HARBIN INSTITUTE OF TECHNOLOGY

STUDY ON IMPACT FACTORS OF CHARGING STATION UTILIZATION USING RANDOM FOREST AND SHAP VALUE METHODS

Haiming Cai[1] hemingway@stu.hit.edu.cn
Fan Wu[2] wu3@ualberta.ca
Zhanhong Cheng[3] zhanhong.cheng@mail.mcgill.ca
Jian Wang[1] wang_jian@hit.edu.cn

[1]Department of Transportation Science and Engineering, Harbin Institute of Technology, Harbin, 150006,China
[2]Department of Civil and Environmental Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada
[3]Department of Civil Engineering and Applied Mechanics, McGill University, H3A 0C3, Canada

TRB ANNUAL MEETING January 8–12, 2023

INTRODUCTION

- The study aims to ***explore relationships*** between charging station utilization and urban form and demand for taxis.
- The study conducts a ***data-driven empirical study*** based on a large-scale electric taxi trajectory dataset.

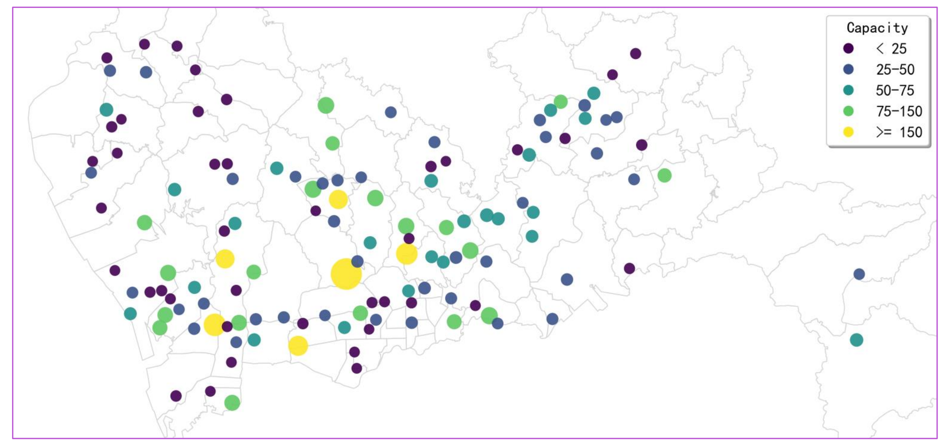

Fig. 1. Spatial distribution and capacities of charging stations.

- A Random Forest model, explained by SHAP value method, is presented to investigate the potential ***non-linear associations*** and ***spatial-temporal interactions***.
- The study enhances the knowledge on ***the extent of the factors*** effectively affect utilization and gives ***guidance*** on how to ***prioritize factors*** in charging station planning.

DATA PREPARATION

Study Area

The data sets are collected from ***Shenzhen***, a pioneer in promoting EVs.

In January 2019, there were more than

20,000 electric taxis
260 charging stations
10,000 charging outlets

Table 1. Description of the data sets

Data set	Data source	Key information	Relevant variable(s)
Trajectory data	Shenzhen Transportation Committee*	taxi ID, longitude, latitude, time, speed, status (occupied/vacant)	hourly utilization rate, pick-up density, drop-off density
Area of intersts (AOI) data	Amap website	AOI ID, longitude, latitude, name, area, category	land-use entropy
Bus and metro station data	Amap website	station ID, longitude, latitude, name, route number	bus station density, metro station density
Population data	WorldPop Dataset	grid ID, longitude, latitude, population density	population density
Road network data	OpenStreetMap	road ID, longitude, latitude, road type	road density
Charging station data	Shenzhen Transportation Committee*	station ID, longitude, latitude, name, number of outlets, opening date	hourly utilization rate

* Non-open-source data, permission required to access and use.

