# 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*—Improved planning, management, and the control of future transportation systems requires passenger related research, such as the analysis of passenger behaviors in a city. However, the passenger data available to researchers and city planners are usually insufficient for this research, either because the data is incomplete with important features missing, or because the data is indirectly related to the task we attempt to achieve. This limits the development of passenger related research. The goal of our work is to generate synthetic passenger data using a novel methodology that leverages joint traffic-passenger modeling and simulation at city scale. A demonstration of such idea in generating synthetic bus passenger data is implemented. Specifically, the method involves 1) a bus passenger demand model, learned from indirect people mobility data, to generate bus passenger demand samples, and 2) a bus passenger behavior model, which jointly runs with a traffic simulator (SUMO), to generate synthetic bus passenger data. We implement the methodology for the 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 the same city. The proposed method may serve as a potential driving force of intelligent transportation system success.**

***Keywords*—transportation, simulation, synthetic data, behavioral modeling, Poisson process, kernel density estimation**

## I. Introduction

To efficiently move city dwellers, given the surge in city population and their mobility needs, is challenging. To address this challenge, *intelligent transportation system* is proposed to use urban informatics and technology to improve transportation efficiency. Public transportation systems, especially bus transportation systems, play an important role in moving passengers in a fast and convenient way. To build an efficient public transportation system, a profound understanding of city-wide passenger behavior is necessary. It has long been thought that the 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. On one hand, with the advance of sensing technologies and the widespread use of automated data collection (ADC) in public transportation, it is possible to collect large quantities of diverse data about urban spaces and city population, e.g., cellphone location data, vehicle GPS data, automated fare collection (AFC) data. With heterogeneous and ubiquitous datasets, researchers have significantly expanded their knowledge about human mobility [1], and plenty of machine learning approaches have been successfully applied to the transportation field (e.g. bus arrival time prediction [2] and vehicle trajectory prediction [3], [4]).

On the other hand, however, when it comes to passenger related research, the data available to researchers are usually insufficient for this research, either because the data is *incomplete* with important features missing, or because the data is *indirectly related* to the topic of focus. Details and examples about terminologies are presented in Table I. Complete data is usually lacking, due to the challenges for urban infrastructures to collect and integrate large-scale multi-source data in a timely and low-cost fashion, and the concern of privacy infringement. This limits passenger related research. For example, to understand passenger demand is key for effective planning of public transportation service. In bus transportation system, the bus passenger data available nowadays are generally provided by automated passenger counts (APC) and AFC system. 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 infrequent, expensive in terms of human effort and financial cost, and prone to response bias. As a result, the development of many state-of-the-art methods for bus passenger estimation and prediction (e.g. [5], [6], [7]) commonly faced an issue while evaluating their validity, due of the lack or missing of necessary features.

To cope with the issue, indirectly related data could be a way out. In general, indirectly related data is from a different source, with some features correlate positively with those in the unknown complete data. Then here comes the opportunity – grounded on those correlated features, we attempt to develop a method to generate synthetic complete data that is most likely to be observed in reality. Inspired by this idea and to move forward in the absence of complete data, as a main contribution of this paper, we propose a methodology to generate synthetic passenger data through joint traffic-passenger modeling and simulation based on indirectly related people mobility data. To be specific, we demonstrate and verified the proposed method in the setting of bus transportation systems:

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 (feature space of the data) and the distribution of the data on those features (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 is in the domain 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 is in the domain which 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 to fully train a cat/non-cat classifier. |
| Indirectly Related Data | The data which is indirect data, and whose domain overlaps with or is similar to the domain of complete data, w.r.t. that some of the features are the same and the distributions on those features are similar. | The dog image set is indirectly related to the cat: its distribution on the features space is more similar to the cat distribution than other images like vehicle, house, etc. Thus the dog image data can help partially solve the cat classifier 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 {(**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**, {(**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 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 transit to get to stop D. Then, experience can be a set of tuples {(**O**, $D_1$, **t**), ($O_2$, $D_2$, $t_2$),…,($O_n$, **D**, $t_n$)}. Here, $t_n$ is the time of arriving at $O_n$. |

