Exploiting Heterogeneous Human Mobility Patterns for Intelligent Bus Routing

Yanchi Liu*, Chuanren Liu†, Nicholas Jing Yuan‡, Lian Duan*,
Yanjie Fu†, Hui Xiong†, Songhua Xu† and Junjie Wu§

*New Jersey Institute of Technology †Rutgers University ‡Microsoft Research §Beihang University
Email: *{yl473, lian.duan, songhua.xu}@njit.edu
†{chuanren.liu, yanjie_fu, hxiong}@rutgers.edu ‡nicholas.yuan@microsoft.com §wujj@buaa.edu.cn

Abstract—Optimal planning for public transportation is one of the keys to sustainable development and better quality of life in urban areas. Compared to private transportation, public transportation uses road space more efficiently and produces fewer accidents and emissions. In this paper, we focus on the identification and optimization of flawed bus routes to improve utilization efficiency of public transportation services, according to people’s real demand for public transportation. To this end, we first provide an integrated mobility pattern analysis between the location traces of taxicabs and the mobility records in bus transactions. Based on mobility patterns, we propose a localized transportation mode choice model, with which we can dynamically predict the bus travel demand for different bus routing. This model is then used for bus routing optimization which aims to convert as many people from private transportation to public transportation as possible given budget constraints on the bus route modification. We also leverage the model to identify region pairs with flawed bus routes, which are effectively optimized using our approach. To validate the effectiveness of the proposed methods, extensive studies are performed on real world data collected in Beijing which contains 19 million taxi trips and 10 million bus trips.

I. INTRODUCTION

More and more people live in metropolitan areas or big cities due in large part to rapid development of urbanization. One major side effect is increasing traffic congestion due to limited space, and consequently unnecessary energy consumption during traffic congestion. Public transportation (e.g., bus, subway) not only saves fuel and reduces congestion, but also offers a safe, affordable, and convenient way to travel. According to American Public Transportation Association, Americans living in areas served by public transportation save 865 million hours of travel time and 450 million gallons of fuel annually in congestion reduction alone. Households that are likely to use public transportation on a given day save more than $9,700 every year. Better public transportation planning can significantly help to foster a more sustainable development and improve quality of life.

Traditional public transportation planning methods have relied on human surveys to understand people’s mobility patterns and their choice among different transportation modes. Despite the substantial time and cost spent on the survey process, these methods do not provide timely and detailed analysis on different Origin-Destination (OD) pairs given the complex and dynamic settings of cities. As a result, such methods cannot accommodate real travel demands which vary over different OD pairs and often make planning inefficient in real life. If transit agencies could have an effective tool to quantify travel demand and a choice model on how people choose public transportation and private transportation (e.g., private car, taxi), then recommendations on how to better design and optimize a given public transportation network could be proposed to attract more people to public transportation. As a result, cities would be able to better support people’s travel demand through a regulated, efficient, more sustainable public transportation system.

Meanwhile, with the wide deployment of Automatic Fare Collection (AFC) systems on bus networks and Global Positioning System (GPS) devices on taxis, large amounts of bus transactions and taxi traces are collected, with which we can provide integrated analysis on the mobility patterns of both public and private transportation. As demonstrated in this paper, this offers the possibility of optimizing public transportation by taking overall city traffic into account. By leveraging mobility patterns of public and private transportation, public transportation services can be designed in a way that accommodates different levels of demands and by doing so, attracts more potential riders and increases utilization efficiency of the public transportation system. As shown intuitively in Figure 1, people may be more willing to choose bus over taxi when better bus routing is provided.

With integrated analysis on the mobility patterns, shows our workflow to detect and re-plan flawed and less effective bus routes for attracting most number of potential bus riders. There are two main challenges to achieve this goal: 1) modeling people’s transportation mode choices for different OD pairs; 2) optimizing bus routes with budget constraints to maximize bus ridership. In order to solve these two challenges, we first partition a city into disjoint regions using bus stops. We then project taxi traces and bus transactions into these
regions and formulate transitions between each pair of regions. With these bus and taxi mobility patterns, we model the transportation mode choice for each OD pair. After that, we detect and optimize the flawed bus routes with budget constraints to increase bus ridership. The three main contributions of this paper are as follows:

- Transportation mode choice modeling: We model transportation mode decisions of OD pairs using a spatio-functionally weighted method, providing the probabilities of taking bus and taxi for each OD pair. Note that we investigate mode choice as an aggregate problem, which means we focus on people’s group behavior of OD pairs other than individual behavior.
- Bus routing optimization: Given limited budgets for bus network restructuring, we propose a method to attract the maximum number of potential bus riders from private transportation.
- Real evaluation: We evaluate our method using a series of large-scale real GPS traces generated by 30,000 taxis and over 10 million bus transactions in Beijing from August to October in 2012. We obtain strong data from the Beijing Bus Company, justifying the effectiveness of our method.

We begin by introducing the preliminaries of this study in Section II. Then the transportation mode choice model is proposed in Section III, followed by the flawed OD pairs detection and bus routing optimization in Section IV. Experimental results and related work are presented in Section V and Section VI respectively, and we summarize the results and give concluding remarks in Section VII.

![Framework of our method](image)

Fig. 2. Framework of our method

II. PRELIMINARIES

We begin by introducing the routing network which provides the platform for bus routing optimization. The routing network is generated from bus stops on the road map of the city. We then generate the human mobility patterns between regions (nodes of the routing network) using taxi traces and bus transactions. These mobility patterns will be modeled in Section III to identify the factors affecting people’s transportation choices.

Unless otherwise stated, we use bold characters to represent non-scalar variables, e.g., vectors, sequences, sets, and graphs. We use a comma in brackets to concatenate row vectors or stack column vectors horizontally, and a semicolon in brackets to concatenate column vectors or stack row vectors vertically. We use \( \langle \cdot, \cdot \rangle \) to represent the inner product of two vectors.