Data and Variables

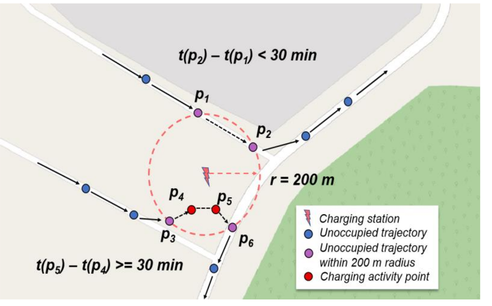

Fig. 2. Charging event identification algorithm.

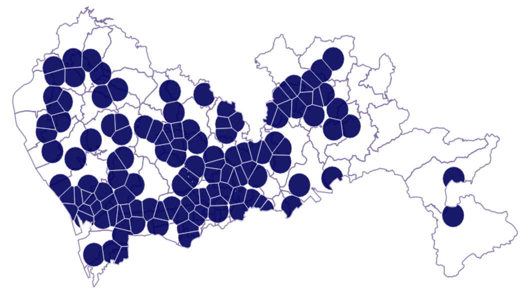

Fig. 3. Station-level buffer zones .

Taxi charging events are identified from trajectory data based on the ***identification algorithm***.

To count collective ***feature variables*** of each charging station, buffer zones are created based on ***Voronoi diagram***.

Dependent variable: hourly utilization rate

$$\text{utilization} = \frac{\sum_{i=1}^{n}\frac{d_i}{60}}{n}\times 100\% = \frac{\sum_{i=1}^{n} d_i}{60\cdot n}\times 100\%$$

where n is the number of charging outlets in the charging station, d_i is the occupied duration (in minute) of charging outlet i.

Independent variables: hourly pick-up density, hourly drop-off density, land-use entropy, bus station density, metro station density, population density, road density

$$\text{land-use entropy} = -\left[\sum_{j=1}^{N} P_j \cdot \ln\left(P_j\right)\right] / \ln N$$

where j represents the number of land use categories; P_j is the percentage of the j-th land use type in the area; N is the number of land use categories.

METHODOLOGY

Random Forests Regression (RFR)

The RFR is an ensemble learning method consisting of a multitude of ***decision trees***.

A ***hyperparameter tuning process*** is applied to acquire a set of optimal hyperparameter combinations with a ***5-fold cross-validation***.

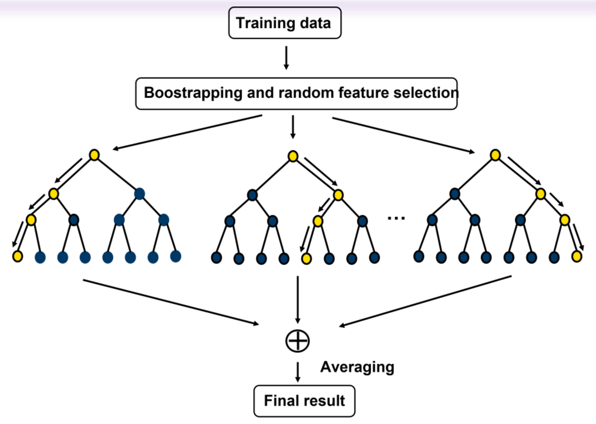

Fig. 4. An illustration of the random forest method.

SHapley Additive exPlanation (SHAP) method

The SHAP can describe the performance of a machine-learning model. All independent variables of the RFR are regarded as ***contributors*** in SHAP. The contribution of ***j-th*** variable x_j is

$$\phi_j = \sum_{S\subseteq\{x_1,\ldots,x_p\}\setminus\{x_j\}} \frac{|S|!\,(p-|S|-1)!}{p!}\left(f_x\left(S\bigcup\{x_j\}\right)-f_x(S)\right)$$

where S is the subset of variable set P, and $\{x_1,\ldots x_p\}\setminus\{x_j\}$ are the all possible sets of variables excluding x_j.