- We learn a bus passenger demand model, from indirectly related people mobility data (taxi data), to generate bus passenger demand samples. This is motivated by the insight that people mobility trend reflected by the data of different sources can correlate positively, to some extent, to the mobility trend of real bus passengers.
- We develop a passenger behavior model to jointly run with a mature traffic simulator (SUMO) for city-wide bus passenger synthetic data generation.
- We implement the methodology using the case study of Porto city, Portugal. The simulation outcomes are validated by measuring the distribution difference between the synthetic passenger data (also called as *simulation data*) and the real bus Automated Fare Collection (AFC) data of the same city.
- Our work is the first successful attempt to transfer indirect people mobility data to complete bus passenger data through joint traffic-passenger modeling and simulation.

## II. Background and Related Work

In this section, we discuss the importance of simulation from the point of view of transportation systems, and review related work on traffic/passenger modeling and simulation.

### A. Simulation and Intelligent Transportation Systems

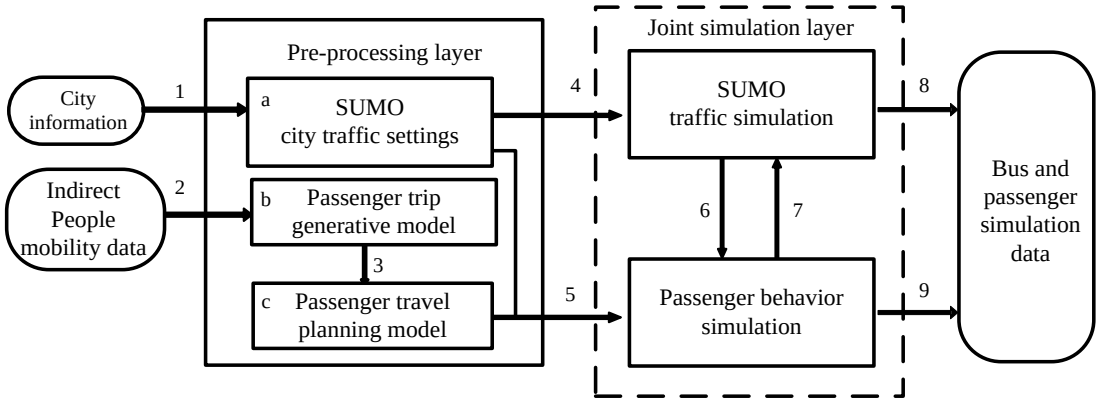
Simulation has proven itself as a prerequisite to building intelligent transportation systems. In traffic control research, instead of directly applying an approach to control the real traffic flow, testing it through simulation and transferring the knowledge to real urban road networks can save a lot of cost and avoid safety problems. For many intelligent transportation system applications that involve interactions between complex spatial traffic networks, vehicles, and humans, to build an effective model requires large-scale real-world data which can be difficult, time-consuming, and even risky to collect. In many cases, simulations are applied as a workaround approach to generate synthetic data for city planners and transportation management practitioners to learn from [8]. In this sense, simulation is a key driving force in the field of intelligent transportation systems.

### B. Traffic, Bus and Passenger Simulators

Passenger behavior modeling and simulations have been involved in a lot of transportation system research. To evaluate the performance of vehicle scheduling and platform deploying (such as selecting bus stop sites), the behaviors of passengers must be simulated and analyzed in detail. Though sophisticated enough to take into consideration individual preference [9], seat allocation process [10], [11], and even pressure from passengers behind [12], most studies are highly microscopic, confining their domains in limited amounts of buses, and not scaling well to provide insights into macroscopic passenger flow in the whole city.

Compared to passenger modeling, traffic simulation (including bus simulation) has gained rapid and significant developments. A lot of road traffic simulators (e.g. VISSIM, AIMSUN, Matsim, SUMO, etc.) are developed with delicate functions and performances. One commonly used open source traffic simulator is SUMO (Simulation of Urban MObility), which provides a platform to explicitly simulate vehicles including cars, buses, and urban trains at 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 to simulate bus passenger behavior in conjunction with the mature traffic simulator (SUMO) at 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.