A. Routing network

As buses can only stop at bus stops and taxis can stop anywhere, we construct a common routing network for both buses and taxis. This network naturally partitions the urban area into disjoint regions served by buses and taxis. Through disjoint regions, we can modify bus routes to attract the corresponding taxi passengers. To this end, we partition the urban area using bus stops \( S = \{s_i | i = 1, \ldots, N\} \). Considering redundancy removing, we have merged stops with same names, or stops with different names but actually share the same place. For instance, for each of the bridges (also called overpasses) there are usually two or four stops around it, e.g., Mingguang Bridge North and Mingguang Bridge South at Xueyuan Road. Buses traveling north through Mingguang Bridge will stop at Mingguang Bridge North, but not Mingguang Bridge South. So we can merge these two stops into one, which represents Mingguang Bridge. In the rest of this paper, we assume the stops in set \( S \) have already been merged.

We then partition the map using a Voronoi-based partitioning method \( [4] \), in which the city is partitioned into regions based on proximity to bus stops. As a result, there is one region formed for each bus stop, and pick-up/drop-off points for taxi trips are mapped to the nearest bus stop. This partition method effectively describes the travel demand around bus stops \( [5] \). In the following sections, we use \( S \) to represent both stops and their respectively associated regions.

Now we define the routing network \( G = (S, E) \) with the merged bus stops \( S \) as nodes. The edges in \( E \) are direct connections of bus stops (i.e., without transiting stops in between), which are generated from existing road segments. Specifically, we have edge \( e = (s_i, s_j) \in E \), if there is a direct road connection between the head stop \( s_i \) and tail stop \( s_j \), where \( s_i, s_j \in S \). Please refer to Figure 3 (a) for an example of region nodes in Beijing, with all the existing bus routes plotted in Figure 3 (b).

![Map of Beijing](image)

(a) Map segmentation of Beijing (b) Bus routes of Beijing

Fig. 3. Map of Beijing.

B. Human mobility pattern

The human mobility patterns contain travel information for both bus and taxi riders, representing public and private transportation respectively. As shown in Figure 4, there is a clear difference between the mobility patterns of these two transportation modes. We retrieve these information by constructing transition records from the taxi traces and bus transactions, and then we summarize these information with a comprehensive set of statistics. Also, we have observed that, people’s behaviors and thus their mobility patterns vary...
significantly over different days and different time periods of a day. Therefore, we apply temporal partition on the transition records before summarizing the statistics. We give details of these three steps as follows.

1) Transition construction: We construct the transition records with the following definition:

Definition 1. A transition \( tr \) contains the following attributes: origin \( o \), destination \( d \), transportation mode \( m \) (0 and 1 stand for taxi and bus respectively), leaving time \( lt \), arriving time \( at \), travel distance \( d \), travel fare \( f \) and number of stops \( sn \). The set of all the transitions is noted as \( TR \).

Specifically, we project each bus and taxi trip to the nodes of the routing network \( G \), turning a trip into a transition. The travel distance of a taxi trip is calculated using the sum of the road distance of all consecutive bus stops traveled through.

2) Temporal partition: People go to different places on weekends (including public holidays in China) in comparison with weekdays. Also, people’s preferences among different transportation modes vary over different time periods of the day. For example, people prefer public transportation to commute, which usually happens during the morning and evening rush hours. In contrast, people often prefer private transportation for business transit during the day.

To incorporate these facts, we segment the transitions \( TR \) based on the leaving time \( lt \) to the temporal slots in Table I, which was proposed by [5] according to the traffic of Beijing. Since few people take buses between 11pm and 5am regularly, this paper focuses on the day bus lines, running from 5am to 11pm. We use \( c = 1, \ldots, 7 \) to represent the temporal slots, and each is associated with its time proportion in \( STime^c \), for example \( STime^1 = 5 \times 5.5 \) hours (5.5 hours every day and 5 days every week).

### Table I. Temporal Slots for Weekday and Weekend.

<table>
<thead>
<tr>
<th>Slot</th>
<th>Weekday</th>
<th>Weekend</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5:00am-10:30am</td>
<td>5:00am-12:30pm</td>
</tr>
<tr>
<td>2</td>
<td>10:30am-4:30pm</td>
<td>12:30pm-7:30pm</td>
</tr>
<tr>
<td>3</td>
<td>4:00pm-7:30pm</td>
<td>7:30pm-11:00pm</td>
</tr>
</tbody>
</table>

3) Statistical summarization: Now we summarize the partitioned transitions \( TR_{ij}^c = \{ tr : tr.o = i, tr.d = j, tr.lt \in c \} \) with statistics defined in Table II for each OD pair \((i,j)\), temporal slot \( c \), and transportation mode \( bus/taxi \), respectively. The definition of an OD pair is given as follows.

Definition 2. An OD (Origin-Destination) pair \((o, d)\) is a pair of regions with origin \( o = s_i \), destination \( d = s_j \), where \( s_i, s_j \in S \). We also write it as \((i,j)\) for short.

### Table II. Statistics of transition records.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume</td>
<td>( BVol_{ij} = \sum_{tr \in TR_{ij}} (tr.t \in c) )</td>
</tr>
<tr>
<td>Travel Time</td>
<td>( BTime_{ij} = \sum_{tr \in TR_{ij}} \frac{\sum_{tr.t \in c} (tr.at-tr.lt)}{</td>
</tr>
<tr>
<td>Travel Distance</td>
<td>( BDist_{ij} = \sum_{tr \in TR_{ij}} \frac{\sum_{tr.t \in c} (tr.d)}{</td>
</tr>
<tr>
<td>Velocity</td>
<td>( BVel_{ij} = \sum_{tr \in TR_{ij}} \frac{\sum_{tr.t \in c} (tr.v)}{</td>
</tr>
<tr>
<td>Fare</td>
<td>( BFare_{ij} = \sum_{tr \in TR_{ij}} \frac{\sum_{tr.t \in c} (tr.f)}{</td>
</tr>
<tr>
<td>Stop Number</td>
<td>( BStop_{ij} = \sum_{tr \in TR_{ij}} \frac{\sum_{tr.t \in c} (tr.sn)}{</td>
</tr>
</tbody>
</table>

Specifically, for each OD pair \((i,j)\) and temporal slot \( c \), we compute \( BVol, BTime, BDist, BVel, BFare, BStop \) for bus and \( TVol, TTime, TDist, TVel, TFare, TStop \) for taxi. For example, \( BTime_{ij}^c \) is the bus travel time from origin \( i \) to destination \( j \) during the temporal slot \( c \). In Section III we will further leverage these statistics to extract features and build the transportation mode choice models. As we have contended earlier, the mobility patterns are significantly different across different temporal slots, and for that reason, we have partitioned the records into different temporal slots. Thus here, the aforementioned statistics are summarized for each temporal slot respectively. As a result, we will build the transportation mode choice model for each slot respectively.

In addition, using the transition records, we also compute some statistics of the routing network, which will be used later for the bus routing optimization. Specifically, for each direct connection edge \( e \in E \), we compute its travel distance \( d \) and travel time \( t \) for bus along the connection edge \( e \). We obtain \( d \) by projecting the head and tail stops of the edge to the map and calculate the shortest travel distance on the road map. To obtain the bus travel time \( t \), we consider the travel speed \( v \) on each edge obtained by using taxis as flowed sensors. Due to the speed difference between taxi and bus, we estimate the bus speed as follows: \( v_{bus} = \lambda \times v_{taxi} \), where \( \lambda \) is a constant [6]. Different cities may have a different \( \lambda \), here we set \( \lambda = 0.7 \) for Beijing. It follows that \( t_{bus} = \frac{1}{\lambda} \times t_{taxi} = t_0 + \frac{1}{\lambda} d/v_{taxi} \), where \( t_0 \) is a constant indicating the time for a bus stop, and we use \( t_0 = 1.5 \) minutes in Beijing. We represent all the edge travel distances in a vector \( EDist^c \in \mathbb{R}^{|E|} \), and all the edge travel time in \( ETime^c \in \mathbb{R}^{|E|} \), where \( c \) signifies the temporal slot when computing the statistics. As noted, these statistics can be specific for each temporal slot. Indeed, when the routing network is considered fixed, \( EDist^c \) is invariant with respect to the slot.

![Figure 4. Trip origin distribution of Beijing. The size of dot is proportional to the related number of trips.](image)
to \( c \); but \( ETime^c \) varies along with the traffic situations in different temporal slots.

### III. Transportation Mode Choice Model

In this section, we learn a transportation mode choice model to estimate the probabilities of people taking bus given origin, destination (the OD pair) and departing time (the temporal slot). To achieve this goal, we first extract features that contribute to the decision process of choosing transportation mode. Then a spatio-functionally weighted regression model is proposed to estimate the probability of taking bus \( p \) given these features.

#### A. Feature extraction

Understanding travel behavior and the reasons for choosing one transportation mode over another is an essential issue. However, travel behavior is complex. The choice of transportation mode is influenced by various factors, such as travel time, monetary cost, accessibility and reliability [7] [8]. Each transportation mode is influenced by various factors, such as travel distance and time, and choose buses for their lower cost. Here we focus on factors related to bus routing and consider other factors like accessibility and reliability remain unchanged.

Given the statistical summarization of an OD pair \((i, j)\) in a temporal slot \( c \), we extract the features \( X^c_{ij} \) to better describe the OD pair and compare the difference between the transportation modes:

\[
X^c_{ij} = \left( TDist^c_{ij}, BDist^c_{ij}, BTime^c_{ij}, \frac{BTime^c_{ij}}{TTime_{ij}}, TTime_{ij}, \frac{TTime_{ij}}{TTime^c_{ij}}, TStop^c_{ij}, \frac{TStop^c_{ij}}{TTime^c_{ij}}, \right)
\]

where details and our motivations are given as follows.

**Distance related features.** Distance influences people’s choice in an intuitive way. It is usually the first factor that comes to mind when traveling, e.g., how far is the destination from the origin. In this paper, distance related features include two parts: shortest road distance and distance ratio of buses and taxis. Here we use \( TDist \) to represent the shortest road distance of the OD pair, since it stands for the choice of experienced drivers which is usually the best in real. As shown in Figure 5 (a), with the increasing of distance of OD pairs, the percentage of people taking bus also increases. Here we use \( TTime \) as a baseline for the travel time of OD pair, and the travel time ratio of bus and taxi \( BTime/TTime \) describes the difference of these two. A larger \( BTime/TTime \), which is larger than 1, indicates a longer travel time by bus than taxi. As shown in Figure 6 (b), with the increase of \( BTime/TTime \) of OD pair, the percentage of people taking bus decreases.

**Monetary cost.** Monetary cost is another factor people need to consider. As shown in Figure 5 (a), with the increase in fare of OD pairs, the percentage of people taking bus increases. That’s because for long distances the taxi fare is much higher than bus. When the taxi fare is fixed, with the fare ratio of bus and taxi \( BFare/TFare \) increasing, we can see from Figure 7 (b) that the number of people taking bus decreases.

**Stop number related features.** Too many stops will affect the riding experience of a trip, not only is the stop a waste of time, but waiting is also an unpleasant process. One main advantage of a taxi is that it has no stop in the middle of a trip, while a bus has many stops. In this paper, we use the bus stop number per kilometer \( BStop/TDist \) to evaluate whether it affects people’s decisions to choose the bus. As shown in...
with an increase in \(B_{\text{Stop}}/T_{\text{Dist}}^T\), the percentage of people taking bus drops quickly.

![Figure 8](image)

**Figure 8.** Trip distribution wrt. stop number

### 2. Spatio-functionally weighted regression

Given the features \(\{X_{ij}^c\}\), and the historical trip numbers of buses and taxis, we are able to build a regression model to connect the features and people's transportation mode choices. Specifically, for a given temporal slot \(c\) and an OD pair \((o, d)\), we estimate the probability of taking bus as 
\[
p_{cd}^e = f((X_{od}^c, \Theta_{od})),
\]
where \(\Theta_{od}\) is the model coefficients. Since we want to estimate a probability distribution, we use the prediction function \(f(z) = \frac{1}{1 + \exp(-z)}\), which leads our model to logistic regression. The regression is fitted with the observations \(\{(X_{ij}^c, P_{ij}^c) : s_i, s_j \in S_i\}\), where \(P_{ij}^c\) is the probability of taking bus from origin \(s_i\) to destination \(s_j\) in temporal slot \(c\), estimated with historical transition records.

As noted, our model is specific to each temporal slot and also each different OD pair. This is desired because we note that transportation mode preferences vary over different OD pairs, due to differences in trip purposes and lifestyle. Indeed, different regions have different functions \([9]\), and the preferences of people from residential areas to commercial areas may differ from that of people from commercial areas to residential areas. On the other hand, travel preferences are more likely to be the same if two region pairs are near each other, sharing similar functions and lifestyles. To take these factors into account, we propose a spatio-functionally weighted logistic regression (SFWLoR). Other than that, a spatio-functionally weighted linear regression model (SFWLiR) which adopts linear regression instead of logistic regression is proposed for more efficient computation.

Specifically, when estimating \(\Theta_{cd}^e\), we define the observation weight for the observation \((X_{ij}^c, P_{ij}^c)\) as
\[
\omega_{od}^{(ij)} = \exp(-\frac{\alpha_{od}^{(ij)}}{2h_{\alpha}}) \cdot \exp(-\frac{\beta_{od}^{(ij)}}{2h_{\beta}}) = \exp(-\frac{\alpha_{od} - \beta_{od}}{2h_{\alpha} + 2h_{\beta}}),
\]
where \(h_{\alpha}, h_{\beta}\) are parameters that control the scaling at which the weights are computed, \(\alpha_{od}^{(ij)}\) is the spatial similarity of \((i, j)\) and \((o, d)\), and \(\beta_{od}^{(ij)}\) is the functional similarity of \((i, j)\) and \((o, d)\). These two are calculated as follows,
\[
\alpha_{od}^{(ij)} = \frac{\text{dist}(s_i, s_o) + \text{dist}(s_j, s_d)}{2},
\]
where \(\text{dist}(s_i, s_j)\) is the Euclidean distance between the bus stops in regions \(s_i\) and \(s_j\).
\[
\beta_{od}^{(ij)} = \frac{\text{dcos}(s_i, s_o) + \text{dcos}(s_j, s_d)}{2},
\]
where \(\text{dcos}(s_i, s_j)\) is the cosine distance calculated by
\[
\text{dcos}(s_i, s_j) = 1 - \frac{\langle n_i, n_j \rangle}{\|n_i\|\|n_j\|}.
\]
Since we want to measure the functional similarity in \(\beta\), the vector \(n_i\) contains the POI distribution of the \(i\)-th region. More details are given in Section V-A.

Note that the observation weight can also be extended by adding other similarities of regions if found to be impacting the choice of transportation mode.

### IV. BUS ROUTING OPTIMIZATION

Routing refers to the specifics of bus service alignment based on certain objective functions and a set of constraints, both as individual routes and as a system of routes working together \([10]\). In this section we start by formulating a bus routing optimization problem, in light of the transportation mode choice model in Section III. Following is our proposed solution to this problem.

#### A. Problem formulation

**A general problem formulation.** One main goal of bus routing optimization is to accommodate bus travel demand \([10]\). In this paper with the transportation mode choice model, we can estimate bus demand dynamically for different routing results, which further allows us to both accommodate and maximize bus travel demand.

Specifically, we denote a bus route by a sequence of bus stops \((\cdots, s_i, \cdots)\) and we search for the optimal bus routes which maximize the total number of bus riders. Given OD pair \((o, d)\) and the transportation mode choice model, one optimized routing \(R_{od} = (s_o, \cdots, s_i, \cdots, s_d)\) maximizes the bus riders of all stops traveled. In other words, \(R_{od}\) is the solution maximizing the objective function:
\[
F(R_{od}) = \sum_c \sum_{(s_i,s_j)\in R_{od}} \text{Vol}_{ij}^c \times f_{ij}^c(R_{ij}),
\]
where \((s_i,s_j) \in R_{od}\) and \(s_i < s_j\) indicate \(R_{od}\) passes \(s_i\) earlier and \(s_j\) later, \(R_{ij} = (s_i, \cdots, s_j)\) is the sub-route of \(R_{od}\) from stop \(s_i\) to \(s_j\). We compute \(f_{ij}^c(R_{ij}) = f((X_{ij}^c), \Theta_{ij}^c)\) as proposed in Section III. Later we will show how to derive the features \(X_{ij}^c\) with the route \(R_{ij}\).

This problem can be well fitted into new bus route design, where there previously were no bus routes. However, in this paper we aim to rework the existing bus routing, in which case, it is unnecessary to change well-designed bus routes but only flawed ones. Hence, the objective function for optimizing a flawed OD pair \((o, d)\) is simplified to
\[
F(R_{od}) = \sum_c \text{Vol}_{od}^c \times f_{od}^c(R_{od}).
\]

Furthermore, in a real application of bus routing optimization, multiple flawed OD pairs need to be considered simultaneously due to various constraints. For instance, the bus company (or government) is constrained by a limited budget which does not always allow for implementation of the identified optimal transit solution and service design. As an example, we have two flawed OD pairs as shown in Figure 9 (a) shows the routing result when optimize the two OD pairs independently,
leading to two routes which exceed the budget constraints; (b) shows the routing results using a multiple optimization method, leading to one route under the budget constraint.

![Image](image_url)

**Fig. 9.** Routing optimization comparison

**Bus routing optimization with constraints.** Following this line, the bus routing optimization problem is formulated as follows. Given choice models $f_k^c$ from Section III (i.e., $f_k^c$ parameterized by $\Theta_k^c = \Theta_k^c_{a_k, d_k}$), for each flawed OD pair $(o_k, d_k)$, $k = 1, \ldots, K$, we optimize the total bus ridership under budget constraints. Supposing the optimal bus routes are $R = \{R_k : k = 1, \ldots, K\}$, where $R_k$ has an origin $o_k$ and a destination $d_k$, our objective function is as follows,

$$F(R) = \sum_k \sum_c Vol_k^c \times f_k^c(R_k).$$

(6)

where $Vol_k^c = Vol_{o_k, d_k}^c$. As stated previously, $f_k^c(R_k) = f((X_k^c, \Theta_k^c))$ and we will show later how to derive the features $X_k^c$ of $R_k$.

We consider multiple budgets (e.g., total route length, total service time) under the following constraints:

$$\text{cost}(R) \leq C,$$

(7)

where the function $\text{cost}$ is calculated with all the bus routes in $R$, and the budgets allowed to stay within are defined in vector $C$. Note that when there is no budget constrain or the budgets are large enough, the above problem becomes an independent routing problem for each OD pair.

Next we will show how to detect OD pairs with flawed bus routing, after which, a solution to the problem defined above is proposed.

**B. Flawed OD pair identification**

To rework the existing bus network, we first detect OD pairs with flawed bus routing as the candidates to be optimized. For instance, people may have to take a long detour traveling on existing bus routes or there may even be no bus routes traveling through two regions with high travel demand. People traveling between these OD pairs have a relatively low probability of taking bus since it is so inconvenient.

Routes traveled by taxi usually indicate the practically best driving directions. It is reasonable for us to use the travel route of taxi for each OD pair $(i, j)$ as the upper bound of the bus route. Then, with the travel route of taxi $R_{T, ij}^c$, in temporal slot $c$, we can derive the features $X_{ij}^c$ of $R_{T, ij}^c$. Finally, with the above information and the transportation mode choice model, we are able to calculate the upper bound of probability of taking bus for every OD pair.

For all OD pairs, we rank them in descending order according to the potentially increased bus rider number, which is calculated as follows:

$$\Delta BVol_{ij} = \sum_c Vol_{ij}^c \times (f_{ij}^c(R_{T, ij}^c) - P_{ij}^c),$$

(8)

and top $K$ OD pairs will be selected as candidates for bus routing optimization.

**C. Problem solution**

To find the optimal route, we consider the routing network $G = (S, E)$. For each edge $e = (i, j) \in E$ connecting the bus stops $s_i$ and $s_j$, we define $R_k \in \mathbb{R}^{|E|}$, where $R_{ke}$ is 1 if and only if route $R_k$ passes edge $e$, and $R_{ke} = 0$ otherwise. Also, for each bus stop $s \in S$, we define $in(s) = \{s', s \in E\}$ and $out(s) = \{s, s' \in E\}$ as the incoming edges and outgoing edges of $s$. Then, for $k = 1, \ldots, K$, we have

$$\sum_{e \in out(o_k)} R_{ke} = \sum_{e \in in(d_k)} R_{ke} = 1,$$

$$\sum_{e \in in(o_k)} R_{ke} = \sum_{e \in out(d_k)} R_{ke} = 0,$$

$$\sum_{e \in out(s)} R_{ke} = \sum_{e \in in(s)} R_{ke}, \forall s \neq o_k, d_k.$$

Should be noted that, $R_k$ passes bus stop $s$ if and only if $\sum_{e \in out(s)} R_{ke} = \sum_{e \in in(s)} R_{ke} = 1$ for $s \neq o_k, d_k$.

To derive the features $X_k^c$ for route $R_k$ at temporal slot $c$, we have

$$BDist_{R_k}^c = \sum_e R_{ke} \times EDist_e^c = \langle EDist^c, R_k \rangle,$$

$$BTime_{R_k}^c = \sum_e R_{ke} \times ETTime_e^c = \langle ETTime^c, R_k \rangle,$$

$$BStop_{R_k}^c = \sum_s \sum_{e \in out(s)} R_{ke} = \sum_e R_{ke} = \langle 1, R_k \rangle,$$

where $EDist^c, ETTime^c \in \mathbb{R}^{|E|}$ are travel distance and time on edges (introduced in Section II), and $1 \in \mathbb{R}^{|E|}$ are vector of ones. We will also use $0 \in \mathbb{R}^{|E|}$ as vector of zeros.

By letting

$$A_k^c = (0; \frac{1}{TDist_k^c} EDist^c; 0; \frac{1}{TTime_k^c} ETTime^c; 0; 0; \frac{1}{TDist_k^c} 1),$$

$$B_k^c = (TDist_k^c; 0; TTime_k^c; 0; TFare_k^c; BFare_k^c / TFare_k^c; 0),$$

we have $X_k^c = A_k^c R_k + B_k^c$.

For the constraints, we limit the service route length and driving time introduced per unit time by the overall routing $R$ on all traveled edges. By letting $WTTime_k^c$ be the service waiting time of route $R_k$ at temporal slot $c$, if $WTTime_k^c = WTTime^c, \forall k = 1, \ldots, K$, this cost can be written as

$$\text{cost}(R) = \sum_c STime^c \times \sum_e (\sum_k R_{ke} > 0) |ECost_e^c|,$$

where $ECost_e^c = (EDist_e^c; ETTime_e^c)$ is a two dimensional column vector encoding both the travel distance and time on
A relaxed calculation which avoids the boolean test operator $(\lnot)$ can be formulated as
\[
\text{cost}(\mathbf{R}) = \sum_{c} \frac{STime_e^c}{WTime_e^c} \sum_{k} R_{ke} \cdot ECost_e^c \\
= \sum_{c} \sum_{k} \frac{STime_e^c}{WTime_e^c} BCost_{R_k},
\]
where $BCost_{R_k} = (BDist_{R_k}^c; BTime_e^c)$ encodes the route travel distance and time of $R_k$. As noted, this also allows us to calculate different waiting time for different bus routes. When we do not focus on the scheduling of bus, we use 15 minutes as the waiting time for all bus routes in this paper. In sum, our constraints in Equation 7 can be linear with respect to the decision variables in $R$. However, the objective in Equation 6 is non-linear with the prediction function $f(z) = \frac{1}{1 + \exp(-z)}$, and the consequent optimization problem is non-convex and the gradient directed searching will result only a local optimal. We also exploit the choice model with a linear prediction function $f(z) = z$, which leads to a constrained linear programming problem. In experiments, we will show results of the routing optimization with both non-linear and linear prediction functions, and it can be seen that the relaxed linear approach can approximate optimal routing effectively.

In general, the resultant integer programming is NP-complete. However, since we optimize only the most flawed OD pairs instead of the overall bus routing, the problem is of a reasonable scale and it turns out that the branch-and-bound algorithm [12] can solve the problem efficiently for flawed bus routing in Beijing. In the more general cases, we can also relax the binary requirements to $R_{ke} \in [0, 1]$, which can be interpreted as the probability of route $R_k$ passing edge $e$. A solution of the relaxed problem signifies how we should route the bus from origin to destination, so that the maximum transportation needs are satisfied by the bus service. To recover the solution for the un-relaxed problem, we can iteratively remove the edge with the smallest probability, until there is a unique route for an OD pair.

V. EXPERIMENTAL RESULTS

In this section we first introduce the data and settings of our experiments. Then we evaluate the results of the proposed transportation mode choice model, followed by the evaluation of flawed OD pairs. Finally we show the results of our bus routing optimization model.

A. Data and settings

**Bus transactions.** Bus transactions are generated by BMAC smart card system installed on all the buses in Beijing. We select the data from the same time span as the taxi data, from August to November, 2012. This dataset contains the following information: card id, bus route number, boarding and alighting, time, fare [13]. Note that a random sampling method is used to recover bus trips to match taxi trips, where the ratio of bus trips to taxi trips is about 3:5.

**Taxi GPS traces.** These taxi GPS traces are generated by about 30,000 taxis in Beijing from August to November, 2012. Each GPS point is associated with a label indicating if the taxi is occupied or not. Here we only focus on the occupied points which form taxi trips of riders, from pick-up points to drop-off points. Table III shows some statistics of the two trip datasets.

**Bus routes and road map.** 1) We have the bus route data, which contains 2,427 stops and 1,058 routes in the urban area of Beijing. After we merge the redundant stops, we obtain 1,250 stops and we partition the urban area into 1,250 regions accordingly. We use the stops/regions as nodes of our routing network. 2) We have the road map data containing 196,307 road segments and their locations. We use this data to construct the connection edges of our routing network. For the 1,250 routing nodes, we have 3,855 connection edges.

**POI data.** A Beijing POI dataset in the year 2012 is employed to compute the functional observation weights. The number of POIs $n_i = \langle n_1, \ldots, n_{10} \rangle$ in region $s_i$ is counted following the categories shown in Table IV.

**Platform.** The algorithms are implemented in Matlab 2013b and C# on Visual Studio 2012. All the experiments are conducted on a 64 bit machine with 3.40 GHz Intel Core i7 CPU and 16GB memory.

### TABLE III. STATISTICS OF THE DATASETS.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Properties</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Taxi GPS Traces</td>
<td>Number of taxis</td>
<td>20,964</td>
</tr>
<tr>
<td></td>
<td>Effective days</td>
<td>114 (77 weekdays, 37 weekends)</td>
</tr>
<tr>
<td></td>
<td>Number of occupied GPS points</td>
<td>333M</td>
</tr>
<tr>
<td></td>
<td>Number of occupied trips</td>
<td>19M</td>
</tr>
<tr>
<td></td>
<td>Total trip distance(km)</td>
<td>156M</td>
</tr>
<tr>
<td>Bus Transactions</td>
<td>Number of bus stops</td>
<td>7,810</td>
</tr>
<tr>
<td></td>
<td>Number of car holders</td>
<td>701,250</td>
</tr>
<tr>
<td></td>
<td>Number of trips</td>
<td>10M</td>
</tr>
</tbody>
</table>

### TABLE IV. CATEGORY OF POIS.

<table>
<thead>
<tr>
<th>Category</th>
<th>Sub-categories</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Home</td>
<td>Apartment building</td>
<td>29,246</td>
</tr>
<tr>
<td>2 Work</td>
<td>Government &amp; office building</td>
<td>71,915</td>
</tr>
<tr>
<td>3 Education</td>
<td>School, training center</td>
<td>15,489</td>
</tr>
<tr>
<td>4 Food</td>
<td>Restaurant</td>
<td>36,723</td>
</tr>
<tr>
<td>5 Shopping</td>
<td>Shop, mall, outlet</td>
<td>56,520</td>
</tr>
<tr>
<td>6 Entertainment</td>
<td>Museum, theater, club</td>
<td>7,897</td>
</tr>
<tr>
<td>7 Outdoor</td>
<td>Park, sports field</td>
<td>2,211</td>
</tr>
<tr>
<td>8 Transportation</td>
<td>Airport, railway &amp; bus station</td>
<td>15,287</td>
</tr>
<tr>
<td>9 Health care</td>
<td>Hospital, medical center, pharmacy</td>
<td>9,768</td>
</tr>
<tr>
<td>10 Car service</td>
<td>Car sale, repair, gas station</td>
<td>10,781</td>
</tr>
</tbody>
</table>

**B. Transportation Mode Choice Model**

**Baselines.** We evaluate the effectiveness of our spatio-functionally weighted regression (SFWLoR, SFWLiR) with a set of baselines, including unweighted logistic regression (LoR), temporal logistic regression (TLoR) and temporal linear support vector machine (TLiSVM).

- A Logistic Regression model (LoR) on the data before segmented to temporal slots. That means we treat the whole day as one temporal slot and it evaluates if the preference changes through the day.
- A Temporal Logistic Regression model (TLoR), which estimates people’s choices in different temporal slots.
• A Temporal Linear Support Vector Machine (TLiSVM), which estimates people’s choices in different temporal slots.

We use the receiver operating characteristics (ROC) curve and the area under ROC (AUC) to evaluate the performance of the transportation mode choice models. The ROC curve is obtained by drawing pairs of sensitivity and false positive rate (1-specificity) at different cutoff points, i.e., every 0.01 from 0 to 1 in our experiments. The sensitivity (sens) is defined as the proportion of true positives as compared to the total positive class, whereas specificity (spec) comprises the proportion of true negatives in relation to the total negative class.

\[
sens = \frac{tp}{tp + fn},
\]

\[
spec = \frac{tn}{tn + fp},
\]

where \(tp\), \(fp\), \(tn\) and \(fn\) are true positives, false positives, true negatives and false negatives, respectively.

**Results.** We evaluate the models with 10-fold cross-validation in each temporal slot separately, and then use the average of different temporal slots as the final result. Figure 10 (a) shows the overall performance of each method, and SFWLoR on each temporal slot. From the figure we can see SFWLoR outperforms other methods. The models perform better on weekdays (Slot 1,2,3,4) than on weekends (Slot 5,6,7), because there is a lot of variation occurs on weekend trips as compared to weekday trips and it increases the difficulty of modeling.

**C. Flawed OD pairs**

Experimental results of flawed OD pair identification are shown in Figure 12: (a) shows the changes of probability (green line) and volume (blue line) of taking bus after using taxi routes as the upper bound of bus routes; (b) shows us top 20 flawed OD pairs.

From Figure 12 (a) we can see with the improvement of bus routing, an average of 5 percent increase of probability taking bus is expected for all OD pairs. Moreover, we find that the bus volume increase follows Zipf’s law, which means most of the volume increase happens among a few OD pairs. This further validate our method which focuses on these flawed OD pairs instead of all of them.

**D. Routing Optimization**

Given the top \(K\) flawed OD pairs with a descending rank of potential increases of bus riders, we evaluate our objective function (Maximum Rider, MR) on different \(K\). Two different solutions for MR, MR-LiR and MR-LoR, are presented, using linear and logistic regression choice models respectively. We use shortest distance (SD) and shortest time (ST), which are widely used routing methods in practice, as baselines of our method. Accordingly, the objective functions of these two
Results of top 100, 200, and 500 flawed OD pairs are shown in Table V, where the columns show average values of statistics of each OD pair. Specifically, we first use these methods to find the best routes $\mathbf{R}$ for our identified flawed OD pairs. For every OD pair in different temporal slots, the transportation mode model is used to predict the probability of taking bus $\hat{p} = f(\mathbf{R})$. Together with the travel demand of each pair, the bus rider number can be obtained. By comparing this to historical bus rider numbers, we then get the change of bus rider number $\Delta BVol$. Please note that here we only use taxis to represent private transportation, and the real effect of buses can be enlarged when other private transportation modes (e.g., private car) are considered.

As shown in Table V, we can see MR-LoR provides routes that lead to highest probabilities of people taking bus, because MR-LoR successfully measures the trade-off between different factors. While MR-LiR obtain second best routing results. On the other hand, we see ST performs better than SD, which indicates people consider time a more important factor than distance. Although some of the routes found by our method are the same as results found by either SD or ST, we can still provide suggestions on the selection of these two. From this point of view, our transportation mode choice model can serve as a criteria for choosing candidate bus routes.