A linear function of binary variables g can be defined by the ***additive feature attribution method***

$$g(z') = \phi_0 + \sum_{i=1}^{p}\phi_i z'$$

where p is the number of variables; $z' \in \{0, 1\}^p$, and equals 1 when a variable is observed; otherwise, it equals 0

RESULTS & DISCUSSIONS

Variable Statistic Results

Table 2. Summary of the model variables

Variables	Units	Time-variants	min	25%	50%	75%	max
Explained Variable							
Hourly utilization rate	%	Yes	0	4.17	31.89	58.87	97.96
Explanatory variables							
Landuse entropy	-	No	0.0376	0.2826	0.3717	0.4608	0.6305
Population density	$10^4/km^2$	No	0.0699	0.5878	1.0957	1.6643	5.4652
Road density	km/km^2	No	1.2313	6.2748	7.8526	10.0873	15.5103
Bus station density	$/km^2$	No	1.2858	22.2790	35.2672	54.1860	114.5796
Metro station density	$/km^2$	No	0	0	0	0.3676	2.0610
Pick-up density	$/km^2$	Yes	0.2743	290.3025	1390.1937	6962.3552	60680.8050
Drop-off density	$/km^2$	Yes	1.1452	387.1054	1717.4702	7419.7860	54147.2941

There are great spatial differences in built environment variables among stations. At the same time, the temporal variables vary greatly in time. These show the ***unbalanced spatial-temporal utilization pattern***.

Feature Importance Analysis

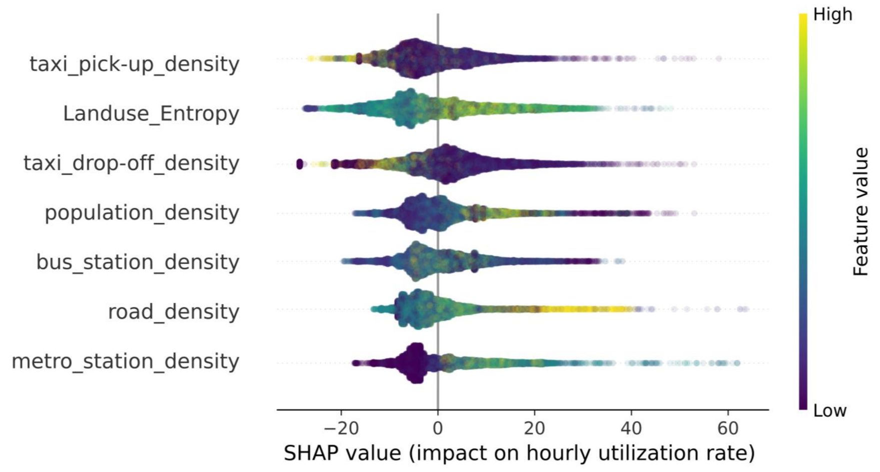

Fig. 5. SHAP Relative Importance plot.

- The ***TOP THREE*** important factors are ***pick-up density***, ***land-use*** and ***drop-off density***.
- The relationship between pick-up / drop-off density and utilization is ***non-linear***.
- Metro station density is in ***positive correlation*** with CS utilization.
- The relationships between utilization and land-use and population, road and bus station densities are ***more complicated***.

RESULTS & DISCUSSION

Relationship of Utilization & Variables

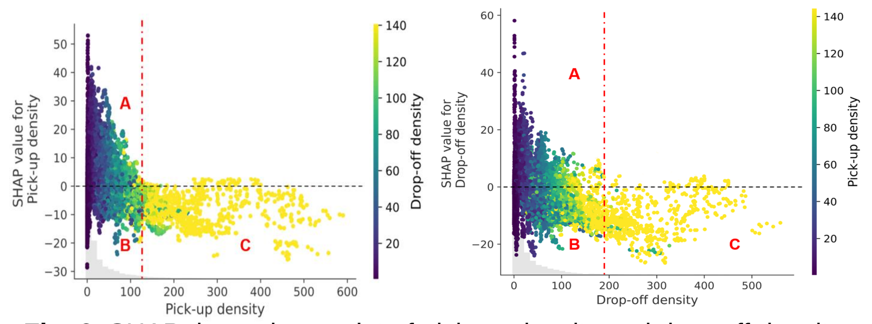

Fig. 6. SHAP dependence plot of pick-up density and drop-off density.

