Accessibility Cluster Analysis of Subway Station Based on Spatial Big Data—A Case Study of Dongcheng and Xicheng Districts in Beijing City

Lingmei Zhao *1,2, Zhenyang Wang 1,2, Mingliang Cao 1,2

1 Beijing Institute of Surveying and Mapping, 60 Nanlishi Road, Xicheng, Beijing, China, 100032
2 Beijing Key Laboratory of Urban Spatial Information Engineering, 60 Nanlishi Road, Xicheng, Beijing, China, 100032

Abstract. Subway is an important means of daily commuting in city life due to its punctuality and speed. Residential accessibility around subway station reflects the transportation convenience and connectivity between the necessary facilities which affecting residents' daily lives. Therefore, this study research on station accessibility factors by improving the walk-score model and establishing a multi-feature integrated transportation model that comprehensively considers the age difference based on spatial big data. Quantitative analysis was conducted on facility and station accessibility. Based on clustering algorithm considering three age groups, subway stations were classified into four types: mature, well-equipped, nurturing, and deficient. Using friendly characteristics, subway stations were categorized into three dominant age types. By integrating the analysis of accessibility, spatial layout, clustering differences and age-friendly characteristics, suggestions were proposed to improve station connectivity and supporting facility development.

1. INTRODUCTION

Accessibility is used to describe the ability of individuals to be easily and conveniently reach a certain location or utilize a particular service. The definition of accessibility was introduced by Hansen in 1959, defining it as the potential for mutual interaction opportunities between origins and destinations [1]. Ingram defined accessibility as the ease or difficulty of travel from an origin to a destination [2]. In 2007, American researchers introduced the concept of "Walk-Score," proposing a quantitative measurement method based on daily amenities, spatial layout, and factors such as intersection density and block length [3]. While, abundant research results have been accumulated through studies conducted by domestic and international scholars including studies on travel accessibility [4-6], community accessibility using indicators such as distance, time, or cost, infrastructure distribution and accessibility [7-9]. Studies on public transportation accessibility [10,11] involved utilizing network analysis, passenger flow models, and trajectory analysis to evaluate the service level and efficiency of public transportation [12,13]. Additionally, some studies have explored the relationship between accessibility and socio-economic impacts.

Referring to previous research findings, this study research on residential accessibility around subway station using quantitative model and clustering algorithm innovatively combining Delaunay Triangulation Algorithm. By improving walk-score model and leveraging spatial big data, the assessments on the convenience of residential neighborhoods facilities and accessibility to subway stations of different age group has been done. Based on the analysis result, planning and enhancement recommendations for subway stations are proposed for low-accessibility stations.

2. RESEARCH AREA AND MATERIALS

The study area is located in the Dongcheng and Xicheng districts of Beijing city having a total area of 92.5 km². It is the earliest developed urban area with abundant educational, medical, commercial, and tourist resources. The study area has an approximate population of 180 million residents, making it the most densely populated area in Beijing. Within the study area, there is Subway Line 1, which was the first subway line established in 1969. The study area has the highest coverage of subway stations and the largest number of subway lines in Beijing. The study focuses on 50 subway stations within the administrative boundaries, targeting study areas on residential neighborhoods within a 1000-meter service area surrounding each subway station (Figure 1).
As the foundation of service area and network analysis, the travel transportation model used in this study is based on the basic transportation road edges in Beijing. It integrates various walking features such as pedestrian overpasses, underground passages, internal roads within residential areas and pedestrian crosswalks. Based on the integrated data, the topological structure of the transportation network is established by identifying and connecting nodes and edges between road edges. Nodes represent the positions of intersections or connecting roads, while edges represent the connections between roads. Relevant attribute information, including road classification, sidewalk width, location, and height of pedestrian overpasses, is associated with the nodes and edges of the transportation network to form a comprehensive transportation network model.