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 in the setting of bus transportation systems.

## III. Methodology

The problem of focus in this paper is that: *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 the passenger behavior nor simulating the traffic can solve the abovementioned problem independently, and this is why the idea of combining the passenger modeling and the traffic simulation comes in. First of all, modeling the 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 provide what passengers actually experience during the travel in the urban traffic environment (e.g., the $t_i$ in TABLE I). This missing experience can be supplemented by the traffic simulation. Second, most traffic simulation provides representation of transportation systems and vehicle behaviors, especially how the public transits operate in the 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 by passenger modeling.

To combine the passenger modeling and the traffic simulation is an effective way to make the best of both approaches and overcome the shortage of either. The method starts from generating the O-D demand of the whole trip, which is based on the knowledge that can be learned efficiently from the people mobility data, and then leverages the joint modeling and simulation to *fill in* the midway travel details in terms of what plan and route to choose, when and where passengers get on and get off, what experience they actually encounter during the trip, and etc. By designing the passenger models, configuring the urban traffic simulation, and combining both appropriately, it is reasonable to expect that the passengers' synthetic experience can be used as the synthetic data that are likely to be observed in real world and upon which some passenger related research can be conducted. On the other hand, the availability of mature traffic simulators and the common sense of passengers' general behaviors in public transits significantly slack the difficulty of implementing the proposed method in practice, making it a potential generic approach to overcome the lack-of-complete-data challenge.

### B. Overview of the Method

We provide a high-level overview of the proposed methodology in the setting of bus transportations systems. 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 between passengers, buses, and traffic. Specifically, it simulates the behavior of bus passengers moving through the urban bus network while having the buses to interact with the urban traffic environment. To avoid misunderstanding, we define that a passenger "trip demand" consists of *trip starting time* and an *O-D pair* (origin-destination pair); while a passenger "travel demand" consists of *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 algorithms (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 next layer. Specifically, *Algorithm a* is 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. *Algorithm b* is to learn a passenger trip generative model from people mobility data and generate passenger trip demand samples. This model should be designed and established according to the type of people mobility data. *Algorithm c* is to model the way a bus passenger *thinks of* a travel plan from the very origin to the final destination through the bus network, and finally this algorithm generates

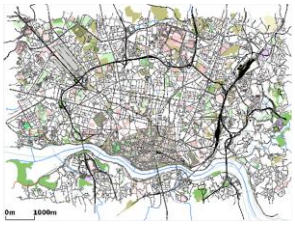


Fig. 2.Virtual Porto traffic network in SUMO.

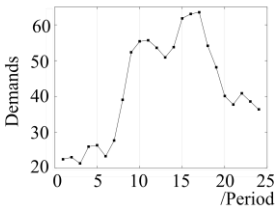



Fig. 3. Expected taxi demand on Wednesday

a passenger travel demand that includes 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 bus states in real time and simulate passenger behaviors accordingly. At the end of the simulation, this layer outputs detailed passenger traveling information and bus state information.

With the rich and informative synthetic passenger data from the simulation, researchers and city planners can make use of them according to needs.

## IV. Implementation

We applied the methodology for a case study of Porto city, the second largest city in 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. The main bus service operator STCP offers a company website, where detailed routes, station geographical location, and timetable information are provided. As shown in Figure 2, using the STCP bus service information and other public resource (OpenStreetMap, etc.), we established the bus transportation system as well as the urban road network within the selected central city area of Porto (E: -8.559543°, W: -8.661915°, S: 41.136044°, N: 41.185110°). The imported bus network contains 136 routes, 855 bus stops, and 5723 bus trips in a normal workday. It is confirmed from the simulation tests that the bus performance matches well to the real Porto bus transportation system: each bus departs at scheduled time, runs along its designated route, and pulls at designated stops correctly.

After establishing the virtual traffic network in SUMO, we turned to learning the passenger trip demand generative model for generating passenger trip samples.