- The non-linear links of utilization and taxi demand can attribute to the ***interaction of built environment*** and drivers' ***profit-oriented operating strategies***.
- There are ***threshold effects*** of both pick-up and drop-off density.

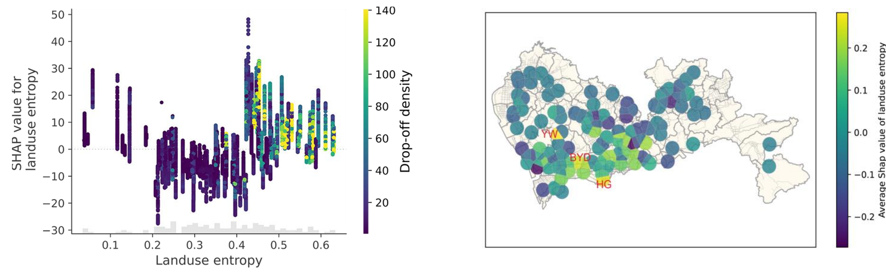

Fig. 7. SHAP dependency analysis on land-use and average SHAP value of land-use entropy.

- The charging stations with entropy value ***over 0.42*** tend to achieve high utilization.
- SHAP values generally ***decrease from downtown*** (southwest of the city) ***to suburbs*** (east and north of the city).

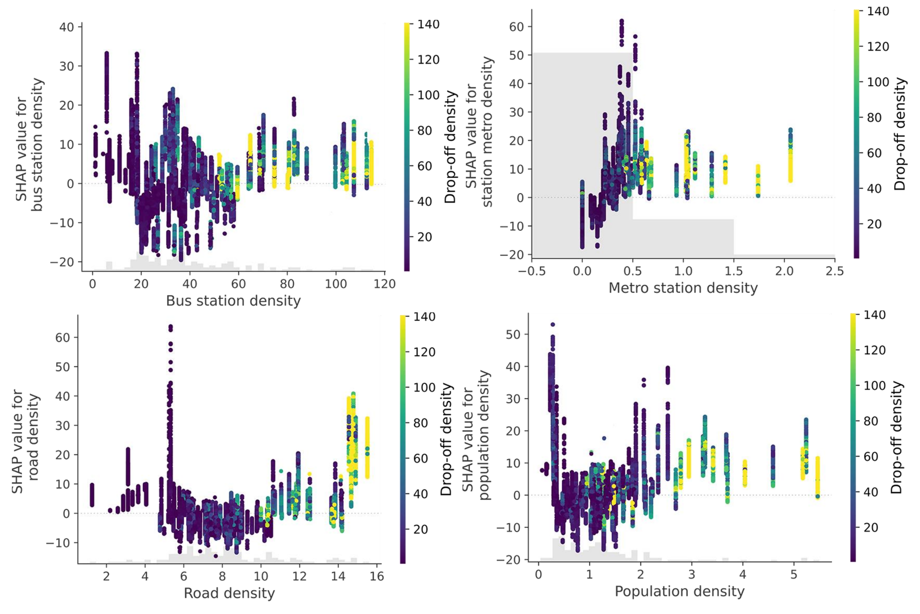

Fig. 8. SHAP dependency analysis on bus station, metro station, population, and road densities.

- Bus and metro station densities have ***different impacting mechanisms*** on utilization.
- ***Low*** values of road, population and bus station densities have ***positive influence*** on charging station utilization.
- Regions with such features above tend to be
  1) ***outer suburbs of the city***,
  2) ***parks and lakes***.

CONCLUSION

- The study reveals different impacting patterns of built environment and taxi demand on charging station usage. The non-linear relationships and threshold effects extend the understanding on the usage of charging stations.
- A more comprehensive study that incorporates more data sets, such as charging transection data and charging price data, could be investigated. This will provide more objective and evidence-based policy indications for charging station planning.

ACKNOWLEDGEMENT

This research work was supported by the *China Scholarship Council* (CSC). The contents of this paper reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein.