Data-driven Location Selection for Battery Swapping Stations

Ming Zeng¹, Yifei Pan², Danyang Zhang³, Ziyang Lu³, and Yong Li³, Member, IEEE
¹School of Information and Electronics, Beijing Institute of Technology, Beijing, 100081 China
²Peking University, Beijing, 100080 China
³Department of Electrical Engineering, Tsinghua University, Beijing, 100084 China

Electric Vehicles (EVs) have been encouraged to penetrate deeper in the vehicle market for the green transportation system. One of the key issues to promote EV industry is to deploy Battery Swapping Stations (BSSs) that can satisfy the electricity demand of EV users. Since large scale data of vehicles such as GPS locations and electricity requests can be collected, the data-driven approach can be a cost-effective and useful method to select the locations of BSSs. In this paper, we propose a data-driven framework to solve the BSS location selection problem based on a large scale of GPS data of taxies in metropolitan area. The proposed solution consists of three main steps: Hidden Markov Model (HMM) based map matching and trajectory extraction, electricity consumption rate model based battery swapping demand estimation and clustering strategy based BSS location determination. Compared to the state-of-art deployment baseline, our proposed scheme is more easily to implement in reality and the mean distance error between the location that battery swapping demand is generated and the nearest BSS is reduced by 52.5% and 62.7%, which will definitely reduce the range anxiety of EV users and help improve the will of using EVs.

Index Terms—Battery swapping station, GPS data, map matching, location selection, clustering.

I. INTRODUCTION

Electric Vehicles (EVs) are the most economic and environment friendly transportation tools. The emission of CO₂ for gasoline car is around 0.3 kg/mile, while substituted by EV, the CO₂ emission can be reduced to about 0.16 kg/mile [1]. Moreover, the energy cost for using EV is significantly decreased [2]. However, driving range and recharging of EVs are key concerns when people decide whether to buy EVs. So far, Tesla Model X [3] has the maximum driving range of 565km, while the average driving range of other EVs is 200km [2]. The exhausted battery of Tesla Model X will be charged to full level in 80 minutes with super charger, while it takes 10 hours to be fully charged if charged at home, and there are only 400 super charger installed in China. Due to the insufficient deployment of recharging infrastructures and the long waiting time for full charging, EVs are not the best choices for people who want a transportation tool both for intercity and intracity commuting. Therefore, deploying Battery Swapping Stations (BSSs) is a promising solution for the development of the EV industry [4], which facilitates people using EVs to change exhausted batteries into full-charged batteries conveniently and reduce the range anxiety [5]. Backup batteries are stored in BSSs, so that EV drivers can replace the drained batteries in a few minutes once EVs are driven into BSSs.

Nevertheless, constructing the battery swapping network is difficult since many factors should be cautiously considered, such as geographic locations of BSSs, initialized battery capacity of each BSS and battery replenishment strategy. Deploying BSSs in reasonable places can increase the probability that EV users find BSSs when their EVs need to be charged. However, deploying BSSs in hot spot regions will introduce some new issues, such as high cost of urban land and traffic congestion. Second, the mismatching between the real time electricity demand and the battery capacity of BSSs can cause power supply failures, which have severe impact on the user experience of EV usage. Third, different EV drivers will choose to replace the battery with different amount of electricity remained according to their diverse anxiety and specific future travel plans. To prevent EVs with urgent electricity demand from not being met, mechanism should be designed to guide the battery replenishment. In this paper, we focus on the BSS deployment problem, which is the initial and critical step in constructing the battery swapping network. However, deploying BSSs according to EVs’ battery swapping demand in reality is non-trivial due to the following five reasons.

First, to the best of our knowledge, there is not any public dataset recording the charging activities of EVs in metropolitan area. Second, it is hard to collect the trajectory data generated by EVs since we cannot tell whether it is the trajectory of an EV or a gasoline vehicle because the sampled GPS data of vehicles does not contain the information about vehicle types. Third, mapping raw GPS trajectories to roads on a digital map is a challenging work because of the positioning drift problem and measurement errors in GPS. Fourth, the electricity consumption rate model of EVs is not perfect yet, which introduces non-negligible errors when estimating the battery swapping demand of EVs. Fifth, the battery swapping demand of EVs in metropolitan level varies temporally and spatially. As a result, existing solutions on BSS placement problem are model based schemes instead of data driven approaches, which are not designed according to the real electricity demand of EVs in metropolitan areas.

To conquer these challenges, we aim to investigate BSS location selection based on the sampled GPS data of taxies in this work. The two main reasons for us to choose sampled GPS data of taxies as follows. On one hand, it is relatively easy to collect GPS data of taxies, which reflects the city-
wide travel demand of some people. One the other hand, it can be well justified to estimate the travel demand of people using EVs based on the raw taxi GPS data because both taxies and EVs are often used as intracity commuting tools. Hence, we assume that all taxies in our dataset used in this paper are substituted by EVs. Nevertheless, it is yet non-trivial to select the BSS locations based on the sampled GPS data of taxies, due to the new emerging challenges introduced by the characteristics of taxi trajectories. The typical route in the taxi trajectory is that after passenger gets off, taxi often turns back. In addition, taxies are often driven into residence communities in which the trails are usually not marked on the digital map. These phenomena further degrades the accuracy when mapping the sampled GPS data to the roads on a digital map. Hence, to obtain the acceptable accuracy when mapping raw GPS trajectory data to roads on a digital map, we first apply a Hidden Markov Model (HMM) based algorithm to decide the actual road segment that each EV is on. Then, after map matching, we could obtain the velocity and accelerated velocity of each EV. The electricity consumption process of each EV can be inferred based on the electricity consumption rate model. Thus, we can determine when and where EVs need to swap the batteries. Gathering the electricity demand of all EVs, different clustering algorithms are applied to find the spatial and temporal distributions of battery swapping demand. According to the distribution of electricity demands, places with intensive battery demand is selected as BSS location candidates.

Our main contributions in this paper are summarized as follows:

- We propose a data-driven approach for solving the BSS location selection problem to satisfy the battery swapping demand of EVs. A large scale GPS data of 13,700 taxies is utilized to estimate the electricity demand of EVs in metropolitan area.
- Inferring from historical trajectories of taxies, we find the temporal and spatial characteristics of the electricity demand of EVs in a metropolitan area in China. Intensive electricity requests occur in the downtown of city and on the way to airports. The peak electricity demand on weekdays emerges between 21:00 and 22:00, while peak demand on weekend comes an hour earlier.
- Our proposed BSS location selection approach consists of HMM based map matching scheme, electricity demand estimation model and two kinds of clustering algorithm (K-Means clustering and hierarchical clustering). Compared to deploying BSSs uniformly, our proposed approach can avoid the construction of super BSSs with more than 2,000 batteries stored. Meanwhile, the average distances to the nearest BSSs when EVs have electricity demand under the K-Means clustering and hierarchical clustering deployment schemes are reduced 62.7% and 52.5%, respectively, compared to the uniform deployment scheme.

The rest of this paper is organized as follows. In Section II, we introduce the existing literatures related to our work from three aspects: charging and battery swapping, facility deployment and data-driven approaches. In Section III, we formulate the BSS location selection problem as an optimization problem, and present the data-driven framework of our solution. Section IV elaborates the details of the proposed solution in each step. In Section V, we describe the dataset we utilize in our experiment and evaluate the performance of the proposed data driven schemes compared with the uniform deployment scheme. Section VI concludes our work.

II. RELATED WORKS

In this section, we discuss the existing literatures related to our work, focusing on three research fields: charging and battery swapping, facility deployment and data-driven approaches.

Charging and battery swapping are two main ways for EVs to extend their driving ranges. On one hand, some researchers focus on charging scheduling scheme design to improve user experience of EV users and reduce the operation cost of power grid. Yang et al. [6] proposed an EV route selection and charging navigation optimization model based on crowd sensing, aiming to reduce the travel cost of EV users and improve the load level of the distribution system concerned. Eljdali et al. [7] proposed a model for EV charging and discharging scheduling at public supply stations, considering demand-supply curve balance as a constraint. Gusriladi et al. [8] developed a strategy consisting of a distributed scheduling algorithm and a cooperative control policy for individual EVs which optimized the operation of the overall charging network on a highway. On the other hand, some other researchers consider battery swapping as a better way for EV users as it saves a lot of waiting time for EV users. Wu et al. [9] assigned an optimized charging schedule for each incoming battery in a BSS to minimize the operation cost of the BSS. You et al. [10][11] assigned a best station to each EV to swap its depleted battery based on the current location and state of charge in a centralized and distributed manner, respectively. However, most existing literatures investigated the impact of the electricity demand of EVs on electric energy networks and ignored the mutual effect between the electricity demand and travel demand of EVs. Since satisfying the travel demand of EV users is a key issue that most EV users concern, in this paper, we consider battery swapping as an effective and fast way to change depleted batteries into full charged batteries, and investigate a BSS location selection problem based on the travel demand of EVs.

Facility deployment has been studied for a long time, including the deployment of billboard [12], ambulance station [13], relays in communication networks and so on. In these deployment problems, the essential problem is to satisfy the service requirements within limited resource constraints. Liu et al. [12] attempted to employ visual analytic techniques to launching a successful billboard campaign by using taxi GPS trajectory data. Li et al. [13] focused on locating the ambulance stations by using the real traffic information so as to minimize the average travel time to reach the emergency requests. Karamshuk et al. [14] extracted features that had impact on the popularity of retail stores and suggested new
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2941901, IEEE Access

retail store locations based on those mined factors. Chen et al. [15] formulated the bike station placement issue as a bike trip demand prediction problem, and proposed a semi-supervised feature selection method to predict the bike trip demand. Zhang et al. [16] proposed a placement planning scheme for a grid-connected charging station to find an optimal combination of photovoltaic panels, local battery cells as well as EV chargers. These related works did not apply data-driven approaches to develop facility deployment schemes, meanwhile, few literature studied the BSS deployment problem. Hence, our paper applies data-driven approach which uses a large scale of sampled GPS data of taxies to find BSS location candidates, which can better match the electricity demand of EVs in reality.

**Data-driven approaches** are increasingly widely used in urban computing research field, since a tremendous amount of data is collected through various types of sensors deployed in a city including smart phones in-hands. Wei et al. [17] reviewed the data-driven approaches that had well addressed a large variety of building energy related applications, such as load forecasting and prediction, energy pattern profiling and regional energy-consumption mapping etc. Eggimann et al. [18] reviewed several data-driven approaches for better rain-data management, urban pluvial flood-risk management and forecasting the usage of drinking water. Li et al. [19] formulated the charging station deployment problem as an integer linear programming problem, with an objective of minimizing the average time for EV. Trajectory data of 490 EV taxies was utilized to evaluate the time of seeking charging stations and waiting time at charging stations in the optimization problem. However, the scale of the dataset utilized in [19] was small, and the noise introduced when extracting charging events was large. In our work, the battery swapping demand is estimated based on the electricity consumption rate model and a dataset including 13,700 taxies.

### III. SYSTEM OVERVIEW

In this section, we model the BSS deployment problem as an optimization problem and give a holistic introduction of our data-driven solution framework.

#### A. Problem Formulation

Since the deployment of a BSS will occur facility and daily maintenance cost, the government or the power utility that is in charge of BSS deployment always has budget constraint. In this paper, we aim to minimize the average distance between the location where battery swapping demand is generated and the nearest BSS though a limited number of BSSs are deployed in the city. It is assumed that the budget is enough for deploying $N_B$ BSSs. Each BSS reserves $C_i$ ($i \in \{1, 2, ..., N_B\}$) kWh electricity each day, consisting of $N_i$ batteries with given specification. BSSs will be replenished with full charged batteries once everyday, hence, the second task in our problem is to determine the capacity $C_i$ of each BSS. In addition, EVs can swap their batteries successfully if at least one of the deployed BSSs is within reachable distance as EVs usually request for battery swapping when the State-of-Charge (SoC) of their batteries is below a threshold $T_h$. Threshold Th is set to be 20% intuitively, since according to average driving range of most EVs, an EV can drive at most 40km with remaining 20% of the EV battery. Sometimes 40km is not even enough for daily commuting in super large-scale city. Meanwhile, we refer to the battery management of cellphone, when the battery level of cellphone drops to 20%, cellphone will automatically remind the user to charge immediately. Thus, we have the following constraints represented by (1) and (2), respectively:

1. **BSS capacity constraint**
   \[
   \sum_j x^j_i \leq N_i, \forall i \in \{1, 2, ..., N_B\},
   \]
2. **Reachability constraint**
   \[
   \sum_i y^{ij}_{j,t} \geq 1, \quad y^{ij}_{j,t} = \begin{cases} 0 & \|l^j_t - l_i\| > d_j \\ 1 & \|l^j_t - l_i\| \leq d_j \end{cases},
   \]

where $x^j_i$ and $y^{ij}_{j,t}$ are indicators, denoting whether EV $j$ swaps battery in station $i$ and whether station $i$ is in the reachable range of EV $j$ at time $t$, respectively. $l^j_t$ and $l_i$ represent the location of EV $j$ at time $t$ and the location of station $i$, respectively. $\|l^j_t - l_i\|$ denotes the geographic distance between EV $j$ and station $i$ at time $t$.

The objective of the BSS deployment problem is to minimize the average distance to the BSS when EVs have electricity demand. Thus, the optimization problem can be formulated as (3)-(5). Since the search space of BSS locations is a continuous space, solving the following optimization problem is NP-hard. Hence, we propose a data-driven approach based on a large scale of taxi trajectory data to find the feasible solution for urban BSS deployment. In the next subsection, we will introduce the framework of our proposed solution.

\[
\begin{align*}
\min_{l_i, i \in \{1, 2, ..., N_B\}} & \frac{1}{\sum_{j \in N_E} x^j_i} \sum_{j \in N_E} x^j_i \|l^j_t - l_i\|, \\
\text{s.t.} & \sum_j x^j_i \leq N_i, \forall i \in \{1, 2, ..., N_B\}, \\
& \sum_i y^{ij}_{j,t} \geq 1, \quad y^{ij}_{j,t} = \begin{cases} 0 & \|l^j_t - l_i\| > d_j \\ 1 & \|l^j_t - l_i\| \leq d_j \end{cases},
\end{align*}
\]

#### B. Solution Framework

Fig. 1 shows our BSS deployment framework. It takes two dataset as input: sampled GPS data of taxies and road network data. The whole system consists of five main steps: trajectory mapping, vehicle velocity estimation, electricity consumption calculation, battery swapping demand inference and BSS location determination. First, in order to map raw GPS data to road networks in a digital map, we design a Hidden Markov Model (HMM) with an adjustable parameter that controls the mapping precision. To solve the HMM, we apply Viterbi algorithm. However, we can only obtain the driving path of taxies without timestamps. Then, we design Trajectory Time Recovery (TTR) algorithm to add timestamps on the pathes. After map matching, the mapped trajectories are utilized to estimate the battery swapping demand of each EV based on the electricity consumption rate. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.
model. Once we determine when and where each EV needs to swap the battery, we apply clustering algorithms to find the spatial and temporal distributions of the battery demand of all EVs. Since the government or the utility companies which are in charge of the infrastructure construction have limited budget, when choosing a suitable number of BSSs \(K\), both the electricity demand of EVs and the construction cost under budget should be considered. The number of clusters determined in the clustering algorithms corresponds to the number of BSSs to deploy. Finally, we determine the BSS location candidates according to the clustering results meanwhile considering system costs. The system cost depends on the number of BSSs to deploy and the construction cost per station. Since BSSs with more batteries stored will occupy larger urban land, which will increase the expense of building BSSs, the capacity \(N_i\) of BSS cannot be enlarged to infinity. Applying clustering algorithms, we can determine the amount of electricity demand in each cluster which refers to the battery capacity of the corresponding BSS. Therefore, through the proposed framework shown in Fig. 1, we find a solution to the optimization problem formulated in (3)-(5).

![Diagram](image_url)

**Fig. 1.** Framework of our proposed solution

**IV. METHODOLOGY**

**A. Introduction of HMM**

In the real world, states and observations are not always one-to-one correspondence. Thus, many casual relationships and inference in real life cannot be modeled by Markov model. Take the famous coin tossing experiment as an example, if there are multiple coins behind a curtain, we cannot tell which coin the experimenter has flipped when we observe head or tail. However, this can be modeled by an HMM, which defines the unobservable states as hidden states. The variation of those hidden states is modeled as a stochastic process. Therefore, HMM is a doubly embedded stochastic process [20] with underlying stochastic hidden states producing stochastic observations. HMM is characterized by the number of hidden states \(N\), the number of observations per hidden state \(M\), the state transition probability distribution \(A\), the observation symbol probability distribution in each hidden state \(B\) and the initial hidden state distribution \(\pi\). Therefore, an HMM \(H\) can be represented by \(H = (N, M, A, B, \pi)\).

**B. HMM based Map Matching**

Due to GPS measurement noises, mistakes, shifting requirement according to local policy and regulation for commercial use and other factors [21], sampled GPS data does not always locate on the actual road. Besides, GPS cannot play a role when taxies are driven into residential communities since GPS signal is degraded due to the loss caused by dense buildings. Thus, map matching is used to map the sampled GPS data of taxies to the road network to infer the real trajectories of taxies [22]. However, the simplest mapping approach that maps the sampled GPS data to the nearest road segment has been proved to cause great discrepancy between the map matching results and the real trajectories [23]. Hence, we apply a HMM based map matching algorithm to find the paths that taxies have passed by, and then recover the timestamp of the points included in those paths. First, we give the definitions about path and trajectory.

**Definition 1:** Path consists of a list of sequential locations without timestamp, denoted as \(P_i = (l_{i1}^p, l_{i2}^p, \ldots, l_{in}^p, t_1, t_2, \ldots, t_n)\), where \(l_{in}^p\) and \(t_{in}\) are the longitude and latitude of the \(n\)th location, respectively.

**Definition 2:** Trajectory consists of a list of sequential locations with timestamp, denoted as \(T_i = (l_{i1}^t, l_{i2}^t, \ldots, l_{in}^t, t_1, t_2, \ldots, t_n)\), where \(l_{in}\) is the time when the taxi locates at the location \(l_{in}^t\).

The process of map matching is divided into two steps. The first step is to figure out the paths that taxies actually were driven on, and in the second step we recover the time of locations in pathes and obtain trajectories. Fig. 2 shows the principle of an HMM based method for finding the path. To figure out the path for each EV, we construct an HMM for each EV. In the HMM, the sampled GPS locations of taxies are the observation states, and the road segments in the road network are hidden states. We apply HMM to infer which road segments taxies are actually on at different time based on the observation states and the reasonability of the path, corresponding to two probabilities. The first probability is the likelihood that the observed GPS location of taxi indicates which road segment the taxi is actually on. The second probability is the transition probability between two consecutive sampling time considering the reasonability of the path. We will introduce the details about how to calculate these two probabilities and how to infer pathes of taxies based on these two probabilities.

![Diagram](image_url)

**Fig. 2.** HMM based map matching method

Sampled GPS location deviates from the actual road segment because of various factors. In [24], the authors listed...
some references where GPS measurement errors follow either zero-mean Gaussian distribution or not zero-mean Gaussian distribution. In this paper, for simplify, we only consider the GPS measurement errors follow a zero-mean Gaussian distribution. In this paper, for simplify, we only consider some references where GPS measurement errors follow either zero-mean Gaussian distribution or not zero-mean Gaussian distribution. In this paper, for simplify, we only consider some references where GPS measurement errors follow either zero-mean Gaussian distribution or not zero-mean Gaussian distribution.

Thus, the first probability, also called emission probability, represents the likelihood that the GPS location \( O_l \) would be observed if the taxi was on road segment \( H_k \). Considering a single sampled GPS location, the higher emission probability is, the taxi is most likely to drive on the corresponding road segment. The probability is calculated for each road segment candidate as following.

\[
P(O_l | H_k) = \frac{1}{\sqrt{2\pi} \delta} \exp\left(-\frac{||O_l - p_l^{k}||^2}{2\delta^2}\right),
\]

where \( \delta \) is the standard variance of the GPS measurement error, \( p_l^{k} \) is the closest point to \( O_l \) on the candidate road segment \( H_k \) regarded as the match point, and \( ||O_l - p_l^{k}|| \) is the spherical distance between point \( O_l \) and \( p_l^{k} \). There is an intuition assumption that the probability that the sampled GPS location deviates from a far road segment is smaller than it deviates from a near road segment. Thus, to reduce the computational complexity, we set a distance threshold \( T_d \) to filter out road segment \( H_k \) where \( ||O_l - p_l^{k}|| > T_d \). After filtering, the number of road segment candidates decreases and we can calculate less emission probabilities to reduce the run time.

Then we utilize the correlation between two consecutive samplings to jointly infer the actual road segments and driving path between them as the GPS locations of taxies are sampled continuously. It is more likely that the length of the actual path between two projection points on the candidate road segments, such as \( ||p_l^{k} - p_{l+1}^{k}|| \) or \( ||p_l^{k} - p_{l+1}^{k+1}|| \) is more close to the spherical distance between \( O_l \) and \( O_{l+1} \). The probability density function of the difference between these two distances \( E_d \) fits the exponential distribution well [26], that is

\[
P(E_d) = \frac{1}{\sigma} e^{-\frac{E_d}{\sigma}},
\]

where \( E_d = \text{abs}(||p_l^{k} - p_{l+1}^{k}|| - ||O_l - O_{l+1}||) \), \( \text{abs}() \) is the absolute value function, \( k \) and \( k' \) are the candidate road segments, and \( \sigma \) is the parameter for the exponential distribution. Fig. 3 sketches the calculation process of the transition probability.

From the above definitions, to determine the paths of taxies, we maximize the product of \( P(O_l | H_k) \) and \( P(E_d) \) and apply Viterbi algorithm [27] to find the shortest path travers all observations in the HMM model shown in Fig. 2 which is the solution in each HMM [28]. Viterbi algorithm is a global optimal method to find the actual path through HMM, rather than the simplest mapping approach which always replaces the observation with the nearest coordinates and is usually trapped in local optimal solution.

Due to the high sampling frequency of GPS data, the information provided by the sampled GPS data is redundant. Each point contained in the pathes of taxies is projected from at least one sampled GPS locations. So far, we obtain the whole path of each taxi. However, a trajectory is not only the path but also contains the information when a taxi occurs at each point. Through the first step in the map matching process, we only know the corresponding time of the sampled GPS locations that just locate on the road networks and the mapping relationships between multiple sampled GPS locations and corresponding points contained in pathes. In the second step, we aim to recover the timestamp of all locations contained in a path. For each point \( (l^n_1, t^n_1) \) in the path, we reserve a list of sampled GPS locations that are mapped into the point in the path with their sampling time, denoting as \( L^n = [(O^n_1, t^n_1), (O^n_2, t^n_2), \ldots, (O^n_{C_n}, t^n_{C_n})] \), where \( C_n \) sampled GPS locations are mapped into the point \( (l^n_1, t^n_1) \), and \( t^n_{C_n} \) is the time that the \( C_n \)th sampled GPS location occurs. So for each path \( P_a \), we have a set of lists \( L = \{ L_1, L_2, \ldots, L_n \} \).

Fig. 4 illustrates the process of time recovery in the second step. A three dimensional space is constructed in Fig. 4 formed by latitude, longitude and time. Green triangles stand for the sampled GPS locations, striated circles stand for points without timestamp in pathes, and solid red circles represent points in trajectories with recovered timestamp. Green triangles and striated circle in the same ellipse exhibit the mapping relationship between the sampled GPS locations and the corresponding point in the path. To determine the timestamp of points in pathes, we design Algorithm 1. The Trajectory Time Recovery
(TTR) algorithm is executed once for each path. For each point \( (l^p_g, l^p_a) \) in the path, we first find the nearest sampled GPS location \( O_k \) in the list \( L \) (Line 2) and then update the list \( L \) of the next point in the path because the occurrence time of point \( (l^p_g, l^p_a) \) can only be later than that of point \( (l^p_g, l^p_a) \) (Line 3). If there is not any element in the list \( L \) updated or the timestamp of all points in the path has been recovered, the time recovery process is finished (Line 4-9). The computational complexity of Algorithm 1 is \( O(\sum L_i \times C_n) \).

\[
\text{Algorithm 1. Trajectory Time Recovery (TTR) Algorithm} \\
\text{Initialization:} \\
P_a; L; \\
1: \text{for} \ (l^p_g, l^p_a) \in P_a \ \text{do} \\
2: \quad (l^p_g, l^p_a) = \arg\min_{O_k} ||O_k - (l^p_g, l^p_a)||; \\
3: \quad \text{Update} \ \hat{L}^{n+1} = L^{n+1} \setminus \{O_k^{n+1}, t^{n+1}_a\}, \text{where} \ t^{n+1}_a \leq t^{n+1}_l; \\
4: \quad \text{if} \ \hat{L}^{n+1} = \emptyset \ \text{then} \\
5: \quad \quad \text{Delete} \ (l^{i=n+1.a,n+2,\ldots}, l^{i=n+1,a,n+2,\ldots}); \\
6: \quad \text{else} \\
7: \quad \quad \text{continue}; \\
8: \quad \text{end if} \\
9: \text{end for} \\
\text{Output:} \\
T_f = (l^p_g, l^p_a, t^p_l) \Rightarrow (l^p_g, l^p_a, t^p_l) \Rightarrow \ldots \Rightarrow (l^p_g, l^p_a, t^p_l)
\]

C. Electricity Demand Estimation

To deploy battery swapping stations, we have to investigate the electricity demand of all EVs in an urban area in a whole day. EV users are assumed to request battery swapping when the SoC of their batteries is below a threshold \( T_h \). The SoC of EV batteries decreases along with the traveling. Thus, estimating when an EV will request for battery swapping service is to determine when the SoC of this EV falls below \( T_h \). Some previous works have focused on modeling the Electricity Consumption Rate (ECR) of EV batteries. [29] declared that the most relevant factors that affect the ECR of EVs are the velocity and accelerated velocity. The authors in [30] proposed the model which is validated by the data collected when EVs are running in different modes through the road networks in a medium-sized city in China, such as starting, acceleration, uniform motion and deceleration. Thus, we apply the ECR model in this paper, represented by:

\[
E_{cr} = \begin{cases} 
3 \sum_{i=0}^{3} \sum_{j=0}^{3} w_{i,j} \times v^i \times a^j & a > 0 \\
\frac{v^1}{c_1} + \frac{v^0}{c_0} & a = 0, v > 0 \\
\exp(\sum_{i=0}^{3} \sum_{j=0}^{3} m_{i,j} \times v^i \times a^j) & a < 0. 
\end{cases} \\
(8)
\]

where \( w_{i,j}, c_0, c_1 \) and \( m_{i,j} \) are parameters, and their values can be found in [30]. The unit of \( E_{cr} \) is Joules per second.

To capture the electricity consumption of each EV based on the above ECR model, we need to obtain the velocity and accelerated velocity of each EV during the traveling. As we have extracted the trajectory of each EV in the step of map matching, we calculate a series of average velocity and accelerated velocity pair for EVs based on their trajectories. Then, for each average velocity and accelerated velocity pair, \( E_{cr} \) is calculated according to (8). The average velocity and accelerated velocity can be calculated as following:

\[
\bar{v}_k = \frac{|| (l^k_g, l^k_a) - (l^{k+1}_g, l^{k+1}_a) ||}{t^{k+1}_l - t^k_l}, \\
(9)
\]

where \( | (l^k_g, l^k_a) - (l^{k+1}_g, l^{k+1}_a) | \) represents the length of path between point \( (l^k_g, l^k_a) \) and point \( (l^{k+1}_g, l^{k+1}_a) \) in the trajectory. Thus, the SoC of an EV can be calculated as:

\[
S_{oc} = C - \sum_{k=1}^{n-1} E_{cr} \times (t^{k+1}_l - t^k_l), \\
(11)
\]

where \( E_{cr} \) is the function of \( \bar{v}_k \) and \( \bar{a}_k \) according to (8), and \( C \) is the capacity of EV battery. When \( S_{oc} \leq T_h \), the EV will request for battery swapping service, and the battery swapping demand is represented by \( (l^d_g, l^d_a, t_d) \), where \( (l^d_g, l^d_a) \) and \( t_d \) denotes where and when the battery swapping demand is generated.

D. BSS Location Determination

The objective of deploying BSSs is to satisfy the battery swapping demand of EVs as much as possible. Thus, we apply clustering algorithms to characterize the spatial distribution of battery swapping demands. Then, we select the location candidates for BSSs based on the spatial distribution of battery swapping demand.

The first proposed BSS location selection scheme shown in Algorithm 2 is based on the K-Means clustering algorithm. We randomly choose \( K \) locations where battery swapping demands are generated as \( K \) initial centroids. Other demand locations are partitioned into the cluster which the centroid is the nearest one, and then update the centroid. Iterate the process until all the centroids are not changed.

\[
\text{Algorithm 2. K-Means Clustering based BSS Location Selection Algorithm} \\
\text{Initialization:} \\
(l^d_g, l^d_a, t_d); \\
1: \text{Randomly choose} \ K \ ((l^d_g(k), l^d_a(k)), k = \{1, \ldots, K\} \text{ as initial centroids}; \\
2: \text{repeat} \\
3: \quad \text{Find} \ \arg\min_{k} ||(l^d_g(k), l^d_a(k)) - (l^d_g, l^d_a)||, \text{allocate} \ (l^d_g, l^d_a) \ \text{to Cluster} \ k; \\
4: \quad \text{Update} \ (l^d_g(k), l^d_a(k)), k = \{1, \ldots, K\}; \\
5: \text{until} \ (l^d_g(k), l^d_a(k)), k = \{1, \ldots, K\} \text{ do not change}; \\
\text{Output:} \\
(l^d_g(k), l^d_a(k)), k = \{1, \ldots, K\}, \text{ amount of demand in Cluster} \ k
\]

The time complexity of Algorithm 2 is \( O(I \times N_d \times K) \), where \( I \) is the number of iterations and \( N_d \) is the amount
Algorithm 3: Hierarchical Clustering based BSS Location Selection Algorithm

Initialization:

\((l^d_1, l^d_2, t_d)\):
1: Each location is initialized as a cluster;
2: repeat
3: Merge Cluster \(k\) and \(k'\) which is \(\arg \min \|k - k'\|_{\text{proximity}}, k, k' \in \{1, \ldots, K_n\}\);
4: Update \(\|k - k'\|_{\text{proximity}}, k, k' \in \{1, \ldots, K_n\}\);
5: Update \(K_n\):
6: until \(K_n = 1\);
7: Choose the number of clusters \(K\)

Output:

\((l^d_1(k), l^d_2(k)), k = \{1, \ldots, K\}\), amount of demand in Cluster \(k\)

V. PERFORMANCE EVALUATION

In this section, we elaborate the dataset used in our simulation, investigate the effect of the precision parameters in map matching, show the visualization results of the proposed BSS location selection schemes and compare the performance of the proposed schemes with baseline scheme.

A. Dataset Description

The dataset utilized in this paper is an anonymous taxi location dataset measured by GPS equipment installed in each taxi. The dataset is collected from April 1, 2015 to April 30, 2015 in Shanghai, China, sampled every 10 seconds. There are usually two public holidays on April. One is the Tomb-sweeping Day in early April, and another one is the Labor Day at the end of April. Thus, we only utilize the taxi data collected from April 13 to 26 lasting for two weeks, since generally the travel demand of people during public holidays is different from that in regular life. 13,700 taxies are involved in our dataset, about 23.6% of taxies in Shanghai according to the statistic on Shanghai’s taxi data. Each entry in the dataset contains the anonymized taxi ID, timestamp, longitude and latitude. There are totally 1,603,578,412 entries in our dataset, about 114.5 million entries in each day. Sampled GPS data of taxies from April 13 to 19 is used as the training set to determine the BSS location candidate selection, and data from April 20 to 26 is the testing set used to evaluate the effectiveness of our selection schemes of the BSS locations.

In addition, we acquire the road network of Shanghai through OpenStreetMap [32] which is represented by nodes, ways and relations. The road network of Shanghai consists of 366,431 nodes and 52,714 ways. Nodes are intersections, dead ends, road name changes and road bend points represented by longitude and latitude pairs. The road bend points are different from intersections as they exist because the ways between the starting point and the end point of road segments sometimes are not straight. A way is labeled by the way index, and consists of sequential node list on the way. Relations describe the relationships among nodes and relationships among ways, such as which nodes are in the same way, which ways cross, and so on. The longitude of the road network in Shanghai varies from 121.34E to 121.66E, while the latitude varies from 31.12N to 31.38N.

B. Schemes to be Compared

The baseline algorithm of BSS selection we choose to compare is the uniform deployment scheme. In the uniform deployment scheme, the bounding box of Shanghai map is equally divided into \(6 \times 8\) grids. In each grid, one BSS will be installed and the battery capacity of the BSS is determined due to the average battery swapping demands generated in the coverage of the grid during one day. For fair comparison, \(K\) in K-Means clustering algorithm is also set to be 48. If the bounding box of investigated area is divided into \(n \times m\) grids, \(K\) can be chosen as \(n \times m\), which will not affect the comparison result. The only thing needs to be noticed is that the number of clusters in these three clustering algorithms should be the same when comparing the clustering results.

We also compare the performance of the proposed map matching scheme with the simplest mapping approach that maps the sampled GPS data to the nearest road segment.
This can be explained that EVs travel farther on weekend than on weekdays. For example, the temporal distribution of battery swapping demand on weekend is different from that on weekday as shown in Fig. 7. The demand in the morning of weekend is more because EV users choose to travel during the morning for different reasons, such as business or leisure. Also, the journey at the beginning of the day, EVs generate a little swapping demand in the morning, for example, about 100 battery swapping demand is generated along the way to the airports. This can be explained that EVs travel farther on weekend.

C. Result Analysis

1) Performance of Map Matching

We compare the performance of the proposed mapping method to the simplest mapping which referred to mapping observed GPS location to the nearest road segment. The performance is evaluated in term of length of all trajectories, number of mapped points, average distance error and distance error per meter. We also investigate the effect of distance threshold $T_d$ on the results of map matching. Based on the matching results, we optimally choose the value of distance threshold $T_d$. Despite of applying our proposed method or the simplest mapping, Fig. 5(a) shows that with the increase of $T_d$, the number of mapped points, i.e., the number of all locations contained in all trajectories after map matching decreases. This is because less sampled GPS locations will be filtered out in the map matching process due to a smaller $T_d$. Thus, longer trajectories are inferred as shown in Fig. 5(b). On the other hand, since more sampled GPS locations are involved in the map matching, it is possible that more sampled GPS locations are mapped to wrong road segments. This may result in longer trajectories due to extra trails such as back and forth. This is another important factor that causes the trajectories of all EVs in a day to be long when $T_d$ is small. At the same time, the derivation from trajectories consisting of raw sampled GPS locations becomes large. Thus, another two metrics such as the average distance error, i.e., the total distance error divided by the number of mapped points and the distance error per meter, i.e., the distance error generated with one meter trajectory mapped first decrease. However, when $T_d$ is large, too much information contained in raw sampled GPS data is lost, hence, the inferred trajectories become short. Nevertheless, the actual trajectories can be longer since there may exist driving behaviors such as turning around or changing direction in trajectories. Hence, both the average distance error and the distance error per meter increase when $T_d$ becomes large. Fig. 5(c) and (d) show the variation of the average distance error and the distance error per meter with threshold $T_d$. From Fig. 5(c) and (d), we can find that the average distance error and distance error per meter is the lowest when $T_d$ is set to be 100 meters. Meanwhile, not too much information contained in raw data is lost according to Fig. 5(a) and (b) when $T_d$ is set to be 100 meters. Therefore, the optimal value of the threshold $T_d$ is set to be 100 meters in our experimental scenario to minimize the average distance error. Furthermore, from Fig. 5(c), we can observe that by applying the proposed mapping method, the average distance error can be reduced to 10.07 meters when threshold is set to be 100 meters, which increases 23.2% accuracy compared to the simplest mapping method. Fig. 5(d) also shows the distance error per meter is reduced 25.4% from 0.114 to 0.085 compared to the simplest mapping method.

![Fig. 5. Effect of distance threshold $T_d$ on the results of map matching](image)

2) Spatial and Temporal Distribution of Battery Swapping Demand

We visualize the battery swapping demand on April 13 to show the spatial distribution of demand as shown in Fig. 6 in which the background map is obtained through the Application Program Interface (API) offered by BaiduMap [33]. It is shown that there is much battery swapping demand in the downtown of Shanghai. We also observe that the battery swapping demand in the West area of Shanghai is larger than that in the East area. Two airports locating in the suburb of Shanghai is far from the urban resident area, thus some battery swapping demand is generated along the way to the airports.

![Fig. 6. Heatmap of battery swapping demand on April 13](image)

3) Performance Evaluation of the Proposed BSS Location Selection Schemes

The spatial distribution of battery swapping demand varies in different time of a day as shown in Fig. 8. Since it is assumed that each EV is fully charged when it starts the journey at the beginning of the day, EVs generate a little battery swapping demand in the morning, for example, about 100 battery swapping demand is generated from 11:00 to 12:00. The amount of battery swapping demand is up to 500 in the afternoon from 15:00 to 16:00, and keeps increasing and reaches the highest intensity between 21:00 and 22:00. The temporal distribution of battery swapping demand on weekend is different from that on weekday as shown in Fig. 7. The demand in the morning of weekend is more because EV users are prone to start a journey in the early morning of weekend. In addition, the peak hour of the demand on weekend is from 20:00 to 21:00 which is an hour earlier than that on weekday. This can be explained that EVs travel farther on weekend.

![Fig. 8. Temporal distribution of battery swapping demand](image)
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2941901, IEEE Access

K-Means clustering algorithm is highly influenced by the selection of initial cluster centroids. Thus, to obtain better clustering result, we initialize charging demand as the cluster centroid as disperse as possible. In addition, we can observe that not all 48 BSSs are in the scope of Fig. 9(c) because part of the BSSs deployed under the uniform deployment scheme are in the suburb area of Shanghai. However, as Fig. 6 shows the battery swapping demand in the suburb area is little. The deployment of public BSSs in suburb area will gain low efficiency in facility utilization. Compared to the uniform deployment scheme, K-Means based scheme and hierarchical clustering based scheme are more cost-effective.

We further apply the metric of Mean Distance Error (MDE) to evaluate the performance of these three schemes. MDE is defined as the average distance from the nearest BSS when a battery swapping demand is generated. If MDE is large, it will take EV users more time to reach the BSSs and which will definitely increase the range anxiety of EV users. Fig. 10 shows the MDE is reduced to around 2.2km and 2.8km by applying the proposed K-Means and hierarchical clustering based BSS location selection scheme, respectively, comparing to 5.9km when applying the uniform deployment scheme. Moreover, testing the three schemes on the battery swapping demand inferred from GPS data from April 20 to 26, the MDE under each scheme on different days is consistent, which shows that the spatial distribution of battery swapping demand does not vary on different days and our proposed schemes are robust.

VI. CONCLUSIONS

In this paper, we propose a data-driven framework for solving the Battery Swapping Station (BSS) location selection problem in the green and intelligent transportation system. To deploy BSSs in a cost-effective way meanwhile satisfy the battery swapping demand of EVs in metropolitan area, massive GPS data of taxis and electricity consumption rate model of EV battery are utilized to mine the spatial and temporal distribution of the battery swapping demand of EVs. Raw GPS dataset is first mapped onto road networks based on Hidden Markov Model (HMM), and trajectories of EVs are extracted along with the map matching. Then, battery swapping demand is estimated based on the electricity consumption rate model. Finally, we propose the location selection schemes of BSSs based on the observations on the spatial distribution of battery swapping demand. Both the distribution of battery swapping demand and BSS locations are well visualized. Our proposed BSS location selection schemes have great advantages compared to the baseline. The implementation of our proposed schemes are more easily and the Mean Distance Error (MDE) between the location that battery swapping demand is generated and the nearest BSS is shorter which can definitely reduce the range anxiety of EV users. Our BSS deployment schemes gives useful guidelines to the planning department of the government.

REFERENCES


Fig. 8. Spatial distribution of battery swapping demand in different time on weekday

(a) 11:00-12:00
(b) 16:00-17:00
(c) 17:00-18:00
(d) 21:00-22:00
(e) 22:00-23:00
(f) 23:00-24:00

Fig. 9. Deployment of BSSs under different schemes

(a) K-Means clustering based scheme
(b) Hierarchical clustering based scheme
(c) Uniform deployment scheme

Fig. 10. MDE under the proposed two schemes and the uniform deployment scheme