### *B. Bus Passenger Trip Demand Generative Model*

The goal of this model is to generate the trip demand tuple $(O, D, t)$ of a passenger. The approaches used to establish the passenger trip demand generative model highly depend on the sources of data available. In many cases, direct and complete data in 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 distributions, based on which the bus passenger trip demand model is established to generate passenger trip samples.

The taxi dataset [13] used in the paper describes a complete year (from 1/7/2013 to 30/6/2014) of the trajectories for all 422 taxis running in Porto city. All the taxis are equipped with mobile data terminals, providing information on GPS localization and taximeter state. Each ride is categorized into three categories: A) Taxi central based – if the taxi received a telephone call for demand, and was dispatched from the operation central; B) Stand based – if the taxi picked up the passenger at one of the 63 taxi stands in the city; C) random street based – if the trip was demanded on a random street. Each data point contains several features, of which we are interested in: 1) Trip starting time; 2) Date type (identifying whether the trip occurred in a holiday or any other special day); 3) Call type (telling whether the trip started from the operation central, one of the taxi stands, or a random street); and 4) Poly line (storing the list of GPS coordinates for the trip trajectory). In data pre-processing, according to the need of our study, we selected out the data in category C and remove special-day samples. The dataset contains detailed trip starting time and O-D pairs of random street passengers, making it a nice resource of city dwellers' travel trend. On the other hand, the area being modeled and simulated is the central city area of Porto, which has a quite dense bus network, and this setting mitigates the negative correlation between taxi and bus demand models by reducing area that is 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 event in time. We use this model to describe how frequent passenger demands will occur across the city in a day. We should be careful about the fact that the Poisson process could be inhomogeneous and the *rate parameter* $\lambda$ may vary in time. However, instead of complete randomness, human trajectories show a high degree of temporal and spatial regularity [1]. In that sense, we fit the rate parameter on hourly basis for a certain weekday. We divided a day equally into 24 periods, and focused on studying all Wednesdays of the year. The averaged taxi demand in each period on Wednesday is shown in Figure 3, according to which we fit the estimated Wednesday rate vector $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_{24})$. The temporal model is then described as: In period $i$, the interval $\tau$ between two consecutive passenger demands follows an 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 $\hat{\lambda}$, since the number of bus demands is usually greater than taxi demands ($a$ is used to scale daily passenger demands up to 150k, which is suggested by the STCP 2016 annual service report [14]). 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 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 effective when prior knowledge about the distribution is unavailable, thus parametric methods do not apply well. The bandwidth is determined based on the normal distribution approximation [15]. Finally, we have the spatial model as:

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

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 use it as the origin/destination stop. We set a cut-off distance of 640 meters with the stop matching outside this region nulled. This distance is from the Public Transport Accessibility Levels (PTAL) methodology, which proposes an 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 [16].

A temporal sample and a spatial sample constitute a passenger trip demand sample. This is the core design of the passenger trip demand generative model.

### C. *Passenger Travel Planning Model*

The passenger travel planning model is to provide a solution to the question: Given an origin bus stop and a destination bus stop (the O-D pair of a trip demand), how does a passenger *think of* a travel plan in the given bus transportation network? Specifically, the model proposes the set $\{(O, D_1), (O_2, D_2), \ldots, (O_n, D)\}$ of midway trips which lead the passenger to destination. Passengers always choose a plan that minimizes cost, distance, and unnecessary route switching. Usually, many web mapping projects (e.g. Google Map) have a route planning service. However, there are many cases where the city bus transport information is monopolized by local institutes. As a result, a universal planning service does not apply to a lot cities in different countries. As to the route planning service developed by local institutes, it could be less developed due to lack of investment. To generalize our methodology to different cities, 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 bus route segments that connect the stops. For example, the blue edge from $v_1$ to $v_2$ indicates that bus route A will pass from stop $v_1$ to stop $v_2$. In the graph, route A (with blue edges) and route B (with red edges) share the hub 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 as walking reachable stops, and passengers would be willing to walk a few more meters to transit at those stops. In each bus route, vertices are designed to be fully connected, and later we will see that this structure will

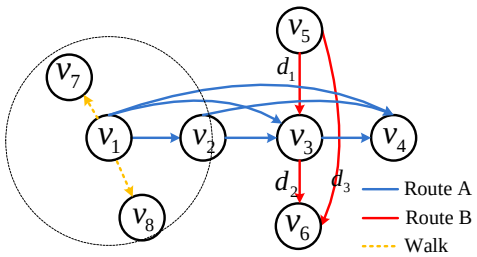



Fig. 4. Graph of bus transportation network.

TABLE II
VERTEX AND EDGE ATTRIBUTES

| | Attribute | Description |
|---|---|---|
| Vertex | ID | Vertex identity |
| | Geo-location | Longitude and altitude of the corresponding bus stop on map |
| | Out-edges | List of edge IDs originated from the vertex |
| | Out-neighbors | List of vertex IDs that terminate out-edges |
| | Route | List of routes passing through the vertex |
| Edge | ID | Edge identity |
| | Source-vertex | The vertex ID the edge originates from |
| | End-vertex | The vertex ID the edge terminates at |
| | Weight | Length of the edge (meter) |
| | Route | The route the edge belongs to |

**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 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/vertices
        for e in E_out do
            update_cost← cost[u]+e.weight+ ε %basic cost
            if u.preRoute ≠ e.route then     %route switched or not
                update_cost← update_cost+Δ
            if cost[v_temp]>update_cost then
                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
```

avoid redundant output for the travel plan. After establishing the bus transportation network in SUMO, we can measure the exact length of each edge (the weight of each edge denoted as $d$). Table II lists the attributes of vertex and edge. Note that the *out-edges* of a vertex indicates the outward edges that originate from the vertex, and the *out-neighbors* indicates the terminal vertices of corresponding outward edges. The *vertex route* is a set of routes the vertex involved in, and the *edge route* is the specific route the edge belongs to. All the vertices and edges uniquely define the whole bus transportation network (graph).

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 want to find an optimal plan that minimizes a certain cost object, and this is a combinatorial optimization problem. In addition to the cost object in traditional shortest path

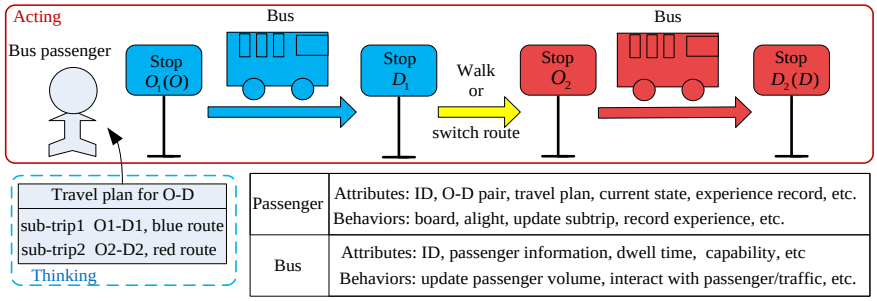



Fig.5. Passenger behavior model.

searching problems, which based on accumulative distance only, we introduced 1) a route-switching penalty $\Delta$ to penalize the route switching of a travel plan, and 2) a penalty $\varepsilon$ to penalize the object when a sub-trip is added into the travel plan. Introducing $\varepsilon$ 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 prefer the optimal travel plan to be represented as {edge( $v_5, v_6$ )} rather than {edge( $v_5, v_3$ ),edge( $v_3, v_6$ )}; By introducing $\varepsilon$, the plan {edge( $v_5, v_6$ )} with cost ( $d_3 + \varepsilon$ ) will win over the plan{edge($v_5, v_3$), 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 [17] 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.

We applied the passenger travel planning model to generate a travel plan for each passenger trip demand. The optimal plan searching algorithm turns out to be very effective in practice. Applying the trip demand generative model and the travel planning model, we can generate passenger travel demand samples on Wednesdays at any size according to need.

### *D. Bus Passenger Behavior Modeling + SUMO Simulation*

Given a bus passenger travel demand (containing trip starting time, an O-D pair, and a travel plan), we need 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, a bus passenger behavior model is 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 is 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. When conducting simulations, we always make a trade-off between simulation speed and granularity depending on research topics. In most cases, to gain macroscopic insights into city-scale passenger behavior patterns, a reasonable strategy is to begin with basic models, and gradually increase the subtlety according to computing capacity and needs.

In our work, a bus passenger behavior model is developed which is illustrated in Figure 5: According to the travel plan, the passenger starts at the origin stop $O_1$ and takes a bus of the blue route to $D_1$ to complete sub-trip 1. Then, it gets to $O_2$ by 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. The 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 channels for boarding and alighting, respectively, and thus, the dwell time $t_{dwell}$ is formulated as $t_{dwell} = max(t\_on, t\_off)$. 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 the travel $L$ is a constant of lost time including pulling, door-open-close time, etc. With this model, the simulation captures primary interactions between passengers, buses and traffic. All bus passengers in the city are treated as agents that 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 city-wide bus passenger behaviors in Porto city of 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 of waiting at the stop, boarding, and alighting at the destination stop. The bus log includes bus arrival time at each stop, on/off passengers ID at each stop, stop dwell time, and passenger volume after accommodating passengers to get on/off. The passenger log and bus log constitute the traffic-passenger joint simulation dataset of the bus transportation system in Porto.

## V. EVALUATION

We evaluated our bus passenger simulation data using real bus Automated Fare Collection (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. The difference in distribution 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 the January, April and May of 2010. They are collected by the AFC system installed in buses

operated by STCP in Porto city. The AFC system called Andante is an entry-only system - a transaction record will be generated once a passenger taps the travel card on the AFC reader. But no additional action is needed when the passenger alights, so the information of destinations is unavailable. Each transaction record contains several attributes among which we are interested in: 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 in the area of interest. The raw data have about 3% fault samples that have illogical or missing attributes. After data recovery, a proportion of 1% remains unsolved, and we removed those records. We selected out transaction records on Wednesdays of the three months, totaling 12 Wednesdays with 2,422,079 transactions. Those Wednesdays are normal weekdays, and local special holidays are avoided.

### *B. Evaluation with Respect to Spatiotemporal Distribution*

The goal of the simulation is 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 (in this paper, we use AFC data). To check the fulfillment of this goal, we quantified the difference between our synthetic passenger data and the real AFC data, and investigated the improvement in distribution similarity achieved by our method by comparing to baseline methods that do not use simulations.

The measure applied in this paper to quantify the difference between two distributions is called Kullback–Leibler (KL) divergence [18]. 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 use of KL divergence in measuring the difference between probability distributions is popular, because of the following reasons: 1) $D_{KL}(P||Q)$ is always non-negative; 2) $D_{KL}(P||Q) = 0$ holds if and only if $P = Q$; and 3) $D_{KL}(P||Q)$ is a convex function with respect to $P$ and $Q$. The convexity implies the fact that 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 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 that get on a bus in period $i$ (e.g., 10 am–11am) on a Wednesday, and $\mathbb{E}(n_i)$ is the average number (over Wednesdays) of passengers that 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, where we can see a nice similarity between them (in this case, the blue one is $P$ and the red one is $Q$).

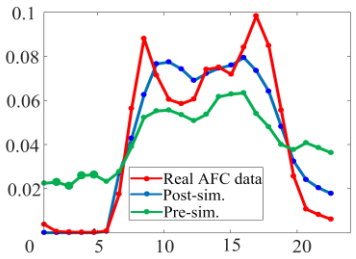



Fig. 6. Temporal distributions of bus passenger demands for pre-simulation data (green), post-simulation data (blue), and real AFC data (red).

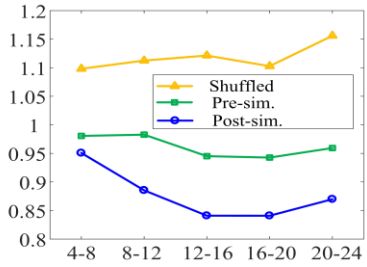



Fig. 7. Spatial KL divergences vs different designated periods.

For comparison purpose, 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. the period $i$. The second one is the temporal distribution estimated directly from the taxi passenger data which is the green distribution in Figure 6. We call such distribution the *pre-simulation distribution*, because the dataset the distribution is based on has not been processed by the simulation. Note, only the first trip demand starting time $t$ of each passenger (see TABLE I) contributes to the pre-simulation distribution. In contrast to the two baseline distributions, $P$ is called *post-simulation distribution* (the blue distribution in Figure 6). Note, not only the first trip starting time $t$, but also the synthetic midway trip starting time $\{ t_i \}$ contributes to the 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 to the shuffled distribution, pre-simulation distribution has an improvement in the similarity to the real data. This supports our assumption that the temporal mobility trend of passengers reflected by the taxi data correlates *positively* to the real temporal mobility trend of bus passengers. On the other hand, the post-simulation distribution achieves more information gain, 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 boarding) between the origin and destination. This contributes to a smaller KL divergence.

***2) KL Divergence in Spatial Distributions:*** We further investigated the difference in spatial distributions of bus passenger demands between simulation data and the real

data. Since the real data only contains boarding information, we should conceptualize the spatial distribution accordingly: because a bus route $R$ consists of a sequence of bus stops $\{s\}$, and each stop corresponds to a spatial location, the spatial probability distribution of passenger boarding demands associated with the route $R$ is essentially the boarding probability distribution over bus stops $\{s\}$. Considering that in different periods, the spatial distribution can vary, we focused on the periods of $\{T\}$ ={4-8, 8-12, 12-16, 16-20, 20-24} and omitted the period 0-4 because buses are mostly off-service during that time. Since the constraint of period $T$ is involved, the spatial distribution is actually evaluated from a spatial-temporal point of view.

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) that 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 Wednesdays) of passengers that 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 followed:

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

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 three pre-mentioned spatial distributions (shuffled, pre-, and post-simulation spatial distributions) are calculated and illustrated in Figure 7. Note, the spatial pre-simulation distribution counts only on the trip demand $(O, D, t)$ of each passenger, and 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 7, we can observe that: from shuffled, to pre-, and then to post-simulation distributions, there shows a decreasing trend in the divergence. The experimental results support that the post-simulation data exhibits 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 to transfer indirect people mobility data to direct and complete bus passenger data.

## VI. Conclusion

In this paper, we proposed a methodology to generate synthetic bus passenger data through joint traffic-passenger modeling and simulation at 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 distribution between the synthetic passenger data and real passenger data. As an echo of our main contribution, our work is a proof-of-concept of how academia and a city planning community can move forward in the absence of direct and complete data in the field of passenger related research by using the joint traffic-passenger modeling and simulation method at city scale. The proposed methodology is expected to serve as a potential driving force of intelligent transportation system success.

## References


[1] M. C. Gonzalez, *et al.*, "Understanding individual human mobility patterns," *Nature*, 453(7196), 779-782, 2008.

[2] M. Chen, X. Liu, J. Xia, and S. I. Chien, "A Dynamic Bus-Arrival Time Prediction Model Based on APC Data," *Computer-Aided Civil and Infrastructure Engineering*, 19(5), 364-376, 2004.

[3] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi–passenger demand using streaming data," *IEEE Trans. on Intelligent Transp. Syst.*, 14(3), 1393-1402, 2013.

[4] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Time-evolving OD matrix estimation using high-speed GPS data streams," *Expert Systems With Applications*, 44, 275-288, 2016.

[5] 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.

[6] 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.

[7] Y. Wang, S. Ram, F. Currim, E. Dantas, and L. Alberto Sabóia, "A Big Data Approach for Smart Transportation Management on Bus Network," in *Proc. of IEEE Intl. Smart Cities Conf.*(ISC2), 2016.

[8] 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), pp. 255-264, 2017

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

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

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

[17] M. Van, "Dijkstra's algorithm," Massachusetts Institute of Technology. Regexstr, 2014

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