the urban bus network while having the buses interact with the urban traffic environment. To avoid any misunderstandings, we define a passenger "trip demand" as consisting of the *trip starting time* and an *origin-destination (O-D) pair*; meanwhile, a passenger "travel demand" consists of the *trip starting time*, an *O-D pair*, and a specific *travel plan*. See TABLE I for details.

The methodology is presented in Figure 1, which is composed of two layers: a pre-processing layer and a joint simulation layer. The *pre-processing layer* is a collection of three components (denoted as *a*, *b*, and *c*), and they conduct city information importing, data learning, and passenger demand generating, respectively, to prepare for the joint traffic-passenger simulation in the second layer. Specifically, *Component a* serves to extract and convert city road infrastructure information from public resources into SUMO formats to establish a virtual city traffic network and define traffic demands. *Component b* serves to learn a passenger



1. Import real city information into SUMO
a. Establish the virtual city traffic network and define traffic demands
2. Load people mobility data to learn a passenger trip generative model
b. Generate passenger trip samples from the generative model
3. Forward passenger trip samples to the passenger travel planning model
c. Travel planning model generates travel plan for each passenger's trip
4. Load city traffic settings to SUMO for city traffic simulation
5. Load passenger demands for passenger behavior simulation
6. Notify passenger simulator to update passenger states
7. Update the bus dwell time according to on/off passengers
8. Store the final traffic/bus simulation log
9. Store the final passenger simulation log

Fig. 1. Joint traffic-passenger modeling and simulation block diagram for bus transportation systems.

Fig. 2. Virtual Porto traffic network in SUMO.

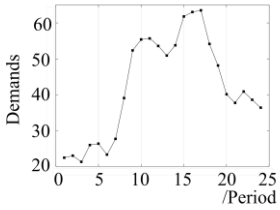


Fig. 3. Expected taxi demand on Wednesday.

trip generative model from people mobility data and generate passenger trip demand samples. *Component c* models the way in which a bus passenger plans to travel from the origin to the final destination, through the bus network. Finally, this component generates passenger travel demand that includes the trip starting time, an O-D pair, and a travel plan.

The traffic settings and bus passenger travel demands are fed forward to the *joint simulation layer*. In this layer, we have SUMO simulate the road traffic, including buses and other vehicles moving through the established urban road network. To be specific, a monitor-control algorithm (the passenger behavior simulation block in the dashed box in Figure 1) runs jointly with SUMO to monitor the states of the buses in real time and to simulate passenger behaviors accordingly. At the end of the simulation, this layer outputs detailed passenger traveling information and bus state information.

IV. Implementation

We applied the methodology for a case study of the bus transportation system in Porto, Portugal. This section details the implementation of the bus passenger modeling and simulation.

A. Bus Transportation System Establishment

The first step is to establish the urban bus transportation system in SUMO, which reflects the exact real world. Main bus service operator STCP provides a company website where detailed routes, geographical locations of bus stops, and timetable information are provided. As shown in Figure 2, using the STCP bus service information and other public resources (e.g., OpenStreetMap, etc.), we established the bus transportation system as well as the urban road network within the selected central city area of Porto. The imported bus network contains 136 routes, 855 bus stops, and 5,723 bus trips on a normal workday. It is confirmed by the simulation tests that the bus performance matches well with the actual Porto bus transportation system: each bus departs at the scheduled time, runs along its designated route, and pulls at designated stops correctly.

B. Bus Passenger Trip Demand Generative Model

The second component is to learn the passenger trip demand generative model for generating passenger trip samples. The goal of this model is to generate the trip demand tuple (O, D, t) of a passenger. The approaches used to establish the model highly depend on the source of data available. In many cases, direct and complete data in the target domain is not available, and a workaround is to proceed with *indirectly related* data. In this implementation, we used Porto taxi trajectory data to learn Porto passenger mobility distribution, and then to generate passenger trip samples from the distribution.