Through web scraping techniques, subway station data was collected from OpenStreetMap. Residential neighborhood data was from census surveys. Residential neighborhood entrances/exits data and facilities data was from Amap (point of interest, POI). Population portrait data (point data from Baidu Huiyan) were obtained. These datasets serve as the foundational database.

3. RESEARCH METHODS

Walk-score is a traditional method for quantitatively analyzing street accessibility. In this study, based on walk-score method, the parameters of age group division, POI density, and category are introduced because the intersection density and road segment length alone cannot meet the analysis requirements. To serve the clustering needs, matrix reverse and standardization algorithms are added to enhance the model.

Spatial clustering algorithms are used to divide spatial discrete point data into clusters (clusters) with similar features or attributes, aiming to explore data features and conduct data analysis. The Delaunay triangulation is an important data structure and algorithm in computational geometry. It is used to connect a set of discrete points on a plane with non-overlapping triangles, generating continuous surfaces based on the discrete points. Therefore, in this study, the Delaunay triangulation is introduced based on the results of spatial clustering result. Discrete points have been transformed into surfaces where each surface represents a clustering type, making the clustering results more intuitive.

3.1. Facility Accessibility Model

Based on the walk-score model, more parameters were added to improve the model and to develop a computational model for the average facility accessibility including population age group segmentation index, POI density, and type richness index to improve the existing walk-score model(Figure 2).

The decay rate values were determined based on pedestrian travel habits. In general, pedestrians are less inclined to walk in areas with more intersections, longer roads, lower POI density, and fewer types of POIs, as these factors are associated with their walking preferences.
Step 1: Calculate the walking coverage circles of residential neighborhood entrances/exits within 500 meters, 1000 meters, and 1500 meters.

Step 2: Based on the differences in human physical capabilities of three age groups, different walking differentiation coefficients $\lambda$ were set for different age groups (elderly, middle-aged, and young).

$$\lambda = T \times (1 + \delta) \quad (1)$$

Among them, $T$ refers to the time required for the baseline walking over the same distance. $\delta$ is the age-related adjustment coefficient, which is a factor adjusted based on the differences in walking abilities among different age groups.

Step 3: Calculate the intersection density (number of intersections within each coverage circle divided by the circle's area), average block length, POI density (number of POI points within each coverage circle divided by the circle's area), and POI richness (total number of POI types within each coverage circle divided by the circle's area). The average block length represents the average length of each street segment within the circle.

$$L = \frac{L'}{n} \quad (2)$$

$L'$ is the total length of intersecting streets. $n$ is the quantity.

The obstruction decay rate is the sum of the decay rates for average intersection density, average block length, POI density index, and POI richness index.

$$y = \sum \text{intersection}_\text{index} + \text{roadlength}_\text{index} + \text{POIdens}_\text{index} + \text{POItype}_\text{index} \quad (3)$$

$\text{intersection}_\text{index}$ represents average intersection density decay rates, $\text{roadlength}_\text{index}$ represents average block road length decay rates, $\text{POIdens}_\text{index}$ represents POI density decay rates, and $\text{POItype}_\text{index}$ represents POI richness decay rates (Table 1).

### Table 1. Decay rates parameters

<table>
<thead>
<tr>
<th>Intersection density (per km²)</th>
<th>Decay rate(%)</th>
<th>Block road length(m)</th>
<th>POI density(per km²)</th>
<th>Decay rate(%)</th>
<th>POI richness(type/km²)</th>
<th>Decay rate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;77</td>
<td>0</td>
<td>&lt;120</td>
<td>&lt;607</td>
<td>0</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>58-77</td>
<td>1</td>
<td>120-150</td>
<td>&lt;500</td>
<td>1</td>
<td>&lt;13</td>
<td>2</td>
</tr>
</tbody>
</table>
Step 4: Calculate the shortest distance from each point to each type of facility within three coverage circles. Find the threshold length for the shortest distance and obtain the decay rate for the shortest distance, according to the POI weight table (Table 2).

