Research on classification of highway service areas based on multifactor clustering

Cao Zhen, Sunyu Hui*, Geng Juan, Tian Zhun
Xi'an University of Architecture and Technology, Department of civil engineering, Xi'an, 710055, China

Abstract. This article proposes a classification method for highway service areas. POI (Point of Interest) data and surveys provide information on the area of highway service areas, distance from city centers, regional economic conditions, and population data. The clustering tendencies are analyzed using the Hopkins statistic, the number of clusters is determined using the elbow method, and the advantages and disadvantages of K-Means, FCM (fuzzy c-means), and HC (Hierarchical Clustering) are assessed using the CH (Calinski Harabasz), SC (Silhouette Coefficient), and DB (Davies-Bouldin) index. Using data from 95 highway service areas in Shaanxi Province as an example, the research findings indicate that the K-Means outperforms the FCM and HC according to all three evaluation indicators. Therefore, the article employs the K-Means to classify the 95 highway service areas in Shaanxi Province into three categories. The classification results obtained from this study provide a basis for the comprehensive development of highway service areas and the surrounding land.

1. Introduction

The classification of service areas on Chinese expressways is mainly based on standard specifications. The current specifications of the Chinese Ministry of Transport define the range of land area for service areas. "General Specifications for Highway Traffic Engineering and Along-Road Facilities Design" (JTG D80-2006) and "Land Use Indicators for Highway Engineering Projects" explicitly state that the land area of service areas is closely related to the number of highway lanes and traffic flow on road sections. The classification includes two major categories: service areas and parking areas, without further subdivisions. In other provinces of China, service areas are generally classified based on their land area.

Tan Huadong proposed a classification of service areas on Hunan's expressways to address the issues of indistinct characteristics and severe homogenization. Aimed at guiding the differentiated development of service areas and complying with the standards and specifications of the Ministry of Transportation, the classification is based on the basic service functions provided by the service areas, with a focus on traffic flow and land scale. The location of the service area is considered as an additional constraint, resulting in three major categories: Key Service Areas, Featured Service Areas, and Ordinary Service Areas [1]. Xu Yingjun and others, drawing insights from the layout planning of expressway service areas in provinces such as Anhui, Henan, and Guangdong, analyzed the urgency of service facility demands from various service objects such as vehicles, drivers, and passengers. They categorized service areas based on functionality, scale, and other factors, discussing the functional configuration of each type of service area [2]. To address issues like unclear functional division and design flaws in open-service areas on expressways, Zhao Mingfei, using a comparative analysis approach, proposed a classification and design approach for open-service areas based on practical case studies of expressway open-service area projects [3]. Liu Yang and others studied the construction status of expressway service areas in Hunan province. They used the Analytic Hierarchy Process (AHP) to establish a hierarchy model and weight model suitable for service areas. Combining quantitative results with qualitative principles, they developed a classification and grading system for Key Service Areas, Featured Service Areas, and Basic Service Areas, offering construction application methods [4]. Du Yanhua, focusing on the operational status of service areas and parking areas on expressways in Shanxi province, conducted research and analysis. They proposed a classification of expressway service areas based on land scale for different types of service areas, dividing those already operational in Shanxi province into Central Service Areas, Ordinary Service Areas, and Parking Areas, providing scientific theoretical guidance for the construction of expressway service areas in Shanxi province [5]. Hu Guisong and others introduced the "Three Types, Nine Points" factor index system, applying a combined classification analysis method to categorize expressway service areas into seven practical application types [6]. Korf JL, based on the Logit model, classifies railway stations into five categories according to the choice characteristics of commuters using urban rail transit. These categories are Central Area Stations, Residential Concentration Area Stations, Residential...
General Area Stations, Commercial Concentration Area Stations, and Undeveloped Stations[7]. Lambooy proposed using the location of rail transit stations within a city as a classification criterion and conducted a comprehensive analysis and categorization of all urban areas. The identified categories include Stations in Major City Centers, Stations in Medium-Sized City Centers with Strong Functional Areas, Stations in Urban Periphery Areas, Stations in Small City Centers, Commuter Area Stations, and Others[8].

In summary, previous studies on the classification of highway service areas have some limitations, such as considering only factors like traffic volume and area size, and conducting simplistic and qualitative classification of service areas without employing relevant clustering algorithms for quantitative analysis. Therefore, this study comprehensively considers the land use characteristics around service areas and the characteristics of highway service areas. It utilizes Points of Interest (POI) that reflect the land use characteristics around service areas, as well as indicators such as service area size, distance to city center, regional economy, and population as classification criteria. The study first employs Principal Component Analysis (PCA) to extract POI feature factors, and then compares the performance of K-Means clustering, FCM clustering, and hierarchical clustering algorithms using metrics such as CH coefficient, silhouette coefficient, and DB index. Finally, the research utilizes data from 95 highway service areas in Shaanxi Province as an example to conduct the relevant analysis and explore the characteristics of different categories of highway service areas. The research results not only provide theoretical references for planning and design departments in their planning and design of highway service areas but also offer technical support for operational management departments in the operation and management of service areas.

2. Clustering algorithm

2.1. K-Means Clustering

It involves analyzing data points by considering their positions and distances to each other. The goal is to divide the objects of study into mutually exclusive clusters (K), where objects within each cluster are as close to each other as possible while being as far away as possible from objects in other clusters. Each cluster is represented by a centroid, also known as a center point. In most cases, the distance used during the clustering process does not necessarily represent spatial distance. Generally, finding the unique solution to the problem of finding the global minimum requires exhaustively exploring all possible choices for initial points. However, repeated use of random starting points often leads to a solution that approximates the global minimum. The centroids are obtained by calculating the average values of each coordinate of the sample points within the cluster.

2.2. FCM Algorithm

FCM is an unsupervised clustering algorithm that is applicable to a wide range of problems related to feature analysis and clustering. FCM has been widely used in fields such as agricultural engineering, transportation, astronomy, and object recognition. With the development of fuzzy theory, the Fuzzy C-Means (FCM) algorithm based on Ruspini's fuzzy clustering theory was proposed. This algorithm analyzes the distances between different data objects and forms clusters based on these distances, with each cluster forming a cluster center. In fact, FCM is a data clustering method that divides a data set into n clusters, where each data point in the data set is associated with each cluster and has a higher degree of membership to the connected cluster, while data points far from the cluster have a lower degree of membership.

2.3. Hierarchical Clustering Algorithm

The hierarchical clustering algorithm is a commonly used unsupervised learning algorithm for partitioning data points into different clusters or groups. It constructs a hierarchical clustering structure based on the similarity or distance between data points, grouping similar data points into the same category and organizing them into a hierarchical structure based on their similarities. There are two types of hierarchical clustering algorithms: agglomerative clustering and divisive clustering. In this paper, we adopt divisive clustering.

3. Highway service area classification process

The classification process of highway service areas can be divided into four stages: data processing and feature extraction, feasibility of clustering, determination of the number of clusters, and selection of clustering methods and classification of types.

3.1. Data processing and feature factor extraction

3.1.1. Identification of land use features

The current study utilizes the open-source platform of Amap (AMap Open Platform) to obtain Point of Interest (POI) data for detailed characterization of land use features at the building level surrounding highway service areas. Referring to existing research classification methods, the POIs are categorized into 17 types of facilities, including food, hotels, shopping, lifestyle services, beauty, tourist attractions, leisure and entertainment, sports and fitness, education and training, cultural and media, healthcare, automotive services, transportation facilities, finance, real estate, companies and enterprises, and government institutions. These facilities are further classified into five land use categories: residential, commercial services, public administration and public services, road and transportation facilities, and green spaces and plazas.
3.1.2. Data standardization processing

To address the issue of comparability and eliminate the interference of different scales of POIs, it is necessary to perform standardization on the POI data obtained from the map open-source platform, as well as the variables such as service area size, distance from the city center, regional economy, and population data collected through surveys. This study employs the Z-score normalization method to standardize the POI data. The calculation of Z-score normalization is as follows:

\[
x'_i = \frac{x_i - \mu_i}{\sigma_i}
\]  

where \( x'_i \) is the Z-score standardized value for the \( i \)-th of POI in the highway service area; \( \mu_i \) is the mean value of the original Z-score for the \( i \)-th of POI; \( \sigma_i \) is the standard deviation of the original Z-score for the \( i \)-th of POI.

3.1.3. Extraction of major POI types

When there are numerous types of obtained POIs, there may be significant correlations among different types, requiring dimensionality reduction to improve the effectiveness of cluster analysis. In this study, the PCA (Principal Component Analysis) method is utilized to extract feature POIs. It involves an orthogonal transformation of potentially correlated POIs, aiming to transform multiple indicators into a few comprehensive indicators while minimizing the loss of data and preserving the original information to the maximum extent. Building upon previous research findings, this study sets a threshold parameter, with the cumulative contribution rate of eigenvalues not less than 85%, to determine the number of clusters.

3.2. Cluster trend

To validate the clusterability of the analyzed data and interpret the results of cluster analysis effectively, this study first performs cluster trend analysis on the research data. The Hopkins statistic is advantageous due to its lack of dependency on true labels, interpretability, scalability, efficiency, and robustness. Moreover, it has gained attention for its simplicity of implementation and low time and space complexity. Therefore, this study employs the Hopkins statistic to analyze the cluster trend. The calculation steps of Hopkins statistic are as follows:

(1) Randomly sample \( n \) points, ..., from the space of \( X \). For each point, find the nearest neighbor of in, and let \( d \) be the distance between and its nearest neighbor in \( X \). The calculation is as follows:

\[
x_i = \min_{v \in X} \{ dist(p, v) \}
\]  

(2) Randomly select \( n \) points, , , ..., from a space of the same size as the original dataset \( X \). For each point, find its nearest neighbor in \( X \), and let \( \delta \) be the distance

between and its nearest neighbor in \( X \). The calculation is as follows:

\[
y_i = \min_{v \in X} \{ dist(q, v) \}
\]  

(3) Hopkins statistic

\[
H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}
\]  

(4) Repeat the above steps multiple times to obtain the average of multiple Hopkins statistics. Based on the size of the average, we can judge the clustering tendency of the dataset. If the Hopkins statistic is greater than or equal to 0.5, the dataset has a good clustering tendency and is suitable for clustering analysis. Otherwise, the sample points in the dataset are relatively scattered and not suitable for clustering analysis.

3.3. Determining the number of clusters

Regardless of which clustering method is used, it is impossible to determine the number of clusters in advance. Therefore, the hierarchical clustering "elbow method" is introduced, which does not require specifying the number of clusters, and uses the "bend point" of the sum of squared errors within each cluster as the basis for determining the number of clusters.

The core indicator of the elbow method is SSE (sum of the squared errors).

\[
SSE = \sum_{i=1}^{k} \sum_{p \in C_i} (p - m_i)^2
\]  

where \( C_i \) is the \( i \)-th cluster; \( p \) is the sample points in \( C_i \); \( m_i \) is the centroid of \( C_i \); \( SSE \) is the clustering error of all samples and represents the quality of the clustering effect.

The core idea of the elbow method is that as the number of clusters \( k \) increases, the sample partition becomes finer, and the degree of aggregation of each cluster gradually increases, so the sum of squared errors (SSE) naturally decreases. When \( k \) is less than the true number of clusters, increasing \( k \) will significantly increase the degree of aggregation of each cluster, so the decrease in SSE will be large. However, when \( k \) reaches the true number of clusters, the return on increasing \( k \) for aggregation degree will rapidly decrease, so the decrease in SSE will suddenly slow down, and then gradually flatten out as \( k \) continues to increase.

3.4. Evaluation of clustering results

(1) CH (Calinski-Harabasz) coefficient

The Calinski-Harabasz coefficient is essentially the ratio of between-cluster distance to within-cluster distance, and the overall calculation process is similar to variance calculation. Therefore, it is also called the
variance ratio criterion. To cluster a dataset into K classes, the compactness within each class (intra-cluster distance) is measured by calculating the sum of squared distances between each point and its class center. The separation between classes (inter-cluster distance) is measured by calculating the sum of squared distances between each class center and the dataset center. The CH coefficient is calculated using the following formula:

\[ s = \frac{tr(B_k)}{tr(W_k)}(N-K) \]  

where \( s \) = the Silhouette coefficient; \( B_k \) = the between-class covariance matrix; \( W_k \) = the covariance matrix of within-class data; \( N \) = data capacity; \( c_q \) = the centroid of class q; \( c_e \) = center point of the dataset; \( q_c \) = the data set of class q.

A higher CH coefficient score indicates better clustering results, with smaller within-cluster covariance and larger between-cluster covariance being desirable.

4.1 Determining the number of clusters

The Hopkins cluster tendency test was conducted on the extracted dataset. MATLAB programming was used to input the data and calculate the statistical value of 0.65 (greater than 0.5). This indicates that the highway service area data exhibits significant clustering tendencies.

In this study, cluster category numbers ranging from 2 to 10 were selected. The within-cluster sum of squares was calculated for different cluster category numbers, as shown in Figure 1.

From the figure, it can be observed that as the number of cluster categories increases, the within-cluster sum of squares gradually decreases. However, the rate of decrease slows down significantly when the number of cluster categories is 3. Therefore, the optimal number of clusters is determined to be 3.

4.2 Clustering Results

Three clustering methods, namely K-Means, FCM, and hierarchical clustering, were employed for clustering analysis. The classification results were obtained using

\[ M_{ij} = \left( \sum_{i=1}^{k} |a_{ij} - a_{ik}| \right)^{\frac{1}{K}} \]  

\[ DB = 1 \sum_{i=1}^{k} \max_{q \neq j \in [1,K]} \left( \frac{s_j + s_i}{M_{ij}} \right) \]
MATLAB. Based on the MATLAB classification results, the clustering results for highway service areas are presented in Table 1.

### Table 1 Classification Results of Three Methods

<table>
<thead>
<tr>
<th>Algorithm name</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Means Algorithm</td>
<td>Qujiang service area,</td>
<td>sanyuan service area...(46)</td>
<td>guanzhuang service area...(47)</td>
</tr>
<tr>
<td></td>
<td>hancheng service area(2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FCM Algorithm</td>
<td>Qujiang service area,</td>
<td>baojibei service area...(28)</td>
<td>zhashui service area...(43)</td>
</tr>
<tr>
<td></td>
<td>hancheng service area...(24)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hierarchical Clustering Algorithm</td>
<td>Qujiang service area;</td>
<td>guanzhuang service area...(47)</td>
<td>pingli service area...(46)</td>
</tr>
<tr>
<td></td>
<td>hancheng service area(2)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2 Clustering Evaluation Results

<table>
<thead>
<tr>
<th></th>
<th>CH</th>
<th>SC</th>
<th>DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Means Algorithm</td>
<td>47.17</td>
<td>0.08</td>
<td>7.16</td>
</tr>
<tr>
<td>FCM Algorithm</td>
<td>23.00</td>
<td>-0.03</td>
<td>3.14</td>
</tr>
<tr>
<td>Hierarchical Clustering Algorithm</td>
<td>46.85</td>
<td>0.07</td>
<td>7.19</td>
</tr>
<tr>
<td>Range of Values</td>
<td>The larger, the better [-1,1], The larger, the better &gt;0, the smaller, the better</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results of the CH coefficient, silhouette coefficient, and DB index for the three clustering methods are shown in Table 2.

After analyzing the results of three clustering methods, it was observed that the Fuzzy C-Means (FCM) algorithm has significant differences compared to the other two algorithms and should not be adopted. On the other hand, the K-Means algorithm and hierarchical clustering algorithm exhibit minor differences. However, the quality of K-Means and hierarchical clustering algorithms can be evaluated using the CH coefficient, silhouette coefficient, and DB index. It was found that the K-Means clustering algorithm outperforms the hierarchical clustering algorithm in all three evaluation criteria. Therefore, the results of the K-Means algorithm are selected as the final classification outcome.

### 5. Results

In this study, the K-Means clustering algorithm was used to classify 95 highway service areas in Shaanxi Province into three categories. The distribution of the three categories is as follows: 2, 46, and 47 service areas, respectively.

Category 1 service areas include Qujiang Service Area located in Yan Ta District, Xi'an, and Hancheng Service Area located in Weiyang District, Xi'an. In terms of the surrounding amenities, Qujiang Service Area ranks first in the number of available services, third in the number of companies, first in regional economy, and first in regional population. On the other hand, Hancheng Service Area ranks second in the number of amenities, second in the number of companies, third in distance from the city center, and sixth in regional population. Therefore, it can be observed that both highway service areas possess excellent geographical features. Qujiang Service Area benefits from convenient transportation, proximity to commercial centers, clustering of cultural industries, high green coverage, and good environmental quality, showcasing significant locational advantages. As for Hancheng Service Area, being the only highway service area on the peripheral and airport expressways, it offers convenient transportation, an elegant environment, comprehensive facilities, and diverse functions. Hence, it is possible to conduct in-depth analysis by considering factors such as geographical conditions, locational advantages, transportation conditions, distribution of surrounding industries, policies, and planning of Qujiang Service Area and Hancheng Service Area.

Category 2 service areas include Huaxu Service Area, Ziwu Service Area, Ankang East Service Area, Hanzhong Service Area, Weinan West Service Area, and 46 other highway service areas. Overall, Category 2 service areas fall between Category 1 and Category 3 service areas in terms of classification indicators. Huaxu Service Area ranks first in the number of companies, Ziwu Service Area ranks third in regional economy, Shangzhou Service Area ranks sixth in distance from the city center, Hanzhong Service Area ranks second in distance from the city center, Baoji South Service Area ranks first in distance from the city center, and Weinan West Service Area ranks fourth in distance from the city center. From the classification indicators, it can be observed that some Category 2 service areas are relatively close to the city center. They show significant differences in economic and population clustering indicators compared to Category 1 service areas, possibly because Category 1 service areas are located in the capital city of Shaanxi Province. As the city expands and land resources become increasingly scarce, some Category 2 service areas may gradually move closer to the city center, and the surrounding land will be utilized more efficiently. Therefore, proper future planning should be carried out for some Category 2 service areas to adapt to future changes.

Category 3 highway service areas include Hancheng Service Area, Qinling Service Area, Baihe Service Area, Xiangshichuan Service Area, Zhidan Service Area, and 47 other service areas. Category 3 service areas generally have lower values across various classification indicators compared to the previous two categories. Most service areas in this category have zero amenities and companies, and they are located far from the city center in remote geographical conditions, locational advantages, transportation conditions, distribution of surrounding industries, policies, and planning of Qujiang Service Area and Hancheng Service Area.

In summary, the K-Means clustering algorithm is the most suitable for classifying highway service areas in Shaanxi Province, followed by the hierarchical clustering algorithm. This study provides a new perspective for regional planning and management of highway service areas, which is of great significance for regional economic development.
geographical areas. The regional economy and population levels are also relatively low. These service areas have not utilized the surrounding land effectively, and they can only meet basic needs such as dining, refueling, and accommodation for incoming travelers. These service areas are less influenced by urban expansion, and there is unlikely to be significant change in the utilization of surrounding land in the future.

6. Conclusions

This study classified highway service areas based on their surrounding land use characteristics using POI data, service area size, distance from the city center, regional economy, and population data as classification indicators. The clustering results in this study provide better interpretability and objectivity compared to previous research. The conclusions of this study on the classification of highway service areas are as follows:

1) By using PCA to reduce the dimensionality of surrounding POI types based on the data of highway service areas, two principal component factors were extracted with a cumulative contribution rate of no less than 85%. Amenities, companies, service area size, distance from the city center, regional economy, and population data were used as clustering indicators.

2) Different numbers of clustering categories (2 to 10) were selected, and the sum of squared errors within clusters was calculated for each number of categories. It was found that when the number of categories was set to 3, a significant inflection point in the sum of squared errors within clusters was observed, meeting the requirements of clustering search range. Therefore, the number of clustering categories was determined to be 3.

3) Using the data of 95 highway service areas in Shaanxi Province as an example, under the same input variables and iteration times, K-Means algorithm, FCM algorithm, and hierarchical clustering algorithm were compared using the CH index, silhouette coefficient, and DB index. The results showed that the K-Means clustering algorithm performed the best, and its results were chosen as the final classification results.

This study examines indicators that were equally important during the clustering process. However, in practical scenarios, different influencing factors are likely to have varying degrees of impact on the classification of highway service areas. Future research will incorporate additional classification metrics and investigate highway service areas based on their distinct levels of influence.

References