The taxi dataset [7] describes a complete year of the trajectories for all 422 taxis running in Porto city. Each data point contains several features in which we are interested: the 1) trip starting time; 2) date type (identifying whether the trip occurred on a holiday day); 3) call type (telling whether the trip started from the taxi operation central or on a random street); and 4) poly line (storing the GPS coordinate sequence of the trip trajectory). We selected the data with random street call types and removed holiday samples. The dataset contained detailed trip starting times and O-D pairs of random street passengers, making it a nice resource of city dwellers' travel trend. On the other hand, the selected area is the central city area of Porto, which has quite a dense bus network, and this setting mitigates the negative correlation between taxi and bus demand models by excluding areas that are poorly served by buses.

The proposed passenger trip demand model consists of two components, a temporal model and a spatial model. The *temporal model* is an inhomogeneous Poisson process model, which is widely used to model the occurrence of events in time. We consider the Poisson process to be inhomogeneous, and *rate parameter* λ varies in time. We fit the rate parameter on an hourly basis for a certain weekday. We divided a day equally into 24 periods and focused on studying all Wednesdays of the year. The average taxi demand in each period on Wednesday is shown in Figure 3, according to which we fit estimated Wednesday rate vector $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_{24})$. The temporal model is then described as:

in period i, interval τ between two consecutive passenger demands follows the following exponential distribution:

$$\tau \sim f\left(t; a(\hat{\lambda}_i + \sigma)\right) = a(\hat{\lambda}_i + \sigma)e^{-a(\hat{\lambda}_i+\sigma)t}. \quad (1)$$

Here, a is a coefficient to scale $\hat{\lambda}$, as the number of bus demands is usually greater than taxi demands (a is used to scale daily passenger demands up to around 150 thousands, which is suggested by the STCP 2016 annual service report [8]). In practice, we also introduced uncertainty into the model by adding small noise $\sigma \sim N(0,1)$ to $\hat{\lambda}$.

The *spatial model* is also learned on an hourly basis. The spatial model is a four-dimensional (4-D) distribution model from which we can generate 4-D samples with the first two components as origin (O^x, O^y) and the last two as destination (D^x, D^y). We applied kernel density estimation to fit the spatial model, using multivariate 4-D normal distribution as a kernel. This method is non-parametric and is effective when prior knowledge about the distribution is unavailable; thus, parametric methods don't apply well. The bandwidth is determined based on the normal distribution approximation [9]. Finally, we have the spatial model:

$$\hat{p}_{\mathrm{H}}(\mathrm{X}) = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{(2\pi)^{4/2}|\mathrm{H}|^{1/2}} \exp\left(-\frac{1}{2}(\mathrm{X}-\mathrm{D}_k)^T \mathrm{H}^{-1}(\mathrm{X}-\mathrm{D}_k)\right), (2)$$

where $\mathrm{H} = \mathrm{diag}(h_1, h_2, h_3, h_4)$ defines the bandwidth of each dimension, and $\mathrm{D}_k = (O_k^x, O_k^y, D_k^x, D_k^y)$ is the O-D demand of taxi demand data point k. The model generates two geographical points, and we searched for the closest bus stop near each point and used it as the origin/destination stop. We set a cut-off distance of 640 meters, with the stop matching outside of this region nulled. This distance is from the public transport accessibility levels (PTAL) methodology, which proposes insight that the longest distance a passenger would normally walk to access a bus service is within the range of an 8-minute walk at the speed of 4.8 km/h [10]. A temporal sample and a spatial sample constitute a passenger trip demand sample.

C. Passenger Travel Planning Model

The passenger travel planning model is to provide the set $\{(O, D_1), (O_2, D_2), \ldots, (O_n, D)\}$ of midway trips which lead the passenger from origin O to destination D. It is assumed that passengers always choose a plan that minimizes cost, distance, and unnecessary route switching, upon which, we designed a built-in bus *passenger travel planning model*.

As illustrated in Figure 4, we considered the bus transportation network as a *directed graph* G with vertices $V = \{v_i\}$ denoting bus stops and edges $E = \{e_i\}$ denoting the *reachability* between bus stops. For example, the blue edge from v_1 to v_2 indicates that stop v_2 is reachable from stop v_1 by taking route A. In the graph, route A (with blue edges) and route B (with red edges) share the transition stop v_3 and passengers can choose to switch routes there. Besides bus route edges, walking edges are introduced into the graph (see the yellow dashed edges in Figure 4). Stops within a certain geographical distance (e.g., 640 meters) are considered to be walking reachable stops, and passengers would be willing to walk a few more meters to transfer at those stops. In each bus route, vertices are designed to be fully connected. After establishing the bus transportation network in SUMO, we can measure the exact length of each edge, i.e., the weight of each edge denoted as d.

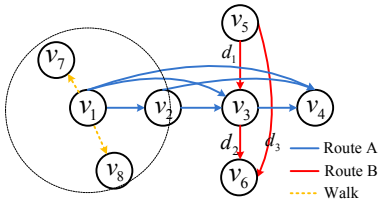


Fig. 4. Graph of bus transportation network.

Algorithm 1 Optimal travel plan searching algorithm

Input: s: source vertex; d: destination vertex;
{V, E}: graph; iniRoute: initial route of s;
Δ: route-switching penalty; ε :adding sub-trip penalty
Output: cost, trace

```
cost[s]← 0                                          %zero the cost of source vertex
s.preRoute← iniRoute                                %preRoute used to judge route switching
for all v in V − {s} do
    cost[v]← ∞                                      %the cost of non-source vertex is set to infinity
    trace.update({v:(s, infEdge)})                  %initialize trace-back record
S← ∅                                                %S: visited vertex set
Q← V                                                %Q: queue set (vertex set to be visited)
while Q ≠ ∅ and d not in S do
    u← minCost(Q,cost)                              %select vertex u in Q with minimal cost
    S← S + {u}
    Q← Q - {u}                                      %move vertex u from Q to visited set S
    for v_temp in u.outNeighbors do
        E_out← getEdges(u, v_temp)                  %examine outward edges
        for e in E_out do
            update_cost← cost[u]+e.weight+ ε        % basic cost
            if u.preRoute ≠ e.route then            %judge route switched or not
                update_cost← update_cost+Δ
            if cost[v_temp]>update_cost then        %when basic cost improves
                cost[v_temp]← update_cost           %store new cost value
                v_temp.preRoute← e.route            %update route information
                trace.update({v_temp:(u,e)})        %store the trace-back record
return cost, trace to d
```

Given the graph structure, a travel plan for an O-D pair consists of a set of edges, and we call each edge a *sub-trip*. We wanted to find an optimal plan that minimizes a certain cost object. In addition to the cost object in traditional shortest-path searching problems, which are based on accumulative distance only, we introduced 1) route-switching penalty Δ to penalize the route switching of a travel plan, and 2) penalty ε to penalize the object when a sub-trip (edge) is added into the travel plan. Introducing ε is beneficial: when considering the travel demand from v_5 to v_6 through route B in the graph (with $d_3 = d_1 + d_2$), we preferred the optimal travel plan to be represented as $\{\mathrm{edge}(v_5, v_6)\}$ rather than $\{\mathrm{edge}(v_5, v_3), \mathrm{edge}(v_3, v_6)\}$. By introducing ε, the plan $\{\mathrm{edge}(v_5, v_6)\}$ with cost $(d_3 + \varepsilon)$ will win over the plan $\{\mathrm{edge}(v_5, v_3), \mathrm{edge}(v_3, v_6)\}$ with cost $(d_1 + d_2 + 2\varepsilon)$. Based on the structure of the graph and the definition of the cost object, we designed the optimal travel plan searching algorithm in Algorithm 1, which is an advanced version of the Dijkstra algorithm [11] with a more sophisticated cost object. The output of the algorithm is the passenger's plan to move from the origin stop to the destination stop through the bus transportation network.

D. Bus Passenger Behavior Modeling + SUMO Simulation

Given a bus passenger travel demand (containing the trip starting time, an O-D pair, and a travel plan), we needed to simulate how the passenger moves through the bus transportation network and interacts with buses and traffic to ultimately reach the destination. For example, one of the core functions of the joint simulation layer is to fill the boarding time t_i for each subtrip tuple (O_i, D_i). In this layer,

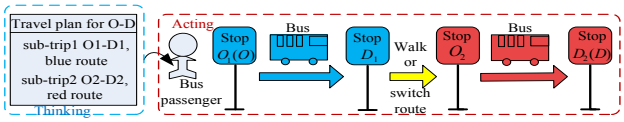


Fig.5. Passenger behavior model.

a bus passenger behavior model was designed and implemented to run jointly with SUMO via a monitor-control algorithm based on TraCI, a traffic control interface of SUMO. Specifically, the algorithm was to 1) modify the conditions and states of buses, 2) simulate passenger behaviors, and 3) record important moments (bus arrival time, etc.) in real time.

In our work, a bus passenger behavior model was developed, which is illustrated in Figure 5: according to the travel plan, the passenger starts at origin stop O_1 and takes a bus on the blue route to D_1 to complete sub-trip 1. Then, the passenger gets to O_2 via walking (if $D_1 \neq O_2$) or route switching (if $D_1 = O_2$) to start sub-trip 2. Finally, the passenger gets to the final destination D_2 through the red route, and the travel demand is fulfilled. Interactions between buses and passengers take place at each stop, where the bus dwell time is affected by the number of boarding and alighting passengers. According to STCP vehicle descriptions, most buses in Porto city have independent boarding and alighting channels for passengers, and thus, dwell time is the maximum of boarding time t_on and alighting time t_off. The interactions between buses and traffic are simulated by SUMO, where the travel time t_bus varies according to traffic conditions on the roads. The time the passenger spends from stop i to stop $(i+1)$ is

$$t_i = L + max(t\_on_i, t\_off_i) + t\_bus_i, \quad (3)$$

where L is a constant of lost time, including the pulling, door-open-close time, etc. With this model, the simulation captures primary interactions among passengers, buses, and traffic. All bus passengers in the city are treated as agents who follow both the travel planning model and the behavior model defined in previous sections.

Loading the passenger travel demands to the joint simulation layer, we simulated the city-wide bus passenger behaviors in Porto city for 90 Wednesdays. For each day, the simulation log stores detailed passenger behavior information and bus state information. For example, the passenger log records the time spent of waiting at the stop, boarding, and alighting at the destination stop. The bus log includes the bus arrival time at each stop, on/off passengers' ID at each stop, stop dwell time, and passenger volume after prompting passengers to getting on/off. The passenger log and bus log constitute the traffic-passenger joint simulation dataset of the bus transportation system in Porto. This synthetic bus passenger data have been successfully applied to validate a semi-supervised learning based method for inferring the passengers' unknown destination [12].

V. Evaluation

We evaluated our bus passenger simulation data using real bus AFC data collected from Porto city. The basic idea is to compare the simulation data with the real data in terms of spatial-temporal distribution, of which the difference is measured by means of Kullback–Leibler divergence.

A. Real AFC Bus Passenger Data

The AFC dataset is the set of bus passenger transaction records that occurred in January, April, and May of 2010. The data were collected by the AFC system installed in buses operated by STCP in Porto city. The AFC system called "Andante" is an entry-only system. Each transaction record contains several attributes of which we are interested in the following: the 1) ID; 2) transaction timestamp; 3) bus stop where the transaction occurred; 4) route; and 5) route direction. We fused the Andate AFC data with additional data sources to obtain the route structure (sequence of stops in a route) and the geographical location of each stop. There are 2,374 bus stops and 66 bidirectional bus routes within the area of interest. The raw data have about 3% fault samples that contain illogical or missing attributes. After data recovery, we selected transaction records on Wednesdays of the three months, totaling 12 Wednesdays with 2,422,079 transactions. Those Wednesdays were normal weekdays, and local special holidays were avoided.

B. Evaluation with Respect to Spatiotemporal Distribution

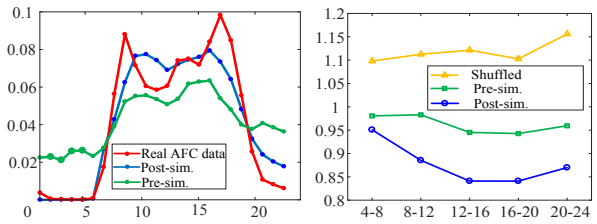
The goal of the simulation was to capture the underlying distributions from which the real observations are generated so that the simulation outcomes can be used as a reasonable approximation of real passenger data. To this end, we quantified the difference between the synthetic passenger data and the real AFC data, and we investigated the validity of our method by comparing the synthetic-real data difference with the data difference of baseline methods (not using the simulation) from the real data. The measurement applied in this paper to quantify the difference between two distributions is called Kullback–Leibler (KL) divergence [13]. For discrete distributions, the KL divergence from distribution $Q(i)$ to $P(i)$ is defined as:

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}. \quad (4)$$

The larger the $D_{KL}(P||Q)$ is, the more difference there will be between P and Q.

1) KL Divergence in Temporal Distributions: We first investigated the difference in *temporal passenger demand distributions* between the simulation data and the real data. Here, P and Q are the expected *temporal* passenger demand distributions for simulation data and real data, respectively. In this paper, we focus on Wednesday data. Specifically, the distribution P is defined as $P(i) = \mathbb{E}(n_i) / \sum_{j=1}^{24} \mathbb{E}(n_j)$, where n_i is the number of passengers who get on a bus in period i (e.g., 10 am–11am), and $\mathbb{E}(n_i)$ is the average number (over all 90 simulated Wednesdays) of passengers who get on a bus in period i. In the same way, we can obtain the distribution Q for the real data. The shapes of both distributions are illustrated in Figure 6. a), where we can see a clear similarity between them.

For comparison purposes, we also considered two baseline distributions. The first one is simply *shuffled* from P, and we call it *shuffled distribution*. The shuffle means a random permutation of $P(i)$ w.r.t. period i. The second baseline distribution is the temporal distribution estimated directly from the taxi passenger data, which is the green distribution in Figure 6. a). We call such a distribution the *pre-simulation distribution* because the data have not been



a) b)

Fig. 6. a) Demand occurrence probability vs. hour for pre-simulation data (green), post-simulation data (blue), and real AFC data (red). b) Spatial KL divergences $D_{KL\_spatial}$ vs. time periods T.

processed by the simulation. Note that only the first *trip demand starting time* t (see TABLE I) contributes to pre-simulation distribution. In contrast to the two baselines, P is called *post-simulation distribution* (the blue distribution in Figure 6.a). Note that not only first trip starting time t but also synthetic midway trip starting time $\{t_i\}$ contributes to post-simulation distribution.

Calculating the KL divergence from the real distribution Q to each of the three distributions, we obtained 1.418, 0.553, and 0.045 for shuffled, pre-, and post-simulation distributions, respectively. Compared with the shuffled and pre-simulation distributions, the post-simulation distribution achieves the best similarity, reducing the distribution difference down to 0.045. The information gain comes from the fact that the simulation can capitalize on the passenger behavior model to effectively fill the midway details (especially the *timing* $\{t_i\}$ of each midway trip) between the origin and the destination.

2) KL Divergence in Spatial Distributions: We further investigated the difference in spatial distributions. Because the real data contains only boarding information, we should conceptualize the spatial distribution accordingly: because a bus route R consists of a sequence of bus stops $\{s\}$, and because each stop corresponds to a spatial location, the spatial probability distribution of passenger boarding demands associated with route R is essentially the boarding probability distribution over bus stops $\{s\}$. Considering that the spatial distribution can vary in difference periods, we focused on the periods of $\{T\}$ ={4-8, 8-12, 12-16, 16-20, 20-24} and omitted the period of 0-4 because buses are mostly off-service during that time. Then, for simulation data, given period T and route R, the spatial distribution over $\{s\}$ is defined as $P_{R,T}(s) = \mathbb{E}(n_{R,T,s})/\sum_k \mathbb{E}(n_{R,T,k})$, where $n_{R,T,s}$ is the number of passengers (on a Wednesday) who get on a bus in period T at stop s of route R, and $\mathbb{E}(n_{R,T,s})$ is the average number (over all 90 simulated Wednesdays) of passengers who get on a bus in period T at stop s of route R. In the same way, we can obtain $Q_{R,T}(s)$ for the real data. The spatial KL divergence for the period T is defined as:

$$D_{KL\_spatial}(T) = \mathbb{E}_R[D_{KL}(P_{R,T}||Q_{R,T})] = \frac{1}{N_R}\sum_R D_{KL}(P_{R,T}||Q_{R,T}), \quad (5)$$

where N_R is the number of routes in the area being investigated. This is the expected KL divergence in spatial distributions over all bus routes during certain period T.

Based on (5), the spatial KL divergences from the real spatial distribution to the shuffled, pre-, and post-simulation spatial distributions are calculated and illustrated in Figure 6. b). Note that the spatial pre-simulation distribution counts only on the trip demand (O, D, t) of each passenger; in contrast, the spatial post-simulation distribution counts on the whole synthetic experience $\{(\boldsymbol{O}, D_1, \boldsymbol{t}), (O_2, D_2, t_2), \ldots, (O_n, \boldsymbol{D}, t_n)\}$ of each passenger. From Figure 6. b), we can observe that from shuffled to pre- and then to post-simulation distributions, there is a decreasing trend in the divergence. The experimental results support that the post-simulation data exhibit a higher degree of similarity to the real bus passenger data in terms of spatial activity. This experimental outcome also supports our claim: the joint traffic-passenger modeling and simulation is a meaningful method for transferring indirect people mobility data to direct and complete bus passenger data.

VI. Conclusion

We proposed a methodology for generating synthetic bus passenger data through joint traffic-passenger modeling and simulation on a city scale. It is the first use of a modeling and simulation approach to transfer the indirectly related people mobility data to direct and complete passenger data. This method is validated by quantifying the similarity of the distributions between the synthetic passenger data and real data. Our main contribution is a proof-of-concept of how academia can move forward in the absence of direct and complete data in the field of passenger-related research by using the indirect people mobility information. The proposed methodology is expected to serve as a potential driving force of intelligent transportation system success.

References

[1] L. Moreira-Matias, and O. Cats, "Toward a Demand Estimation Model Based on Automated Vehicle Location," *Transp. Research Record: J. of Transp. Research Board*, 2016(2544), 141-149, 2016.

[2] A. A. Nunes, T. G. Dias, and J. F. Cunha, "Passenger Journey Destination Estimation from Automated Fare Collection System Data Using Spatial Validation," *IEEE Trans. on Intelligent Transp. Syst.*, 17(1), 133-142, 2016.

[3] T. Schelenz, A. Suescun, M. Karlsson, and L. Wikstrom, "Decision making algorithm for bus passenger simulation during the vehicle design process," *Transport Policy*, 25, 178-185, 2013.

[4] A. Sumalee, Z. Tan, and W. H. Lam, "Dynamic stochastic transit assignment with explicit seat allocation model," *Transp. Research Part B: Methodological*, 43(8), 895-912, 2009.

[5] J. D. Schmocker, A. Fonzone, H. Shimamoto, F. Kurauchi, and M. G. Bell, "Frequency-based transit assignment considering seat capacities," *Transp. Research Part B: Methodological*, 45(2), 392-408, 2011.

[6] Q. Zhang, B. Han, and D. Li, "Modeling and simulation of passenger alighting and boarding movement in Beijing metro stations," *Transp. Research Part C: Emerging Technologies*, 16(5), 635-649, 2008.

[7] Trajectory - Prediction Challenge Dataset, ECML/PKDD 2015: http://www.geolink.pt/ecmlpkdd2015-challenge/dataset.html

[8] "STCP Annual Report and Accounts", STCP, Porto, Portugal, 2016.

[9] B. W. Silverman, "Density estimation for statistics and data analysis," Vol. 26, CRC press, 1986.

[10] "Transport Assessment Best Practice: Guidance Document," Transport for London, London, U.K., April 2010.

[11] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, 1(1), 269-271, 1959

[12] R. Shi, P. Steenkiste, and M. Veloso, "Second-order destination inference using semi-supervised self-training for entry-only passenger data," In *Proc. of the 4th IEEE/ACM International Conference on Big Data Computing, Applications and Technologies* (BDCAT), 2017

[13] T. M. Cover, and J. A. Thomas, "Elements of information theory," Wiley press, 1991