<table>
<thead>
<tr>
<th>First-level category</th>
<th>Weight1</th>
<th>Second-level category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic living facilities (Transportation, convenience commercial, educational, healthcare facilities)</td>
<td>0.4</td>
<td>Old</td>
</tr>
<tr>
<td>Subway station</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Parking space</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Supermarket</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Daily commodities market</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Convenience store</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Restaurant</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Kindergarten</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>Primary school</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>Middle school</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>3A grade hospital</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>Health service station</td>
<td>0.4</td>
<td>0.3</td>
</tr>
<tr>
<td>Security and safety facilities</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Emergency shelter</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>Fire station</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Police station</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Living quality improvement facilities</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Park</td>
<td>0.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Cultural facilities</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>Sports facilities</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Leisure facilities</td>
<td>0.1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

The distance decay rate is calculated based on the Dijsktra’s shortest path algorithm using a Python program to compute the shortest paths. Over 5,000 residential neighborhoods and approximately 60,000 points representing 27 subcategories of POIs were calculated which involves approximately 360,000,000 calculations. The shortest distance threshold is divided into three intervals: less than 500m, 500-1000m range, and greater than 1000m range. The segmented function assigns path decay coefficients accordingly.

\[
\sigma(x) = \begin{cases} 
  x < 500, & \text{distance}_{index} = 0 \\
  500 \leq x < 1000, & \text{distance}_{index} = 0.25 \quad (4) \\
  1000 \leq x, & \text{distance}_{index} = 0.88 
\end{cases}
\]

\[
\text{Distance}_{index}(x) \text{ represents path decay rate, variable } x \text{ represents each shortest path, distance}_{index} \text{ represents the value of path decay rates for the three intervals.}
\]

Step 5: Calculate the facility accessibility score. \( \omega_0 \) represents the weight of the major category of indicators, \( \omega_1 \) represents the weight of the subcategory of indicators, \( \gamma \) represents the distance decay rate, \( \sigma \) represents the distance decay rate, and \( \lambda \) represents the walking differentiation coefficient.

\[
S_{score} = \omega_0 \times \omega_1 \times (1 - \gamma) \times (1 - \sigma) \times \lambda \quad (5)
\]

Step 6: According to Step 5, calculate facility accessibility score for the three population groups: elderly, middle-aged, and young.

3.2. Station Accessibility Algorithm

Apply Dijkstra’s shortest path algorithm to calculate the average distance from residential neighborhoods within a 1000-meter walking radius of subway stations to the stations themselves. Normalize the distance into \([0, 100]\) which is regard as the average station accessibility score.

3.3. Standardization and Normalization Algorithm

The cluster result is based on the theory that the closer the scores are to the origin and the better the facility accessibility and station accessibility. There for, the facility accessibility score matrix is reversed. Then, the reversed matrix and the original station accessibility matrix are standardized and normalized to the range \([0, 100]\)\(^{[14]}\). Finally, a normalized coordinate mapping is performed\(^{[15]}\).

(1) Array reverse function

```c
void reverse (double *array, int size) {
    int g = size;
```
int box[g]; int j = g-1; for(int i=0;i<g;i++) { box[i]=array[j]; j--; } For(int i=0;i<g;i++) array[i]=box[i];

(2) Data standardization
Let \( E \) represent the standardized matrix, where \( x_{ij} \) represents each element in the matrix. The standardized matrix is obtained through the standardization algorithm as follows:

\[
E = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}^2}
\]  

(6)

(3) Data Normalization
The standardized matrix is positively oriented by transforming it to the minimum value and then normalizing it to the [0, 100] interval. The function used for this purpose is:

\[
f(x) = \frac{\text{Max}(x) - x}{\text{Max}(x) - \text{Min}(x)} \times 100
\]  

(7)

(4) Normalized coordinate mapping
The normalized station accessibility scores of 50 subway stations are taken as the x-axis, and the facility accessibility scores are taken as the y-axis. The matrix data is mapped to a two-dimensional geographic space, where the bottom-left corner is (0,0) and the top-right corner is (100,100), forming the bounding rectangle for the two-dimensional data. This creates spatial scatter data.

3.4. Spatial clustering algorithm
Let’s assume dataset matrix \( S \) in which \( A_1 \) and \( A_2 \) are two points. Euclidean distance between the two points is:

\[
\text{Dis}(A_1, A_2) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \text{Dis}(A_{1i}A_{2j})}{m+n}
\]  

(8)

Where \( m \geq 1 \) and \( n \geq 1 \); when \( m = 1 \) and \( n = 1 \), it calculates the distance between two points.

Establishing Voronoi polygons based on the threshold between points. The definition of the distance threshold during the clustering process is:

\[
\varphi = \eta \times \text{Dis}(P_1, Z_2)
\]  

(9)

Where \( P_1 \) is the centroid or center point of the first category, \( Z_2 \) refers to the point farthest from \( P_1 \), and \( \eta \) is a constant with a range of \( 0 < \eta < 1 \). The calculation of the centroid within the class is:

\[
\text{midpoint}(Q) = \left( \frac{\sum_{i=1}^{m} x_i}{m}, \frac{\sum_{i=1}^{m} y_i}{m} \right)
\]  

(10)

Where \( m \) is the number of points or objects in a dataset \( Q \), and \( m \geq 1 \). \( i \) represents a point in the dataset \( Q \).

The technical roadmap for the developed program is as follows:

Step1: Input matrix \( S \) and clustering threshold \( \varphi \);
Step2: Establish Voronoi polygons based on dataset \( S \);
Step3: Combine the function \( \text{Dis}(A_1, A_2) \) to calculate the distance between each pair of points (or two classes) and place the resulting distances in the matrix \( D_C \);
Step4: Sort the matrix \( D_C \) and select the minimum distance value. If the minimum distance \( \text{min}(\text{dis}) < \varphi \) merge the two points (or two matrices) to form a new class \( W_i \);
Step5: Calculate the centroid of the new class \( W_i \) and add the coordinates of the centroid to the matrix \( S \);
Step6: Recursively repeat steps 3-5 until \( \text{min}(\text{dis}) < \varphi \), then stop.

4. ANALYSIS AND EVALUATION
The total population of residents around the subway stations is around 8.5 million. There are 18 stations serving a permanent resident population of over 200,000 people (Figure 3). The proportion of car owners in residential communities is approximately 31.76%, 64.19%, and 4.05% for the ranges of 10%-20%, 20%-30%, and above 30%, respectively. The proportions of elderly (60 years and above), middle-aged (30-60), and young (below 29) populations are 25%, 34%, and 41%, respectively. In terms of car ownership, around 70% of residential communities around the stations rely on public transportation, indicating a high probability of choosing subway stations for commuting.

Figure 3. The number of permanent residents covered within service areas of subway stations
The threshold for POI density within a 1km service area of residential neighborhoods is in the range of 19 to 607. 40% of the total number of neighborhoods have a POI density over 300 per km². POI category richness ranges from 5 to 13 with 89% of neighborhoods having 13 different types of facilities. Average shortest paths from residential neighborhoods to each subcategory of POIs range from 25 to 5400. The proportion of shortest paths less than 500 meters is 20%, while the proportion of paths within 500m-1000m is 67% and those above 1000m account for 13%. Average shortest paths from residential neighborhoods to subway stations is in the range of 2500~4300 with 15% of neighborhoods located within a 3000m radius.

In terms of the average facility accessibility scores for the elderly population, there are 68 residential areas with scores below 60, and approximately 59% of the residential areas have scores above 80, with around 480 areas scoring above 90. For the middle-aged population, there are 29 residential areas with scores below 60, approximately 60% of the areas have scores above 80, and there are 224 areas with scores above 90. For the young population, there are 44 residential areas with scores below 60, approximately 67% of the areas have scores above 80, and there are 550 areas with scores above 90. As for three age groups, the residential areas in the north and south have higher proportions than those in the east and west (Figure 4).

By calculating the average facility and station accessibility for three age groups and normalizing the scores to the range of [0, 100], clustering was performed using the 3.4 algorithm using three score thresholds: 35, 25, and 20 units. The resulting set of graphs corresponds to (a), (b), and (c) in Figure 5, where the horizontal axis represents the station accessibility and the vertical axis represents the facility accessibility. From the graphs, it can be observed that a lower value on the horizontal axis indicates easier access to nearby subway stations. Similarly, a lower value on the vertical axis indicates more comprehensive types of facilities in the vicinity of subway stations and easier access to these facilities. From graph (a) in Figure 5, obtained with a cluster size of 35 units, it can be seen that most stations are concentrated in the bottom-right corner. The mapping for these stations is closer to the region labeled ① in graph (a) indicating the presence of good facilities and convenient access to subway stations in the surrounding area, with relatively little variation in average indicators. In graph (a) zone ②, which includes Andingmen, Dongsi Shitiao, and Dongzhimen, it can be observed that the average facility distribution in the surrounding neighborhoods is uneven, resulting in poor facilities convenience. This may be due to inadequate supporting facilities or longer distances required to reach these facilities. In graph (a) zone ④, represented by the Tiananmen West Station, the average station accessibility score is relatively high, indicating higher costs and longer distances required to reach the subway station involving much more detours.

Figure 4. The average Facility Accessibility score for the old, middle-aged, young

Figure 5. Clustering analysis result using thresholds: 20(left), 25(middle), and 35(right)
Graphs (b) and (c) were generated using cluster thresholds of 25 and 20 units, respectively. The results showed a more detailed division of regions compared to graph (a). From zone (b) ①, it can be observed that the stations in this area perform better in both indicators. This zone further splits into (c) ① and (c) ②, with (c) ② exhibiting the highest level of facility completeness and convenience. Subway stations such as Chongwenmen, Xisi, and Jingtai are located in this zone. In (c) ①, the Ping'anli subway station stands out with the best values for facility completeness, facility convenience, and distance to the subway station among all the stations.

According to the clustering results based on graph (a), the subway stations are classified into four categories: mature type, well-developed type, nurturing type, and deficient type. Among the 18 subway stations with a large permanent population over 200,000, 13 stations belong to mature type with optimal facility convenience and station accessibility. Three stations are classified as well-developed type, including Chegongzhuang, Caishikou, and Guangqumen Nei. Two stations are classified as emerging type, namely Ciqikou and Xinjiekou indicating lack of facilities or too far away from the station. (Table 3):

<table>
<thead>
<tr>
<th>Station Type</th>
<th>Station Name</th>
<th>Type Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mature</td>
<td>Dongsi, Dengshikou, Zhangzizhonglu, Beixinqiao, Yonghegong, Hepinglibei, Chongwenmen, Beijing, Wangfujing, Dongdan, Nanluoguxiang, Zhushikou, Qiaowan, Jingtai, Nanlishiul, Fuchengmen, Changchunjie, Xuanwumen, Taoranting, Lingjinghutong, Xisi, Pinganli, Shichahai, Wanzi, Guanganmen, Daguanying, Hufangqiao</td>
<td>Best facilities and convenient access to subway stations</td>
</tr>
<tr>
<td>Well-developed</td>
<td>Tiananmendong, Tiantandongmen, Qianmen, Guangqumen, Yongdingmenwai, Chegongzhuang, Hepingmen, Caishikou, Dongwuyuan, Chegongzhuang, Guloudajie, Xizhimen</td>
<td>Good facilities and convenient access to subway stations</td>
</tr>
<tr>
<td>Nurturing</td>
<td>Ciqikou, Xidan, Xinjiekou, Beihaibei, Andingmen, Andelibeijie</td>
<td>Low facility accessibility or low station accessibility</td>
</tr>
<tr>
<td>Deficient</td>
<td>Fuxingmen, Dongzhimen, Dongsishitiao, Tiananmenxi</td>
<td>Low facility accessibility and low station accessibility</td>
</tr>
</tbody>
</table>

Additionally, the normalized mapping results can also reflect the friendliness of facility convenience for different age groups. A higher indicator value of indicates less friendliness for the particular group. From Figure 6, we can see that the South Lishi Road station provides the most friendliness services and facilities for the elderly population. The surrounding area of this subway station has a rich variety and total quantity of facilities such as bus stops, top-tier hospitals, health service stations, and park green spaces, with low travel costs to reach these facilities. However, the number of office buildings, primary and secondary schools, and sports and leisure facilities is relatively small. Similar stations in this category include Chongwenmen, Xisi, and Chegongzhuang station. The Daguanying station has high convenience for accessing primary and secondary schools in the surrounding residential area, and also has abundant medical facilities, making it relatively friendly for both young and elderly populations. This category also includes the Yonghegong station. The Beijing station has a diverse range of facilities in its vicinity, such as bus stops, medical facilities, health service stations, and park green spaces. However, it lacks convenience stores, shopping malls and bookstores. The surrounding hutong parking lots are relatively scarce. While, the cost of walking or using public transportation to reach these facilities is low. These reasons make Beijing station relatively friendly for middle-aged and elderly populations, followed by young people.

Additionally, park green spaces provide a place for leisure and exercise for the elderly. The abundance and high walkability of these facilities enable the elderly to easily meet their needs. Furthermore, the educational facilities,

**Figure 6. Typical age characteristic instance**

The Beixinqiao station provides a friendly facility allocation for the elderly population. It has daily service facilities such as bus stops, medical facilities, and health service stations, which are crucial for the elderly.
entertainment venues, and recreational facilities in the vicinity of Beixinqiao subway station are also friendly for young people, as shown in Figure 6. However, for the middle-aged population, the convenience of parking facilities and sports facilities in the vicinity of Beixinqiao station is relatively low, making it difficult to meet their travel needs. On the other hand, Beihai North station shows relatively balanced facility convenience for each demographic group.

5. CONCLUSION

This article is based on a quantitative method studying on subway accessibility. Introduce age weight, POI density and category parameters to improve traditional walk-score model. By using segmented functions to assign weights, the convenience of facilities in residential areas around subway stations is quantitatively evaluated. Theoretically, the innovatively integration of spatial clustering algorithms and the Delaunay triangulation spatial shape creation method generates block-based classification results, intuitively dividing station categories and evaluating accessibility indicators.

The preliminary results of this study indicate a consistent distribution between the permanent population around stations and the clustering results of station accessibility. Among the 18 high-density population stations with a permanent population of over 200,000, mature and well-developed subway stations account for 89%. Ciqikou and Xinjieqiu stations are classified as emerging type, indicating a need for improvement in facility convenience. Regarding the residential areas surrounding the stations, the facility convenience in the north and south parts is higher than that in the east and west parts about the study area, highlighting the need to prioritize supplementary facilities in station upgrades. From the perspective of age-friendliness, there is a limited number of elderly-friendly stations. Moreover, there is an apparent imbalance in the distribution of facilities for three age groups Which shows a necessity to optimize and adjust the facility categories. By improving the level of station accessibility based on above findings, the well-being of residents in the surrounding areas can be enhanced, contributing to the construction of a livable city based on communities.

ACKNOWLEDGMENT

This paper is the result of the research project founded by Beijing Key Laboratory of Urban Spatial Information Engineering, named ——Research on Quantitative Evaluation Model of Slow Traffic System in Historical and Cultural Districts in Beijing Based on Spatiotemporal Big Data.

REFERENCES