Impact of Regional Driving Behavior Differences on Traffic Flow

Yuting Wang[1,2], Zhaocheng He[1,2,3,*], Lijian Zhuang[4]

[1] School of Intelligent Engineering, Sun Yat-sen University
[2] Guangdong Provincial Key Laboratory of Intelligent Transportation System Sun Yat-Sen University
[3] The Pengcheng Laboratory
[4] Shenzhen Urban Transport Planning Center Co., Ltd.

* Corresponding author. E-mail: hezhch@mail.sysu.edu.cn

Supplementary Information

. Supplementary Tables

Supplementary Table 1. Mean values of performance metrics for each functional station type.

Cluster ID	Functional Type	No. of Stations	F_1	F_2	T_1	T_2	M_1	M_2	E_1	E_2	Connectors	Power (kW)
0	5413	71.49%	0.2	0.13	0.84	0.48	4.45	2.66	3.58	2.09	26.72	46.01
1	1236	16.32%	0.6	0.54	3.5	3.11	12.87	11.06	10.74	9.25	20.17	46.79
2	91	1.20%	3.16	1.08	2.26	0.93	63.94	9.36	120.74	51.98	19.34	132.43
3	832	10.99%	0.53	0.28	1.5	0.7	7.31	3.39	7.83	3.69	80.89	17.15

Supplementary Table 2. Description of charging station construction and operational data.

Field	Data Type	Description
station_id	Integer	A unique identifier for each physical charging station.
station_type	Enum	The category of the station (e.g., public, private, battery-swapping).
num_connectors	Integer	The total number of charging connectors (outlets) at the station.
num_piles	Integer	The total number of charging piles (physical units) at the station.
address	String	The detailed street address of the station.
site_context	String	The physical environment of the station (e.g., underground parking, roadside, shopping mall).
electricity_fee	Decimal	The price per kilowatt-hour (kWh) of electricity.
service_fee	Decimal	The service fee, charged either per session or per kWh.
longitude	Decimal	The longitude of the station's location (WGS84).
latitude	Decimal	The latitude of the station's location (WGS84).
total_power	Decimal	The total rated power of all charging equipment at the station, in kilowatts (kW).

Supplementary Table 3. Description of charging transaction data.

Field	Data Type	Description
station_id	Integer	A unique identifier for the charging station.
equipment_id	Integer	A unique identifier for a specific piece of charging equipment within a station.
connector_id	Integer	A unique identifier for a specific charging connector on a piece of equipment.
order_id	String	A unique identifier for each charging session.
start_time	Datetime	The start time of the charging session (to the second).
end_time	Datetime	The end time of the charging session (to the second).
energy_delivered	Decimal (kWh)	The total energy delivered during the session, in kilowatt-hours (kWh).

total_fee	Decimal	The total fee for the session, including both electricity and service charges.
electricity_fee	Decimal	The portion of the total fee corresponding to the electricity consumed.
service_fee	Decimal	The portion of the total fee corresponding to the service charge.

Supplementary Table 4. Description of network resilience metrics.

Category	Metric	Definition	Interpretation & Role in Analysis
Structural Integrity	Network size	The total number of nodes in the network.	Indicates the extent of node removal or failure.
	Number of connected components	The number of maximal subgraphs in which any two nodes are connected to each other by paths.	An increase signifies network fragmentation and reduced interconnectivity.
	Size of the largest connected component (LCC)	The number of nodes in the network's largest connected component.	A high value indicates the presence of a giant component maintaining overall function; a sharp decrease signals the collapse of the network backbone.
Core Robustness	Maximum k-core number	The largest integer k for which a k-core (a maximal subgraph where every node has a degree of at least k) exists.	A high value indicates a densely interconnected core. A rapid drop to zero during an attack signifies the failure of this core skeleton.
Accessibility & Efficiency	Average path length	The average of the shortest path lengths over all pairs of reachable nodes.	Reflects the average travel distance between nodes in a connected network. A sudden drop after fragmentation indicates that measurements are confined to small, local components.
	The average of the inverse of the shortest path lengths over all pairs of nodes (unreachable pairs contribute zero).	Measures the efficiency of information or resource flow across the entire network. A high value indicates high accessibility and efficiency. May exhibit complex behaviour post-fragmentation.	The average of the inverse of the shortest path lengths over all pairs of nodes (unreachable pairs contribute zero).

Algorithm 1 Charging Station Clustering and Rating Algorithm Based on Principal Component Dimensionality Reduction

Input: Feature matrix $X \in \mathbb{R}^{n\times d}$; KMO threshold τ; significance level α; cumulative explained variance threshold η; set of candidate cluster numbers $\mathcal{K}$; candidate algorithms $\mathcal{A} = \{\text{K-means, Agglo, DBSCAN}\}$
Output: Cluster label c_i for each station and corresponding clustering evaluation metrics

▷ Feature standardization, see equation (A1)
Perform Z-score standardization on X according to equations (3–5), obtain matrix Z, and construct correlation matrix R
▷ KMO and Bartlett's tests for sampling adequacy, see equations (A2)–(A3)
Calculate KMO statistic and Bartlett's test statistic χ^2 with its p-value
if KMO $\le \tau$ **or** $p \ge \alpha$ **then**
    return "Data is not suitable for principal component dimensionality reduction"
end if
▷ Principal component extraction and score calculation, see equations (A4)–(A6)
Perform eigendecomposition of R as $Rw_q = \lambda_q w_q$, sorted by λ_q in descending order
Calculate cumulative explained variance $\eta(Q)$ using equation (A5), select the number of principal components Q that satisfies $\eta(Q) \ge \eta$ and corresponds to an inflection point in the scree plot
Construct projection matrix $W = [w_1, \ldots, w_Q]$ and calculate principal component score matrix $Y = ZW$
▷ Clustering parameter determination and modeling, see equation (A7)
for all $A \in \mathcal{A}$ **do**
    if A is K-means or Agglomerative **then**
        Calculate WCSS(k) using equation (3–11) for $k \in \mathcal{K}$ and determine k^* using the elbow method
        Perform clustering on Y using algorithm A with cluster number k^*, obtaining label vector $c^{(A)}$
    else ▷ A is DBSCAN
        Select $(\varepsilon, \text{MinPts})$ based on the k-distance curve and run DBSCAN on Y, obtaining $c^{(A)}$
    end if
    ▷ Clustering quality evaluation, see silhouette coefficient equation (A8)
    Calculate the average silhouette coefficient $S^{(A)}$ for the current clustering result
end for
Select algorithm A^* with the maximum $S^{(A)}$ and its parameters, set final cluster labels $c_i \Leftarrow c_i^{(A^*)}$
return $\{c_i\}_{i=1}^n$ and corresponding clustering quality evaluation results

Detailed Steps and Formulas for the Charging Station Clustering and Rating Algorithm Based on Principal Component Dimensionality Reduction

(1) Symbol Definition and Data Standardization

Let there be n charging stations and d evaluation features, forming the feature matrix $X = [x_{ij}]_{(n\times d)}$, where x_i is the feature vector of station i. Considering the differences in scale among various indicators, Z-score standardization is applied:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{A1}$$

where μ_j and σ_j are the mean and standard deviation of the j-th feature, respectively. The standardized matrix is denoted as Z, which is used for correlation structure testing and dimensionality reduction.

(2) KMO and Bartlett's Sphericity Tests

Since there may be correlations among feature indicators, it is necessary to test whether the data structure is suitable for extracting latent variables before dimensionality reduction.

1) KMO test measures whether the correlations between variables are suitable for principal component/factor extraction. Let r_{jk} be the correlation coefficient and p_{jk} be the partial correlation coefficient, then:

$$\text{KMO} = \frac{\sum_{j\neq k} r_{jk}^2}{\sum_{j\neq k} r_{jk}^2 + \sum_{j\neq k} p_{jk}^2} \tag{A2}$$

2) Bartlett's sphericity test examines whether the correlation matrix R is an identity matrix. Given a sample size of n and d variables, the test statistic can be written as:

$$\chi^2 = -\left(n - 1 - \frac{2d+5}{6}\right) \ln |R| \tag{A3}$$

with degrees of freedom $d(d-1)/2$. When $p < 0.05$, the null hypothesis that "the correlation matrix is an identity matrix" is rejected, indicating sufficient correlation between variables for dimensionality reduction. In this study, KMO = $0.63 > 0.5$, and Bartlett's sphericity test $P < 0.05$, indicating that the data is suitable for dimensionality reduction clustering.

(3) Dimensionality Reduction: Principal Component Extraction and Score Calculation

After passing the applicability tests, the correlation matrix R (or covariance matrix) is constructed based on the standardized matrix Z, followed by eigendecomposition:

$$Rw_q = \lambda_q w_q, \quad q = 1, \ldots, d \tag{A4}$$

where λ_q is the eigenvalue and w_q is the corresponding eigenvector. The number of principal components Q is determined using both the scree plot and explained variance criteria. The cumulative explained variance is defined as:

$$\eta(Q) = \frac{\sum_{q=1}^{Q} \lambda_q}{\sum_{q=1}^{d} \lambda_q} \tag{A5}$$

Q is determined when the cumulative explained variance reaches the threshold and there is a clear inflection point in the scree plot. Subsequently, the low-dimensional representation (principal component scores) of stations in the latent space is calculated:

$$y_i = W^T z_i \tag{A6}$$

where $W = [w_1, \ldots, w_Q]$ is the projection matrix composed of the first Q principal component loading vectors, and y_i is the Q-dimensional latent representation of station i. The matrix Y, formed by all stations' y_i, serves as the input data for clustering.

(4) Determination of Optimal Cluster Number

After feature dimensionality reduction, the cluster number k needs to be determined based on the data itself. For center-based clustering methods like K-means, the elbow method is used to analyze the relationship between Within-Cluster Sum of Squares (WCSS) and k:

$$\text{WCSS}(k) = \sum_{c=1}^{k} \sum_{y_i \in C_c} \|y_i - \mu_c\|^2 \tag{A7}$$

where C_c represents the c-th cluster and μ_c is its centroid. Typically, WCSS(k) decreases as k increases. When the downward trend of the curve shows a clear "inflection point," the corresponding k can be considered as a reasonable number of clusters.

(5) Clustering Model Construction

After determining the clustering parameters, this study employs three typical unsupervised clustering methods for comparative analysis:

1. K-means: Suitable for approximately spherical cluster structures;
2. Agglomerative Clustering (hierarchical clustering): Merges clusters bottom-up, capable of revealing hierarchical structures;
3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Based on density, can identify clusters of arbitrary shapes and demonstrates strong robustness against outliers/noise points.

(6) Clustering Effect Evaluation

The Silhouette Coefficient is used to evaluate clustering quality. For a sample y_i, define: $a(i)$ as the average distance between y_i and samples in the same cluster; $b(i)$ as the average distance between y_i and samples in the nearest other cluster, then the Silhouette Coefficient is:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{A8}$$

Its value range is $[-1, 1]$. The closer $s(i)$ is to 1, the more compact within clusters and more separated between clusters; if $s(i) < 0$, it usually indicates that the sample is closer to another cluster, suggesting poor clustering results.

In this study, the Silhouette Coefficients for K-Means, DBSCAN, and Agglomerative Clustering are 0.46, -0.46, and 0.41, respectively. The K-means clustering algorithm performs best.

Algorithm 2 Progressive Node Failure Strategy Based on Power Ranking

Input: Initial network $G^{(0)} = (V^{(0)}, E^{(0)})$; a node power function sum_power(v); total number of steps T.

Output: The sequence of remaining networks $\{G^{(t)}\}_{t=0}^{T}$ and the corresponding sets of removed nodes $\{V_{\text{removed}}^{(t)}\}_{t=1}^{T}$.

$N \Leftarrow |V^{(0)}|$ ▷ Store the initial number of nodes

▷ Step 1: Define node power, see $P(v) = \text{sum\_power}(v)$

for all $v \in V^{(0)}$ **do**

  $P(v) \Leftarrow \text{sum\_power}(v)$

end for

▷ Step 2: Generate attack sequence by sorting nodes by power in descending order

Sort $V^{(0)}$ in descending order of $P(v)$ to obtain the sequence $\pi = (v_1, v_2, \ldots, v_N)$, such that $P(v_i) \geq P(v_{i+1})$.

▷ Step 3: Perform progressive attacks and update the network

$G^{(0)} \Leftarrow (V^{(0)}, E^{(0)})$

for $t = 1$ **to** T **do**

  $p_t \Leftarrow 0.1 \times t$ ▷ In this study, the attack fraction increases by 10% at each step

  $n_{\text{removed}}^{(t)} \Leftarrow \lfloor p_t \cdot N \rfloor$

  $V_{\text{removed}}^{(t)} \Leftarrow \{v_1, \ldots, v_{n_{\text{removed}}^{(t)}}\}$

  $V^{(t)} \Leftarrow V^{(0)} \setminus V_{\text{removed}}^{(t)}$

  $E^{(t)} \Leftarrow \{(u, v) \in E^{(0)} \mid u \in V^{(t)} \text{ and } v \in V^{(t)}\}$

  $G^{(t)} \Leftarrow (V^{(t)}, E^{(t)})$

  Calculate resilience metrics on $G^{(t)}$.

  if $V^{(t)} = \emptyset$ **then**

    break ▷ Terminate early if the network is completely disabled

  end if

end for

return $\{G^{(t)}\}_{t=0}^{T}$, $\{V_{\text{removed}}^{(t)}\}_{t=1}^{T}$, and the resilience metrics.

Algorithm 3 Phased Targeted Attack Strategy Based on Land-Use Zones

Input: Initial network $G^{(0)} = (V^{(0)}, E^{(0)})$; a land-use type function land_type$(v) \in \{A, B, R\}$; an attack sequence $\mathcal{L} = (L_1, \ldots, L_M)$.

Output: The sequence of remaining networks $\{G^{(s)}\}_{s=0}^{M}$ and the corresponding sets of removed nodes $\{V_{\text{removed}}^{(s)}\}_{s=1}^{M}$.

▷ Step 1: Define Land-Use Types

for all $v \in V^{(0)}$ **do**

  Assign land_type$(v) \in \{A, B, R\}$. ▷ e.g., A: Public Admin., B: Commercial, R: Residential

end for

▷ Step 2: Define the Phased Attack Sequence

Given an attack sequence $\mathcal{L} = (L_1, L_2, \ldots, L_M)$, where $L_s \in \{A, B, R\}$ is the land-use zone to be attacked in stage s.

▷ Step 3: Perform Phased Attacks and Update the Network

$G^{(0)} \Leftarrow (V^{(0)}, E^{(0)})$

for $s = 1$ **to** M **do**

  $V_{\text{removed}}^{(s)} \Leftarrow \{v \in V^{(s-1)} \mid \text{land\_type}(v) = L_s\}$ ▷ Remove all nodes in the specified zone

  $V^{(s)} \Leftarrow V^{(s-1)} \setminus V_{\text{removed}}^{(s)}$

  $E^{(s)} \Leftarrow \{(u, v) \in E^{(s-1)} \mid u \in V^{(s)} \text{ and } v \in V^{(s)}\}$

  $G^{(s)} \Leftarrow (V^{(s)}, E^{(s)})$

  Calculate resilience metrics on $G^{(s)}$.

  if $V^{(s)} = \emptyset$ **then**

    break ▷ Terminate early if no nodes remain

  end if

end for

return $\{G^{(s)}\}_{s=0}^{M}$, $\{V_{\text{removed}}^{(s)}\}_{s=1}^{M}$, and the resilience metrics for each stage.