TRB Annual Meeting

A Two-Stage Demand Prediction and Optimization Modeling Framework for Charging Station Location Problem

--Manuscript Draft--

Full Title:	A Two-Stage Demand Prediction and Optimization Modeling Framework for Charging Station Location Problem
Abstract:	The optimal planning of charging infrastructure, termed charging station location problem (CSLP), aims to satisfy the EV charging demand to the maximum extent under the limitation of the investment budget. While abundant studies on CSLP exist, most of them do not explicitly model the charging demand of an EV charging station (CS). As a result, explicit demand measurements, such as the predicted number of charging events per day, although frequently sought for by the practitioners, are unavailable. Another common issue is they usually assume a known EV OD demand matrix and mobility patterns of EV drivers, which is, however, not readily available. Last but not the least, the measurement of induced demand in response to new charging infrastructure is not well understood. To fill these research gaps, a two-stage CSLP model is proposed. In stage one, a gradient boost-based model is developed to generate accurate predicted charging demand, which is explicitly measured as the charging usage rate per week. Next, in the stage two model, a demand-supply coupled CS location selection model is developed with the objective of maximizing the total charging demand of both existing and newly selected CSs. The proposed CSLP model is solved with a greedy-based stochastic spatial search algorithm. A case study using multi-source real-world data from Kansas City Missouri is performed to test the effectiveness of the proposed model. Results show that the stage one model generates a satisfactory charging demand prediction result, and the stage two model is demonstrated to outperform two benchmark models
Manuscript Classifications:	Data and Data Science; Information Systems and Technology AED30; Data and Technology Services Related to CAEV (Connected, Automated, and Electric Vehicles); Sustainability and Resilience; Transportation and Sustainability; Alternative Transportation Fuels and Technologies AMS40; Electric and Hybrid-Electric Vehicles
Manuscript Number:	TRBAM-23-04248
Article Type:	Presentation
Order of Authors:	Yang Song
	Qing Tang
	Yiyang Wang
	Xianbiao Hu
Additional Information:	
Question	**Response**
The total word count limit is 7500 words including tables. Each table equals 250 words and must be included in your count. Papers exceeding the word limit may be rejected. My word count is:	7453
Is your submission in response to a Call for Papers? (This is not required and will not affect your likelihood of publication.)	No

A Two-Stage Demand Prediction and Optimization Modeling Framework for Charging Station Location Problem

Yang Song
Department of Civil and Environmental Engineering
The Pennsylvania State University, University Park, PA, 16802-1408
Email: ysong@psu.edu

Qing Tang
Department of Civil and Environmental Engineering
The Pennsylvania State University, University Park, PA, 16802-1408
Email: qingtang@psu.edu

Yiyang Wang
Department of Industrial and Manufacturing Engineering
The Pennsylvania State University, University Park, PA, 16802-4400
Email: ykw5222@psu.edu

Xianbiao Hu, Corresponding Author
Department of Civil and Environmental Engineering
The Pennsylvania State University, University Park, PA, 16802-1408
Email: xbhu@psu.edu

Word Count: 7,203 words + 1 table (250 words per table) = 7,453 words

Submitted for Presentation at the 102nd Annual Meeting of Transportation Research Board

Submitted [08/01/2022]

ABSTRACT

The optimal planning of charging infrastructure, termed charging station location problem (CSLP), aims to satisfy the EV charging demand to the maximum extent under the limitation of the investment budget. While abundant studies on CSLP exist, most of them do not explicitly model the charging demand of an EV charging station (CS). As a result, explicit demand measurements, such as the predicted number of charging events per day, although frequently sought for by the practitioners, are unavailable. Another common issue is they usually assume a known EV OD demand matrix and mobility patterns of EV drivers, which is, however, not readily available. Last but not the least, the measurement of induced demand in response to new charging infrastructure is not well understood. To fill these research gaps, a two-stage CSLP model is proposed. In stage one, a gradient boost-based model is developed to generate accurate predicted charging demand, which is explicitly measured as the charging usage rate per week. Next, in the stage two model, a demand-supply coupled CS location selection model is developed with the objective of maximizing the total charging demand of both existing and newly selected CSs. The proposed CSLP model is solved with a greedy-based stochastic spatial search algorithm. A case study using multi-source real-world data from Kansas City Missouri is performed to test the effectiveness of the proposed model. Results show that the stage one model generates a satisfactory charging demand prediction result, with R square values reaching 0.72 and 0.63 on the training and testing datasets, respectively. The stage two model is demonstrated to improve charging usage by 14%, and outperform the results of two benchmark models, namely naïve greedy selection strategy and naïve neighbor swap strategy.

Keywords: electric vehicle; charging station location problem; charging demand; demand-supply interactions; two-stage model

1. INTRODUCTION

Global climate change has encouraged government agencies to adopt more environmental-friendly policies regarding sustainable and cleaner energy consumption. As transportation activities account for over 25% of total greenhouse gas (GHG) emissions, and internal combustion engine vehicles (ICEVs) are responsible for around 75% of transport-related emissions [1], it is essential to replace ICEVs with Electric vehicles (EVs) that are propelled by electric motors without any carbon-based emission. However, due to the current technical limitations of short driving range and long recharging time [2], the market penetration of EVs is still relatively low [3]. Thus, it is important to increase recharging infrastructures to satisfy the growing charging demand of EV users.

The optimal planning of charging infrastructure, termed charging station location problem (CSLP), is commonly defined as the selection of optimal locations for charging stations (CSs) by service providers, so that the recharging demand of EV drivers could be satisfied to the maximum extent under the limitation of the investment budget [3]. Abundant studies on CSLP exist and as reviewed by Kchaou-Boujelben [3], can be categorized into flow-based and node-based approaches. Flow-based models [4-8] typically represent charging demand as a set of origin–destination (OD) trips where fast CSs have to be adequately placed so that distance between two consecutive stations does not exceed the driving range, allowing drivers to travel from their origin to their destination and back without running out of charge. This type of study, however, does not explicitly model the charging demand of an EV charging station, instead, they aim to optimize an aggregated system measurement, for example, to maximize the flow captured on the network, or minimize the system cost. As a result, explicit demand measurements, such as the predicted number of charging events per day, although frequently sought for by the practitioners, are unavailable. Another common issue is the flow-based models assume a known EV OD demand matrix and mobility patterns of EV drivers, for the purpose of developing and calibrating the traffic assignment model. However, such information of EV drivers is usually not available, creating barriers to real-world applications of flow-based models.

Node-based models [9-14], as another popular approach, assume that charging demand does not depend on EV trips but simply arises at the nodes of a road network. Following such assumption, the node-based demand usually originates from a cluster of EV drivers who prefer to recharge their vehicles in a station located close to their home, workplace, or any service facility such as a shopping mall, and the charging demand is commonly represented by the size of CS service area. For this type of research, the objective is to maximize the area that nodes could cover where CSs are constructed, which is defined as the area within an arbitrary short walking distance or driving time from the CS location. Similar to the flow-based models, an explicit prediction of the charging demand of the CSs is still not available.

To overcome these issues, in this manuscript a new demand prediction and optimization modeling framework is proposed to determine the optimal location for CSs. Compared with the literature, the proposed modeling approach does not require a pre-determined EV OD demand matrix, but rather, learns from the field-collected dataset and explicitly predict the charging demand of each charging station. Further, for the charging station site selection problem, we model the demand-supply coupled interactions. Meaning, when a new candidate site is chosen, the new charging supply will inevitably lead to an updated charging demand profile for the existing CSs in the area. As such, instead of naively selecting the candidate sites with the highest predicted usage rates, how to quantify the demand-supply coupled interactions and develop CSLP solution with consideration of such interactions becomes an important question.

To this end, a two-stage CSLP model is proposed. In the stage one model, instead of using the assigned traffic flow or service coverage area to represent charging demand, we use the weekly charging usage rate as an explicit measurement. A gradient boost-based model is developed to generate charging demand predictions. Next, in the stage two model, a demand-supply coupled CS location selection model is developed with the objective of maximizing the total charging demand of both existing and newly selected CSs. The proposed model is solved with a greedy-based stochastic spatial search algorithm. To analyze the effectiveness of the proposed model, real-world data from Kansas City Missouri is used. We compared our modeling results with those from two benchmark models, with the first one simply selecting the candidate locations with the highest predicted charging demand (i.e., naïve greedy selection strategy), and the second model that updates selection only from neighboring areas but ignores demand-supply coupled interactions (i.e., naïve neighbor swap strategy). Results reveal that the stage one model is able to predict charging demand with R square values reaching 0.72 and 0.63 on training and testing sets, respectively, indicating satisfactory training performances. The stage two model is demonstrated to improve charging usage by 14%, and outperform the results of both benchmark models.

The remainder of this manuscript is organized as follows. A literature review is summarized in Section 2. Section 3 presents the formulation of the proposed two-stage model. The greedy-based spatial stochastic search algorithm to solve the optimization model is shown in Section 3. Section 4 introduces a real-world instance to test the performance of the proposed model. Conclusions are summarized in Section 5.

2. LITERATURE REVIEW

A significant number of studies on CSLP have been found in the literature, with charging demand modeling being a key element of research. As reviewed by Kchaou-Boujelben [3], according to the method of representing recharging demand, CSLP formulation can be mainly categorized into flow-based and node-based models *[3]*.

Flow-based models assumed that EV drivers need to recharge their vehicles during the trip from origin to destination if they traveled for long distances, especially when the mileage exceeds the limited driving range. In this case, the charging demand was represented by a set of origin-destination (OD) trips of EVs. Thus, the typical model setting was to optimize the CS location along these tips, with the objective of maximizing the number of EV drivers who can complete their journey from origin to destination and return without running out of battery. The common methods to describe charging demand coverage included flow capturing, flowing refueling, arc covering, path-segment, and battery state-of-charge (SOC) tracking [3]. Hodgson [4] firstly proposed the flow capturing location model, which assumed a trip was captured if at least one CS was placed along the path from origin to destination. Further, the flow refueling location model was introduced by taking EV driving range limitations into consideration [5]. To address the issue of the high computational cost of flow refueling location models, a new arc covering-based formulation approach was presented, with the assumption that a trip of EV was covered as long as every arc consisting of the trip was covered by at least one CS [6]. Unlike the arc covering method that decomposed trips into arcs, the path-segment-based coverage method divided each trip into a sequence of path segments that was composed of consecutive arcs [7]. If a trip was considered to be covered, the length of each path segment should be no more than the driving range of EVs, given that CSs were deployed at the starting node of each path segment. The battery SOC tracking approach was based on the idea of tracking the EV SOC value along

every visited node, to ensure the remaining charge was always positive, otherwise recharging was triggered [8]. Although details may vary, flow-based research did not explicitly model the charging demand of an EV charging station. Instead, the covered EV traffic flow was frequently used as an optimization target, although the correlation between the charging usage rate of a CS and the bypassing traffic volume has not been clearly demonstrated. In addition, they usually assumed a known EV OD demand matrix and mobility patterns of EV drivers, which is, however, not readily available in reality due to privacy concerns.

Unlike flow-based models that used EV trips to represent charging demand, node-based models, on the other hand, considered charging demand as a classical facility location modeling problem [9], i.e., charging demand generated at nodes in a planning area [3]. Such nodes may include many different types of points of interest (POIs), where a cluster of EV drivers prefer to recharge, such as home, workplace, or restaurants. In terms of model formulation, existing node-based studies adopted approaches including location allocation, set covering, and maximal covering methods. Similar to the idea of the p-median problem [10], a location-allocation approach aimed to minimize the total system cost with all charging demands satisfied. The system costs typically included travel costs, which can be measured by the distances between EV users and the assigned CS, and fixed construction costs [11]. The idea of the set covering approach was to cover every charging customer within the service area of CSs, which can be defined as the maximum driving time [12] or maximum walking distance [13]. Then, in some cases when not all EV users could be covered due to a limited budget, a maximal covering approach was adopted with the objective of maximizing the service coverage area of planned CSs [14]. However, existing node-based models typically used the size of the CS service area to represent the charging demand, and an explicit prediction of charging demand is not available.

Aside from charging demand representation, the measurement of induced demand in response to new charging infrastructure is not well understood. The involving parties included CS location planners like government agencies and private investigators, as well as EV drivers who were also charging customers. For example, Guo et al. [15] took the competitive interaction between charging service providers in the EV CS market into account. Under the condition that all private institutions aimed at maximizing their profit while satisfying the charging customers, the equilibrium of pricing could be achieved. Then, the interaction between EV drivers was introduced by Bernardo et al. [16], by assuming all charging customers wanted to maximize their utilities. In this way, EV drivers may deviate from the shortest path considering charging costs and amenities at CSs. Moreover, the interaction between planners and users was more particularly described in the form of bi-level flow-based models, where the decision process of EV drivers in the lower level was incorporated into the CS location optimization process of planners in the upper level [17, 18]. However, limited research considered the effect of charging facility supplies on charging demand. To be specific, the deployment of a new CS may increase or decrease the charging demand of its neighboring existing CSs. For example, if demand remains the same or only increases slightly, an increased number of charging stations may lead to lower average usage rates at each charging station, whereas in some other cases, the denser charging stations may make this area more attractive to EV drivers, or encourages higher EV ownership, as such the average usage rates at each charging station may increase. Modeling of such demand-supply interactions becomes important to the formulation of CSLP.

To summarize, the existing studies did not explicitly model EV charging demand and assumed the EV OD demand and mobility patterns are readily available. In addition, the demand-supply coupling relationship, although important to the formulation of CSLP, is ignored.

To address these research gaps, a demand-supply coupled CSLP model is proposed to determine the optimal location for CSs. The stage one model explicitly predicts the charging demand of each CS with an ensemble learning approach, whereas the stage two model selects the optimal CS location considering the demand-supply coupled interactions to maximize the charging demand of both existing CSs and selected charging locations.

3. PROBLEM FORMULATION

In this section, a two-stage model is developed to identify optimal locations to build new charging stations, so that the charging demand of both existing and planned CSs are maximized. The purpose of the stage one model is to predict charging demand with explicit measurement of weekly charging rate, and this is achieved by a boosting-based ensemble learning approach. Then, to consider the demand-supply coupled interactions, the stage two model is developed to determine the optimal CSs location to maximize the charging demand of both existing CSs and selected charging locations. Figure 1 illustrates the workflow of the proposed model.

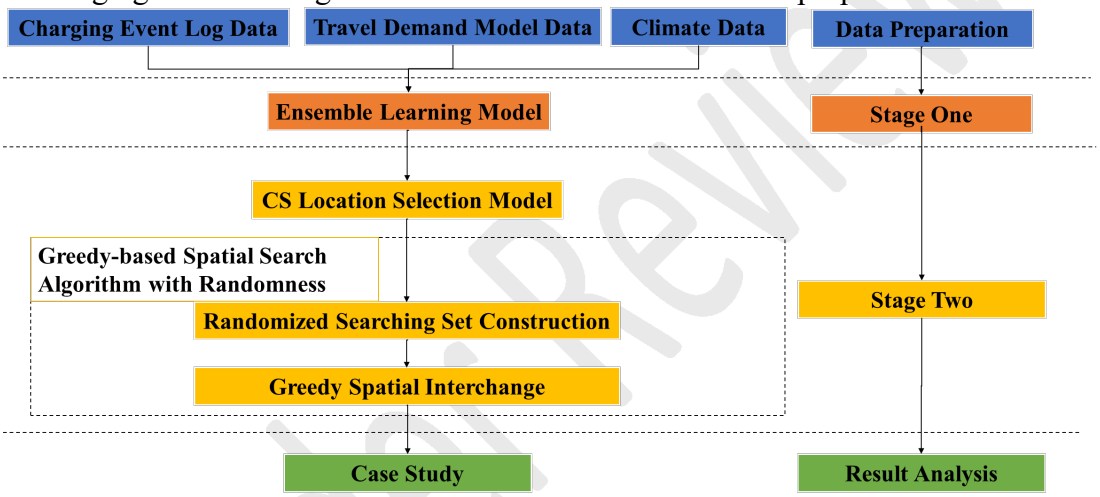


Figure 1 The workflow chart of the proposed model

As a convenient reference, the mathematical notations used in this section are presented below.

i,j: Index of the location of CS
w: Index of the week when charging events happen for one charging station
k: Index of the sub-regression trees
K: Total number of the sub regression trees in the ensemble learning model
L: The objective function of the ensemble learning model
Θ: The set of all regression trees f_k in the ensemble learning model
f_k: The kth sub-regression tree
l: A training loss component
$\widehat{D}_{i,w}$: The observed charging demand of site i at week w
$D_{i,w}$: The predicted charging demand of site i at week w
Ω: A regularization component
γ: A penalty parameter to control the complexity of the tree structure
T: The number of leaf nodes of a base tree
λ: A parameter to control the regularization degree of f_k
Wt: The weight of leaf nodes
ln: The index of leaf nodes

N_{ln}: Total number of leaf nodes
$p_{i,w}$ is an array with the independent features of site i at wth week
N_e: Number of the existing EV charging stations
N_c: Number of candidate EV site locations
N_d: Number of new EV charging stations to be deployed
$MAXN_z$: The max number of selected new charging stations in one zip code area
x_i: A binary decision variable, $x_i = 1$ when there exists or will be an EV charging station at location i, otherwise $x_i = 0$. The same applies with x_j.
$NHCS_i, PRCP_w, TMP_w, WD_w, PT1{\sim}2_i, LDU1{\sim}7_i, AADT_w, TP_i$: Features will be introduced in Table 1
$CSNUM_i$: The total number of charging stations in the zip code where location i is located
z_j^i: A binary decision variable, $z_j^i = 1$ if the zip code of CS location j and i are the same, otherwise $z_j^i = 0$.

3.1 Stage one ensemble learning model for explicit charging demand prediction

In this section, a demand prediction model is developed. Among different kinds of machine learning techniques, ensemble learning models have been demonstrated to improve prediction performance over traditional regression or classification models by training multiple sub-models and combining their results [19]. In particular, the LightGBM method, as a boosting-based ensemble learning algorithm, sequentially fits a new weaker regression tree on the prediction residual of the previous sub-trees [20]. In this way, the training loss can be continuously reduced when adding the new prediction result until convergence. In addition, when compared with other boosting models, LightGBM applies the second-order gradient statistics to optimize the objective with exact solution formulation.

The input data of this model is $O = \{(p_{1,1}, \widehat{D}_{1,1}), (p_{1,2}, \widehat{D}_{1,2}), \dots, (p_{i,w}, \widehat{D}_{i,w})\}$, where $p_{i,w}$ is an array of features shown as Eq.(6), including spatial context information, weather information, charging port type, and traffic information. A detailed definition of each feature will be given in the case study section. $\widehat{D}_{i,w}$ is the response variable, which is the weekly charging rate in our context. The target of this optimization process is to find optimal functions that minimize the prediction error between predictions and actual demand.

The formulation of the boosting-based charging demand prediction model is given from Eq. (1) to Eq. (6). The objective function is presented by Eq. (1), with the goal of minimizing the loss function with a regularization term. Therein, Θ is the set of a total number of K regression sub-trees, as shown in Eq. (2). l is the loss function in the form of squared error as shown in Eq. (3). By decreasing the difference between the predicted charging demand $D_{i,w}$ and actual value $\widehat{D}_{i,w}$ per week, the bias of the charging demand prediction model could be reduced. The regularization term Ω is computed by Eq. (4), where T is the total number of leaf nodes in the base tree f_k. Parameter γ is designed to avoid tree structure from becoming too complex; Wt is the weight of leaf nodes; the parameter λ controls the regularization level of f_k. The calculation of Wt will be discussed in the solution algorithm section. Then, the predicted charging demand $D_{i,w}$ can be calculated by Eq. (5), which is the summation of the output of all base regression trees given the input feature $p_{i,w}$.

$$Min\ L(\Theta) = \sum_{i=1}^{N_e} \sum_{w=1}^{n} l(D_{i,w}, \widehat{D}_{i,w}) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

where

$$\Theta = \{f_1, f_2, \dots f_K\} \tag{2}$$

$$l\left(D_{i,w}, \widehat{D}_{i,w}\right) = \left(D_{i,w} - \widehat{D}_{i,w}\right)^2 \tag{3}$$

$$\Omega(f_k) = \gamma * T + \frac{1}{2}\lambda * \|Wt\|^2 \tag{4}$$

$$D_{i,w}(p_{i,w}) = \sum_{k=1}^{K} f_k(p_{i,w}) \tag{5}$$

$$p_{i,w} = [NHCS_i, PRCP_w, TMP_w, WD_w, PT1{\sim}2_i, LDU1{\sim}7_i, AADT_w, TP_i] \tag{6}$$

With a sufficient number of sub-regression trees built, the charging demand model in stage one could converge to the optimal solution with the minimum objective function value. Then, this ensemble model could be integrated into the stage two model to generate an accurate charging demand value for each CS location. The optimization process will be discussed in the section of Solution algorithm.

3.2 Stage two demand-supply coupled model for site selection optimization

The objective of the stage two model is to choose the optimal candidate sites so that the total charging demand, represented by the weekly charging rate of both planned and existing charging stations, can be maximized. The model formulation can be expressed as follows.

$$Max\ Obj = \sum_{i=N_e+1}^{N_e+N_c} D_{i,w} * x_i + \sum_{i=1}^{N_e} D_{i,w} \tag{7}$$

$$s.t.\ D_{i,w}(p_{i,w}) = \sum_{k=1}^{K} f_k(p_{i,w}) \tag{8}$$

$$NHCS_i = CSNUM_i / ARSZ_i \tag{9}$$

$$CSNUM_i = \sum_{j=1}^{N_e+N_c} x_j * z_j^i \tag{10}$$

$$\sum_{j=N_e+1}^{N_e+N_c} x_j * z_j^i \le MAXN_z \tag{11}$$

$$z_j^i = \begin{cases} 1 & if\ i\ and\ j\ are\ in\ the\ same\ area \\ 0 & Otherwise \end{cases} \tag{12}$$

$$\sum_{i=N_e+1}^{N_e+N_c} x_i = N_d \tag{13}$$

$$x_i = 1, i \in \{1,2, \dots, N_e\} \tag{14}$$

$$x_i = \begin{cases} 1 & if\ \ i\ is\ selected \\ 0 & Otherwise \end{cases}, i \in \{N_e + 1, N_e + 2, \dots, N_e + N_c\} \tag{15}$$

Eq. (7) shows that there are two terms in the objective function. The first component is the summation of the charging demand of the newly planned CSs, while the second component is the summation of the charging demand of the existing CSs. Eq. (8) describes the predicted charging demand of the CS location i from the stage one model. Eq. (9) calculates the density of CSs in the neighborhood area of the CS candidate site i, which is an input feature of Eq. (8). Eq. (10) calculates the number of CSs in the same area of the CS candidate site i. Eq. (11) is a

constraint that the number of new CSs, deployed in the same area where the CS location i is located, should not be greater than $MAXN_z$. Eq. (12) gives a binary variable that indicates the spatial relationship of two candidate sites, i and j. In Eq. (13), the constraint of the total number of newly planned CSs is presented. Eq. (14) and Eq. (15) define the decision variable of the stage two model, which is a binary, and equals 1 when there is an existing CS or newly planned CS at location i, otherwise 0.

4. SOLUTION ALGORITHM

In this section, the solution methods of both stage one and stage two models are presented. Therein, the stage one model is solved by an exact approach with explicit optima. The stage two model is solved with a proposed stochastic spatial searching algorithm.

4.1 Solution algorithm for stage one ensemble learning model

Combining Eqs. (3) and (5), and with some reorganization, the objective function in Eq. (1) can be re-formulated as Eq. (16).

$$\begin{aligned} &Min\ L(\Theta) = \sum_{i=1}^{N_e}\sum_{w=1}^{n}\left(f_k(p_{i,w})^2 + 2\left(\sum_{k=1}^{K-1} f_k(p_{i,w}) - \widehat{D}_{i,w}\right) * f_k(p_{i,w})\right) + \Omega(f_K) + \\ &\sum_{i=1}^{N_e}\sum_{w=1}^{n}\widehat{D}^2{}_{i,w} + \sum_{k=1}^{K-1}\Omega(f_k) \cong \sum_{i=1}^{N_e}\sum_{w=1}^{n}\left(f_k(p_{i,w})^2 + 2\left(\sum_{k=1}^{K-1} f_k(p_{i,w}) - \widehat{D}_{i,w}\right) * \right. \\ &\left. f_k(p_{i,w})\right) + \Omega(f_K) \end{aligned} \tag{16}$$

In Eq. (16), the two terms $\sum_{i=1}^{N_e}\sum_{w=1}^{n}\widehat{D}^2{}_{i,w}$ and $\sum_{k=1}^{K-1}\Omega(f_k)$ are skipped as they are constant and thus do not contribute to the problem solution. Then considering Eq. (4), Eq. (16) can be reformulated by leaf nodes as Eq. (17).

$$\begin{aligned} &Min\ L(\Theta) = \sum_{ln=1}^{N_{ln}}\left((Wt_{ln})^2 + 2\left(\sum_{k=1}^{K-1} f_k(p_{i,w}) - \widehat{D}_{i,w}\right) * Wt_{ln}\right) + \gamma * T + \frac{1}{2}\lambda * \\ &\sum_{ln=1}^{N_{ln}}(Wt_{ln})^2 = \sum_{ln=1}^{N_{ln}}\left((1+\frac{1}{2}\lambda)(Wt_{ln})^2 + 2\left(\sum_{k=1}^{K-1} f_k(p_{i,w}) - \widehat{D}_{i,w}\right) * Wt_{ln}\right) \end{aligned} \tag{17}$$

In Eq. (17), since the objective function is in the form of the quadratic function of Wt_{ln}, the optimal solution can be obtained with the first derivative $\partial L(\Theta)/\partial Wt_{ln} = 0$, as shown in Eq. (18).

$$\frac{\partial L(\Theta)}{\partial Wt_{ln}} = \sum_{ln=1}^{N_{ln}}\left((2+\lambda) * Wt_{ln} + 2\left(\sum_{k=1}^{K-1} f_k(p_{i,w}) - \widehat{D}_{i,w}\right)\right) = 0 \tag{18}$$

Thus, when $Wt_{ln}^* = \frac{2\left(\sum_{k=1}^{K-1} f_k(p_w) - \widehat{D}_{i,w}\right)}{2+\lambda}$, the optimal solution can be obtained via Eq. (19).

$$L(\Theta)^* = \frac{4*\left(\sum_{k=1}^{K-1} f_k(p_{i,w}) - \widehat{D}_{i,w}\right)^2}{2+\lambda} + \gamma T \tag{19}$$

4.2 Solution algorithm for stage two site selection model

The stage two site selection model is solved with a proposed greedy-based stochastic spatial search algorithm. It consists of two components, randomized searching set construction and greedy spatial interchange, which is presented by the Framework below. The basic idea is that starting from an initial solution $S = \{s_p\}$, construct a stochastic search set for each charging station s_p, and perform greedy-based spatial interchange to identify a candidate site with the highest charging demand improvement. This candidate site will be used to swap out s_p in the previous solution, and finally, iteratively repeat this process until the stopping criteria is satisfied.

Framework: Greedy-based stochastic spatial search algorithm

Input: initial solution: $S = \{s_p | p = 1,2 \dots, p_{max}\}$
Repeat
  For s_p where $p = 1$ to p_{max}, do
    Construct stochastic searching set (Component 1)
    Perform greedy-based spatial interchange (Component 2)
Until stopping criteria
Output: optimal solution: $S^* = \{s_p^* | p = 1,2 \dots, p_{max}\}$

The first component, stochastic searching set construction, is introduced below. For an CS location in the existing solution, its searching set is composed of all candidate locations within its adjacent areas (which could be traffic analysis zones or zip code areas), and randomly select candidate locations from the non-adjacent areas. The number of non-adjacent candidate locations is relatively high, but gradually decreases, which is controlled with a decay coefficient. The searching in non-adjacent areas introduces stochasticity to avoid a quick convergence to local minima. The output of stochastic searching set construction is a set of candidate sites for component 2 - greedy spatial interchange.

Component 1: Stochastic searching set construction

Input:
  A set of neighborhood candidate locations for s_p in current solution S: $NB(s_p)$
  A set of non-neighborhood candidate locations for s_p in current solution S: $NNB(s_p)$
Randomly select several n_{rd} locations from $NNB(s_p)$ to get $SUB\_NNB(s_p)$
Construct searching set for s_p: $SS(s_p) = NB(s_p) \cup SUB\_NNB(s_p)$
Update n_{rd} by multiplying a decay coefficient
Output: Searching set $SS(s_p)$

The second component, the greedy spatial interchange is presented in component 2 below. For a CS location s_p in the existing solution, a candidate location is firstly selected from the searching set and then inserted into the current solution set to replace s_p. If the total charging demand of the updated solution set is higher than before, such interchange between s_p and the candidate location is considered favorable and will be added to the potential interchange set. After every candidate location in the searching set has been iterated, the optimal interchange in the potential interchange set will be used to update the current solution set, which can greedily improve the value of the objective function.

Component 2: Greedy spatial interchange

Input:
  Current solution: $S' = \{s_p | p = 1,2 \dots, p_{max}\}$
  Existing CSs set: $E = \{e_q | q = 1,2 \dots, q_{max}\}$
  The searching set for s_p: $SS(s_p)$
  Initial potential interchange set for s_p: $IC(s_p) = \emptyset$
Repeat
  Select a candidate location c'_{s_p} from $SS(s_p)$
  Let $S'' = S' - s_p + c'_{s_p}$

If $obj(E, S'') > obj(E, S')$, $IC(s_p) = IC(s_p) \cup c'_{s_p}$

Until every c'_{s_p} in $SS(s_p)$ has been iterated

Select the potential interchange location $c^*_{s_p}$ from $IC(s_p)$ that could maximize $obj(E, S'') - obj(E, S')$

Perform the interchange by $S' = S' - s_p + c^*_{s_p}$

Output: Updated solution set: S'

5. CASE STUDY

In this section, we test the performance of the proposed two-stage model with a real-world dataset from Kansas City Missouri (KCMO). We will first present the data with some of its basic characteristics, and define the features to be used in the stage one model. Then, we show the numerical analysis results for stage 1 and stage 2 models, respectively.

5.1 Data description and feature definition

This research uses a multi-source dataset collected in KCMO, which includes the charging event log data, travel demand model data, and climate data. A detailed description of each data source is provided below. In total, 67,576 data samples are extracted, and 15 features are defined in Table 1. These features can be categorized into four main categories: 1) spatial context information, 2) weather information, 3) port type and 4) traffic information. Notably, some variables are continuous numerical data, while others are categorical data.

Table 1 Definition of 15 Features and Response Variable

Category	Notation	Definition and Unit	Data Type
Response Variable	$D_{i,w}$	The number of charging events of CS i on week w	Numerical
Spatial Context Information	$NHCS_i$	The density of charging stations in the same zip code area of CS location i (#CSs/sq mile)	Numerical
	$LDU1_i$	Land use type: if CS location i is in an institutional area	Binary
	$LDU2_i$	Land use type: if CS location i is in a transportation area	Binary
	$LDU3_i$	Land use type: if CS location i is in a commercial area	Binary
	$LDU4_i$	Land use type: if CS location i is in a residential area	Binary
	$LDU5_i$	Land use type: if CS location i is in a recreational area	Binary
	$LDU6_i$	Land use type: if CS location i is in a vacant area	Binary
	$LDU7_i$	Land use type: if CS location i is in an industrial area	Binary
Weather Information	$PRCP_w$	Weekly precipitation of week w (mm)	Numerical
	TMP_w	Weekly average temperature of week w $(°C)$	Numerical
	WD_w	Weekly average wind speed of week w (m/sec)	Numerical
Port Type	$PT1_i$	Port type of DC fast charger for CS location i	Binary
	$PT2_i$	Port type of level 2 charger for CS location i	Binary
Traffic Information	$AADT_i$	Annual average daily traffic on CS location i's nearby roads	Numerical
	TP_i	Trip production of the TAZ where CS location i is located	Numerical

5.1.1 Charging event log data

The charging event log dataset includes a total number of 22,0231 charging records, collected at 444 public charging stations in KCMO, from January 2014 to December 2019. The spatial

distribution of existing CSs is shown in Figure 2. We can find that most CSs are concentrated in the downtown area, where most of the high-demand CSs are also observed.

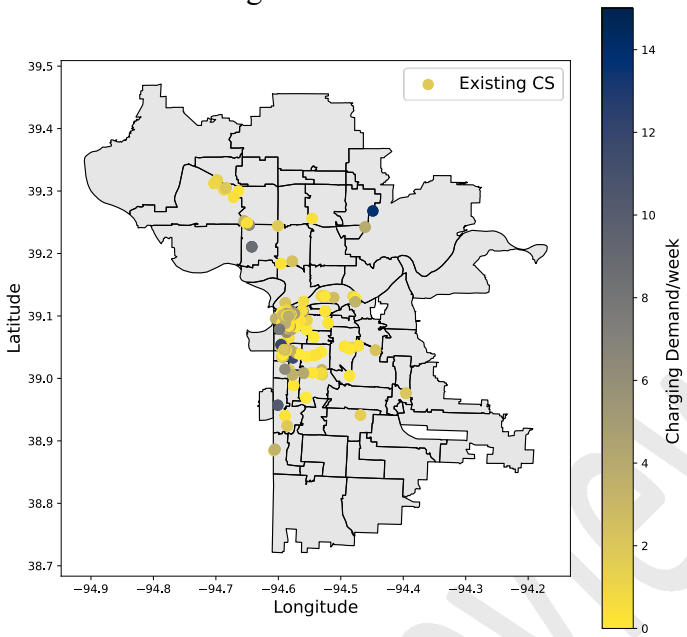


Figure 2 The spatial distribution of existing CSs in Kansas City, MO

For the charging event dataset, the attributes include start date, end date, latitude, longitude, plug type, and postal code of a charging event. So, we can derive the response variable and features including spatial context information and port type. The response variable is the weekly charging rate, which is shown in Eq. (20), where $count(CE_{i,w})$ denotes the total number of charging events from the charging station i on the week w. Its distribution is shown in Figure 3. It can be observed that for each CS, the number of charging events is mostly under 10 times per week, with a longtail where the maximum reaches 58.

$$D_{i,w} = count(CE_{i,w}) \tag{20}$$

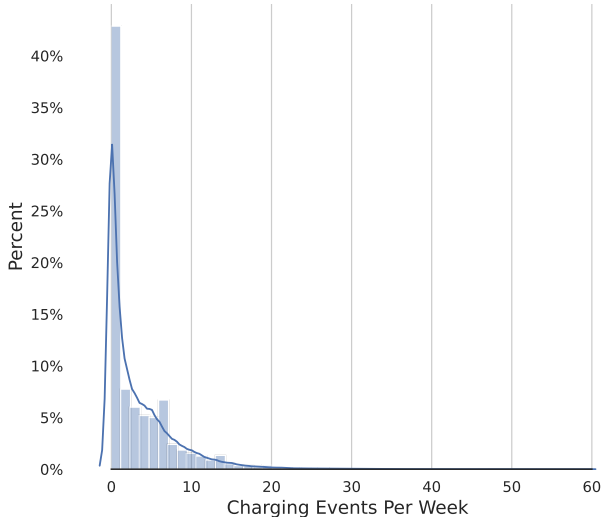
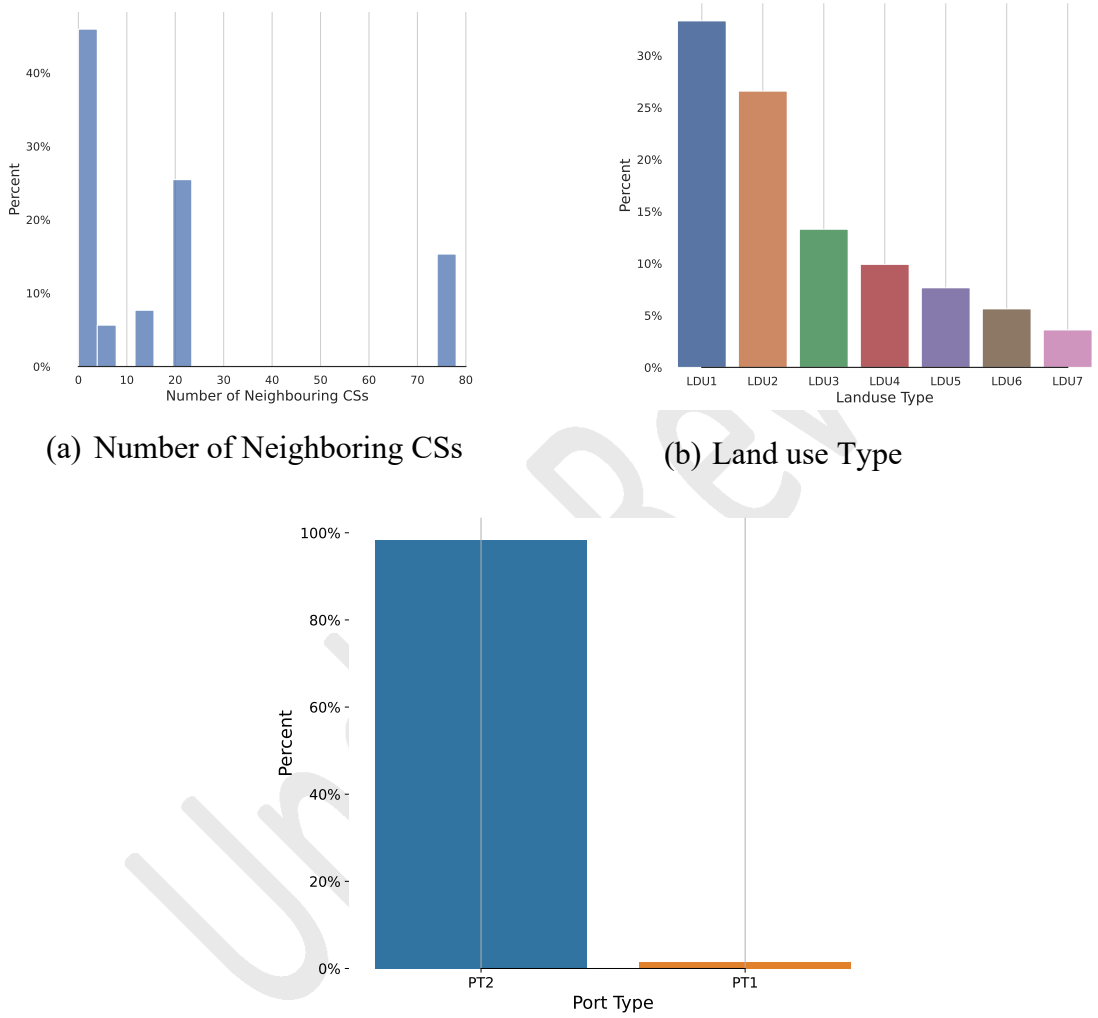


Figure 3 The distribution of response variable, charging demand

When it comes to relevant features, the number of neighboring existing CSs, $NHCS_i$, is defined as the density of charging stations (#CSs/sq mile) in the same zip code area of CS

location i, as shown in Eq. (9). Of all 444 stations, the distribution of $NHCS$ is given in Figure 4-(a), where we can find three clusters. Such clusters denote that the density of CSs may be below 10, around 20, or over 70 per sq mile. $LDU1{\sim}7$ are the land use type of the parcel where the CS i is located. In Figure 4-(b), the distribution of seven land use types of 444 CS locations is presented. Most of the CSs are located in $LDU1$ and $LDU2$, i.e., institutional areas like offices and schools, transportation areas like stations and airports. As Figure 4-(c) illustrates, majority charging stations are equipped with the level 2 chargers ($PT2_i$), whereas $PT1_i$ is the fast charger and accounts for only about 5% of the charging stations.



(a) Number of Neighboring CSs

(b) Land use Type

(c) Port Type

Figure 4 The distribution response variables

5.1.2 Travel demand model data

The travel demand model data is provided by Mid-America Regional Council (MARC), which is the MPO for the bi-state Kansas City region. The data includes a daily OD demand matrix and a road network of Kansas City with the assigned peak hour traffic volume. Kansas City is divided into 2,510 Traffic Analysis Zones (TAZs). By summarizing the OD demand matrix by the row, the trip production of each TAZ can be obtained. Consequently, the feature TP_i, trip production of the TAZ where charging station i is located, can be calculated by Eq. (21), where $TPSUM_i$ is

the summation of trip production generated in the zip code area where CS i is located; $ARSZ_i$ is the size of the zip code area of CS i.

$$TP_i = TPSUM_i / ARSZ_i \quad (21)$$

Next, the road network of Kansas City with the assigned peak hour traffic volume is given in Figure 5, which includes a total of 10, 533 nodes and 24, 601 links. On each link, the darker color denotes heavier traffic while the lighter color represents smaller traffic volume.

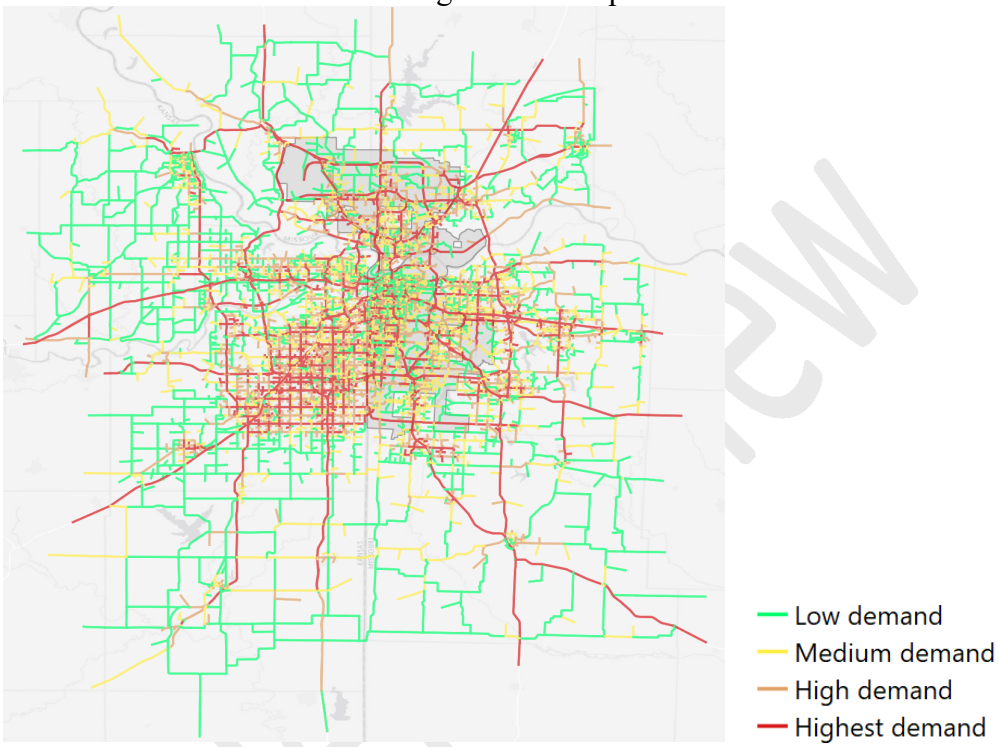


Figure 5 Traffic Distribution on 24,601 Links of Entire Kansas City

Then, the feature $AADT_i$, annual average daily traffic on CS location i's nearby roads, can be calculated by Eq. (22), where $AADTSUM_i$ is the summation of daily traffic volume on every roadway segment in the zip code area where CS i is located. Though TP_i and $AADT_i$ are all traffic-related features, it should be noted that TP_i is focused on the number of vehicles originating from the a particular zone, while $AADT_i$ includes also the bypass traffic.

$$AADT_i = AADTSUM_i / ARSZ_i \quad (22)$$

5.1.3 Climate data

The climate data is provided by NCEI via open access API, and includes weather station location, record date, average wind speed in m/sec, precipitation in mm, and the average temperature in $°C$. Then, as for the weather-related features, the weekly average temperature of week w, TMP_w can be calculated by Eq. (23), where $Tavg_w$ is the average daily temperature of date in week w. Similar logic can be adopted to get WD_w, the weekly average wind speed of week w.

$$TMP_w = sum(Tavg_w)/7 \quad (23)$$

The weekly precipitation of week w, $PRCP_w$, can be derived by Eq.(24), where $Prcp_w$ is the precipitation value of each day in week w.

$$PRCP_w = sum(Prcp_w) \quad (24)$$

After data processing, the total number of data records is 67,576, each including the 15 defined features and the response variable. Such dataset is randomly divided into a training subset which has 80% of the data, and a testing subset with 20% data.

5.2 Test scenario setting

To validate the two-stage model, we randomly generated 5 points of interest from each zip code area, and use them as candidate CS locations. In the end, we have N_c=355 (i.e., 5*71 zip code areas). N_d is set to be 50, i.e., 50 optimal sites need to be selected from 355 candidate locations, with the goal of maximizing the charging demand of both 444 existing (i.e., N_e) and 50 newly built CSs. The spatial distribution of 355 candidate locations is shown in Figure 6.

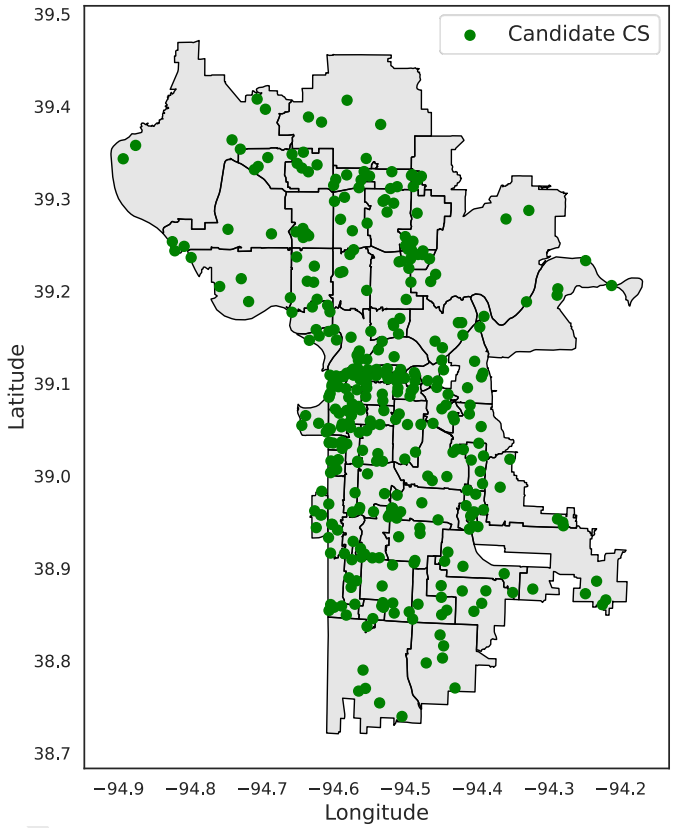


Figure 6 The distribution of 355 candidate CS locations

Then, we set the other initial parameters as follows. The upper bound of new-built CSs in a single zip area (i.e., N_z) is 3. In the solution process, the initial number of candidate locations selected from $NNB(c_m)$, i.e., n_{rd}, is 15. The convergence criteria between two iterations to terminate searching (i.e., err) is set to 0.01. The numerical analysis is implemented in a Python environment on a desktop computer with Intel (R) Xeon (TM) W-2295 CPU 18 Core@ 3.00 GHz.

5.3 Stage one ensemble learning model training results

With the scenario setting as described above, a total of 51,896 sub-regression trees are built. The prediction performance of the model on the training and testing dataset are measured by R square, root mean squared residuals (RMSE), and mean absolute error (MAE). As shown in Figure 7, R square values are 0.72 and 0.63, RMSE are 2.38 and 2.77, and MAE are 1.62 and

1.85, on the training and testing subsets, respectively. Thus, the model performance is satisfactory in generating accurate charging demand prediction values.

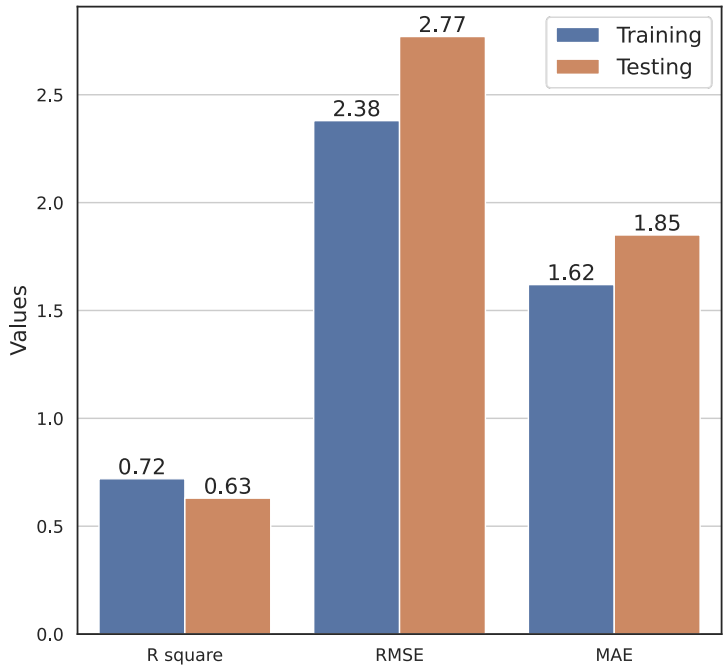


Figure 7 The prediction performance on training and testing sets

To understand the impact of the defined features on charging demand, the importance of each feature is measured by the Shapley Additive Explanation (SHAP) method, which produces consistent output by averaging all possible orderings of input feature permutations [21]. The top ten features with the largest importance value are given in Figure 8, with X-axis being the mean SHAP value that measures the average impact of a specific feature on modeling output, and Y-axis being the features. It is found that the two most important predictors are all traffic-related, nearby traffic $AADT_i$ and trip production TP_i. The number of neighboring CSs, $NHCS_i$ ranks the fourth, which validates the demand-supply coupled relationship, as charging demand is shown to be affected by the density of CSs in the same neighborhood.

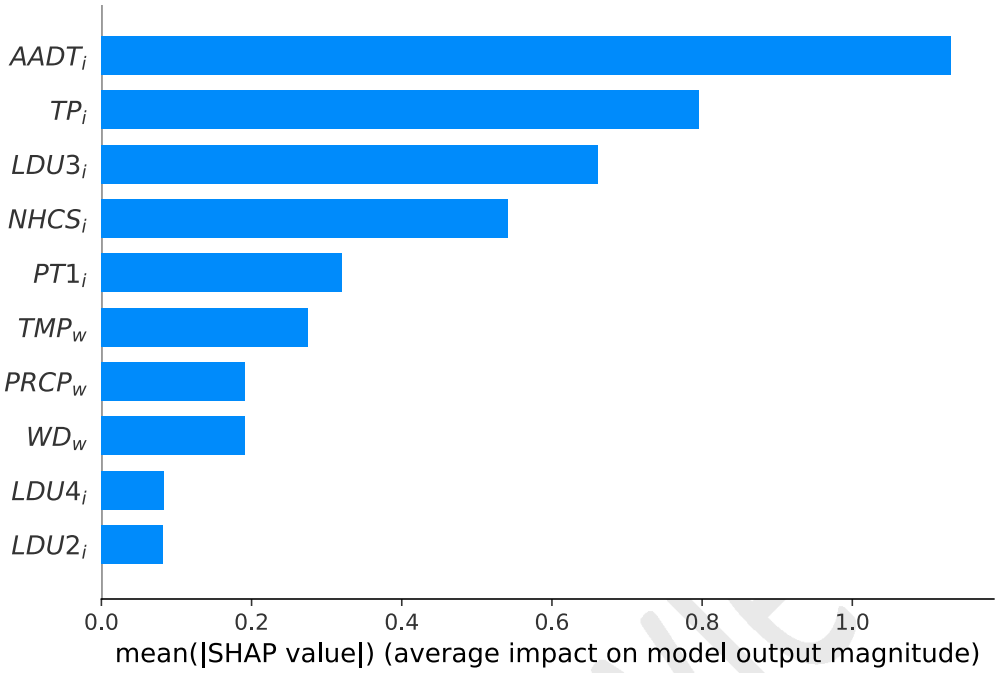


Figure 8 The mean SHAP value for the 10 most important features

5.4 Stage two site selection model results

The solution process of the stage two model is shown in Figure 9, with the X-axis being the number of iterations and Y-axis being the objective function value. The objective function value, predicted weekly charging rates, has been improved from the initial solution of 1,278 to the final optimal solution of 1,459 after eight iterations. The Figure also shows the objective function does not change much after the third iteration, indicating the proposed model has a fast convergence property.

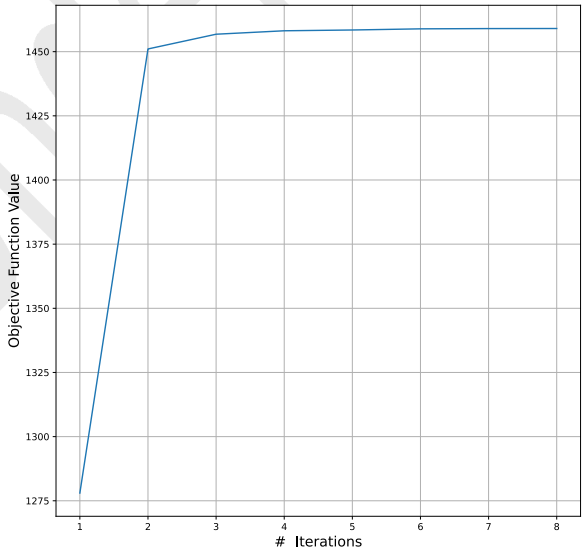
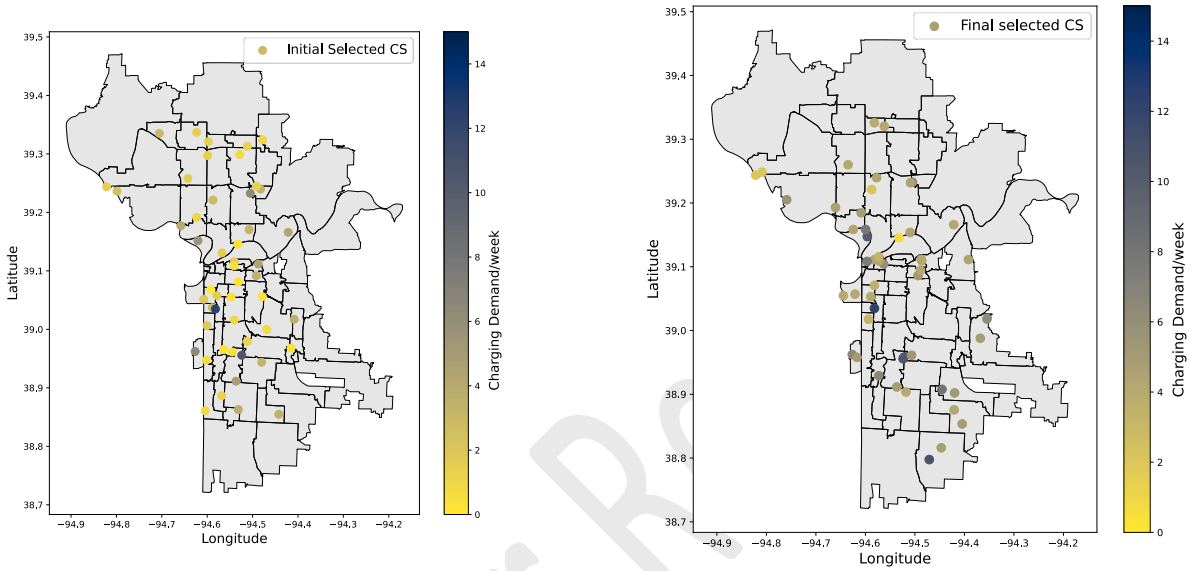


Figure 9 Solution iteration process

Then, the spatial distribution of CSs in the initial and final solution is shown in Figure 10 (a) and (b), respectively. The fifty charging stations in the initial solution are evenly distributed on the map, with at most one charging station in each zip code area. However, in the optimal

solution, the final 50 CSs are more concentrated in certain zip code areas. For example, one zip code area has three sites selected, and 18 zip code areas have two sites selected. It can also be observed that the zip code areas with more than one CS are more concentrated in the downtown area. In terms of the changing demand, we can see that the color of CSs in Figure 10 (b) is darker than that in Figure 10 (a), indicating the charging demand of the final selected CSs are higher than the initial solution. The explicit charging demand values are given in Figure 11, where the charging demand of initial and final selected CSs are 125 and 264 charging events per week, respectively.



(a) Initially selected CSs

(b) Finally selected CSs

Figure 10 Distribution of initial and final selected CSs

To test the effectiveness of the proposed solution algorithm, we compared our modeling results (indicated as "final solution") with those from two benchmark models, with the first one simply selecting the candidate locations with the highest predicted charging demand (i.e., naïve greedy selection strategy), and the second model that updates selection only from neighboring area but ignores demand-supply coupled interactions (i.e., naïve neighbor swap strategy). The result is presented in Figure 11. It can be observed that (1) for the 444 existing charging stations in Kansas City, their charging usage rates are impacted by the strategies of selecting new charging stations. The proposed model and the two benchmark algorithms generate a demand of 1,195, 1,163, and 1,195, respectively. In other words, the development of new charging stations increased the usage frequency of existing charging stations, and the effectiveness of the proposed algorithm is equivalent to that of benchmark algorithm two; (2) for the 50 newly developed charging stations, these three algorithms can all significantly increase charging rates (over the initial solution), with benchmark model one achieving the best performance; (3) however, when all charging stations are combined (444 existing charging stations plus 50 selected new charging stations), the proposed algorithm generate the highest usage rate, at 1,459 charging events, which is significantly higher than the initial solution (1,278), and the two benchmark algorithms (at 1,429 and 1,451, respectively). In other words, the proposed algorithm is able to identify the optimal charging locations in a way to better serve the charging demand, not only for the newly selected charging stations but also for the existing charging stations that have been built. The

results of the proposed algorithm can improve charging usage by 14% over the initial solution, and outperforms the two benchmark algorithms by 2% and 0.5%, respectively.

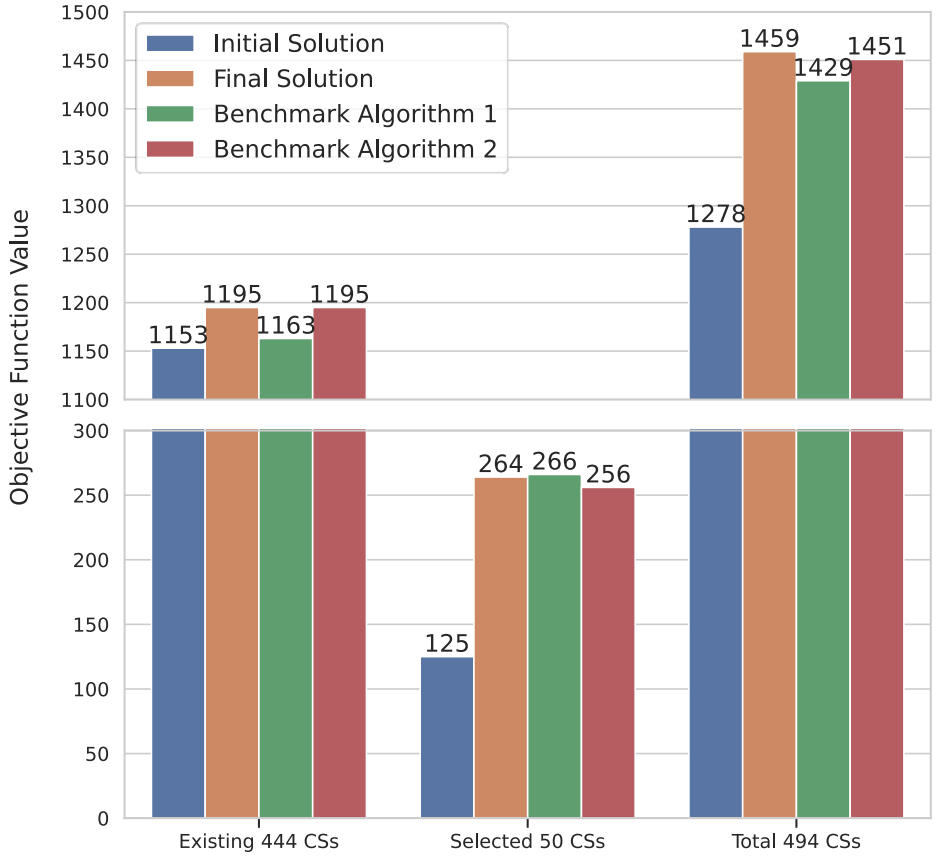


Figure 11 Solution Results

6. CONCLUSION

This manuscript develops a two-stage demand prediction and optimization modeling framework for the charging station location problem. In stage one, a gradient boost-based model is developed to generate accurate predicted charging demand, which is explicitly measured as the charging usage rate per week. Next, in the stage two model, a demand-supply coupled CS location selection model is developed with the objective of maximizing the total charging demand of both existing and newly selected CSs. The proposed CSLP model is solved with a greedy-based stochastic spatial search algorithm.

A case study using multi-source real-world data from Kansas City Missouri is performed to test the effectiveness of the proposed model. The used dataset includes a total number of 22,0231 charging records, collected at 444 public charging stations in KCMO, from January 2014 to December 2019. The travel demand model data and climate data are also retrieved for research purposes. Based on these multi-source data, a total of 15 features are extracted. Results show that the stage one model generates a satisfactory charging demand prediction result, with R square values reaching 0.72 and 0.63 on the training and testing datasets, respectively. The stage two model is demonstrated to improve charging usage by 14%, and outperform the results of two benchmark models, namely naïve greedy selection strategy and naïve neighbor swap strategy.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy's (EERE) Vehicle Technologies Office under the Award Number DE-EE008474. The authors are also thankful for the support from the Metropolitan Energy Center (MEC), the City of Kansas City Missouri (KCMO), Lilypad, Mid-America Regional Council (MARC), and Evergy (formerly Kansas City Power and Light Company (KCP&L)).

8. AUTHOR CONTRIBUTIONS

The authors confirm their contribution to the paper as follows: study conception and design: Yang Song, Xianbiao Hu; data process: Yang Song, Yiyang Wang; analysis and interpretation of results: Yang Song; draft manuscript preparation: Yang Song, Qing Tang, Xianbiao Hu. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Zhang, X., D. Rey, and S.T. Waller, *Multitype recharge facility location for electric vehicles.* Computer-Aided Civil and Infrastructure Engineering, 2018. **33**(11): p. 943-965.
2. Jung, J. and R. Jayakrishnan, *High-coverage point-to-point transit: electric vehicle operations.* Transportation Research Record, 2012. **2287**(1): p. 44-53.
3. Kchaou-Boujelben, M., *Charging station location problem: A comprehensive review on models and solution approaches.* Transportation Research Part C: Emerging Technologies, 2021. **132**: p. 103376.
4. Hodgson, M.J., *A flow-capturing location-allocation model.* Geographical Analysis, 1990. **22**(3): p. 270-279.
5. Kuby, M. and S. Lim, *The flow-refueling location problem for alternative-fuel vehicles.* Socio-Economic Planning Sciences, 2005. **39**(2): p. 125-145.
6. Capar, I., M. Kuby, V.J. Leon, and Y.-J. Tsai, *An arc cover–path-cover formulation and strategic analysis of alternative-fuel station locations.* European Journal of Operational Research, 2013. **227**(1): p. 142-151.
7. Yıldız, B., O. Arslan, and O.E. Karaşan, *A branch and price approach for routing and refueling station location model.* European Journal of Operational Research, 2016. **248**(3): p. 815-826.
8. Wang, Y.-W. and C.-C. Lin, *Locating road-vehicle refueling stations.* Transportation Research Part E: Logistics and Transportation Review, 2009. **45**(5): p. 821-829.
9. Klose, A. and A. Drexl, *Facility location models for distribution system design.* European journal of operational research, 2005. **162**(1): p. 4-29.
10. He, S.Y., Y.-H. Kuo, and D. Wu, *Incorporating institutional and spatial factors in the selection of the optimal locations of public electric vehicle charging facilities: A case study of Beijing, China.* Transportation Research Part C: Emerging Technologies, 2016. **67**: p. 131-148.
11. Cui, Q., Y. Weng, and C.-W. Tan, *Electric vehicle charging station placement method for urban areas.* IEEE Transactions on Smart Grid, 2019. **10**(6): p. 6552-6565.
12. Stephens-Romero, S.D., T.M. Brown, J.E. Kang, W.W. Recker, and G.S. Samuelsen, *Systematic planning to optimize investments in hydrogen infrastructure deployment.* International journal of hydrogen energy, 2010. **35**(10): p. 4652-4667.
13. Huang, K., P. Kanaroglou, and X. Zhang, *The design of electric vehicle charging network.* Transportation Research Part D: Transport and Environment, 2016. **49**: p. 1-17.
14. Cavadas, J., G.H. de Almeida Correia, and J. Gouveia, *A MIP model for locating slow-charging stations for electric vehicles in urban areas accounting for driver tours.* Transportation Research Part E: Logistics and Transportation Review, 2015. **75**: p. 188-201.
15. Guo, Z., J. Deride, and Y. Fan, *Infrastructure planning for fast charging stations in a competitive market.* Transportation Research Part C: Emerging Technologies, 2016. **68**: p. 215-227.
16. Bernardo, V., J.-R. Borrell, and J. Perdiguero, *Fast charging stations: Simulating entry and location in a game of strategic interaction.* Energy Economics, 2016. **60**: p. 293-305.
17. Chen, Z., F. He, and Y. Yin, *Optimal deployment of charging lanes for electric vehicles in transportation networks.* Transportation Research Part B: Methodological, 2016. **91**: p. 344-365.
18. Wang, C., F. He, X. Lin, Z.-J.M. Shen, and M. Li, *Designing locations and capacities for charging stations to support intercity travel of electric vehicles: An expanded network approach.* Transportation Research Part C: Emerging Technologies, 2019. **102**: p. 210-232.
19. Dong, X., Z. Yu, W. Cao, Y. Shi, and Q. Ma, *A survey on ensemble learning.* Frontiers of Computer Science, 2020. **14**(2): p. 241-258.
20. Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, *Lightgbm: A highly efficient gradient boosting decision tree.* Advances in neural information processing systems, 2017. **30**: p. 3146-3154.
21. Lundberg, S.M., G.G. Erion, and S.-I. Lee, *Consistent individualized feature attribution for tree ensembles.* arXiv preprint arXiv:1802.03888, 2018.