### Table V. Results of bus routing on top K flawed OD pairs.

<table>
<thead>
<tr>
<th>K</th>
<th>Methods</th>
<th>BDist (km)</th>
<th>BTime (hour)</th>
<th>BStop (#)</th>
<th>BFare (CNY)</th>
<th>Prob.</th>
<th>$\Delta BVol$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>SD</td>
<td>3.83</td>
<td>0.28</td>
<td>3.91</td>
<td>0.44</td>
<td>0.83</td>
<td>445.6</td>
</tr>
<tr>
<td></td>
<td>ST</td>
<td>3.84</td>
<td>0.27</td>
<td>4.02</td>
<td>0.44</td>
<td>0.83</td>
<td>446.7</td>
</tr>
<tr>
<td></td>
<td>MR-LiR</td>
<td>3.65</td>
<td>0.29</td>
<td>3.83</td>
<td>0.44</td>
<td>0.85</td>
<td>669.3</td>
</tr>
<tr>
<td></td>
<td>MR-LoR</td>
<td>3.45</td>
<td>0.28</td>
<td>3.67</td>
<td>0.44</td>
<td>0.86</td>
<td>731.2</td>
</tr>
<tr>
<td>200</td>
<td>SD</td>
<td>4.23</td>
<td>0.53</td>
<td>4.86</td>
<td>0.48</td>
<td>0.81</td>
<td>245.8</td>
</tr>
<tr>
<td></td>
<td>ST</td>
<td>4.30</td>
<td>0.30</td>
<td>4.52</td>
<td>0.49</td>
<td>0.82</td>
<td>301.1</td>
</tr>
<tr>
<td></td>
<td>MR-LiR</td>
<td>4.41</td>
<td>0.34</td>
<td>4.33</td>
<td>0.49</td>
<td>0.84</td>
<td>427.3</td>
</tr>
<tr>
<td></td>
<td>MR-LoR</td>
<td>4.52</td>
<td>0.35</td>
<td>4.40</td>
<td>0.49</td>
<td>0.85</td>
<td>481.2</td>
</tr>
<tr>
<td>500</td>
<td>SD</td>
<td>4.31</td>
<td>0.34</td>
<td>4.77</td>
<td>0.48</td>
<td>0.79</td>
<td>198.4</td>
</tr>
<tr>
<td></td>
<td>ST</td>
<td>4.51</td>
<td>0.32</td>
<td>4.78</td>
<td>0.49</td>
<td>0.81</td>
<td>225.8</td>
</tr>
<tr>
<td></td>
<td>MR-LiR</td>
<td>4.54</td>
<td>0.33</td>
<td>4.64</td>
<td>0.49</td>
<td>0.83</td>
<td>316.4</td>
</tr>
<tr>
<td></td>
<td>MR-LoR</td>
<td>4.58</td>
<td>0.34</td>
<td>4.71</td>
<td>0.49</td>
<td>0.84</td>
<td>342.3</td>
</tr>
</tbody>
</table>

A real example of bus routes found for flawed OD pairs is shown in Figure 13 where includes two flawed OD pairs (Xiaohongmen, Qianmen) and (Shazikou, Qianmen). The routes generated by SD,ST, and MR are shown in green, red, and blue lines respectively. From the figure we can see SD and ST both generate two routes, which are similar to each other, while MR generates a single route traveled through these two OD pairs. Moreover, we found that this route share same subroutes with bus line 93 which is newly added by the Beijing Bus Company from March, 2013.

**Efficient Study.** Figure 14 presents the efficiency of the four methods for different $K$. From this figure we can see ST and SD are the fastest among these four, while MR-LoR costs the most time for computing results. We note that MR-LiR is much faster than MR-LoR but the performance is not much worse. In real applications, MR-LiR would be recommended for large-scale bus routing. Since this application usually works in an offline manner, MR-LoR would also be used for better planning results.

### VI. Related Work

Our work is related to two research areas, the first one is human mobility pattern mining and the second one is bus routing optimization.

#### A. Human mobility pattern

Understanding human mobility in urban environments is central to traffic forecasting, location-based services, and urban reconstruction. A significant number of papers on human mobility analysis have been published in recent years thanks to widely available mobility data, such as GPS data [17], cellular network data [18], [19], and public transportation transaction data [20], [21], [22], [23], [24], [25]. suggest that human mobility patterns follow a high degree of spatial and temporal regularity and are thus predictable. [26] suggest that human mobility has a predictability of 93%. [27] also reports on the consistency of daily travel patterns with public transportation transaction data.

Other than theoretical analysis, mining human mobility data has also enabled a variety of emerging applications, such as urban planning [5], region function analysis [9], hotspots detection [18], driving route recommendation [11], [28], and bus route planning [29]. Recently, Chen et al. [6] investigated night bus route planning using large scale taxi GPS traces, which aims to find a bus route with a fixed frequency, maximizing the number of passengers expected along the route subject to the total travel time constraint. Unlike the above mentioned work focusing on single mode human mobility pattern, we integrate heterogeneous human mobility data together to better represent the mobility of a city. Thus we are able to analyze the difference and relation...
between different human mobility patterns and make planning in the city level.

B. Bus routing optimization

Bus network design addresses the problem of how to design a new bus network or to redesign an existing network [1] [8]. It is an intensive studied area in the urban planning and transportation field, known to be a complex, non-linear, non-convex, multi-objective NP-hard problem [6]. Certain objectives are optimized, subject to constraints and travel demands. Some widely used objectives include shortest distance, shortest travel time, maximum passenger flow, and maximum area coverage while constraints include time, capacity and resources [3] [30]. However, the selection of objectives should take care of the operator as well as user requirements which are often conflicting, leading to design trade-off rather than an optimal solution. Various models have been proposed in this area, including cost-oriented models, passenger-oriented models, game-theoretic models and location-based models [3]. The above work assumes that travel demands are determined by user survey or population estimation, and provide sub-optimal solutions based on heuristic procedures. Different from that, we integrate the bus network design problem with real mobility demand, also the prediction of travel demand is based on different routing results.

VII. CONCLUSION

In this paper, we focused on the identification and optimal planning of the flawed bus routes to improve the utilization efficiency of the public transportation service, according to the transportation mode choice model built on real data. First, we partitioned the urban area into disjoint regions on which an integrated analysis of the taxi traces and the bus transactions is conducted. Second, based on the integrated analysis, we proposed a localized transportation mode choice model, with which we can dynamically predict the bus travel demand for different bus routing. Then, we leveraged this model to optimize the bus routes by maximizing the bus ridership with budget constraints. At last, we provided a solution for the identified most flawed region pairs in the urban area. Extensive studies, which validated the effectiveness of our methods, were performed on real world data collected in Beijing which contains 19 million taxi trips and 10 million bus trips.

ACKNOWLEDGMENT

This research was partially supported by the National High Technology Research and Development Program of China (863 Program) via grant number SS2014AA012303.

REFERENCES