Trip purpose inference and spatio-temporal characterization based on anonymized trip data - an empirical study from dockless shared bicycle dataset in Xi'an, China

Jianwei Wang  
Chang'an University

Mengyan Cui  
Chang'an University

Hao Yang  
Department of Civil and Environmental Engineering

Sen Fu  
Shanghai Radio Equipment Research Institute

Xuxu Liu  
Chang'an University

Xin Fu (✉️ fuxin@chd.edu.cn)  
Chang'an University

Research Article

Keywords: Trip activity inference, Gravity model, Bayesian rules, DBSCAN, Dockless shared bicycle, Anonymized trip data

Posted Date: May 16th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2889533/v1

License: ☑️ This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License

Additional Declarations: No competing interests reported.
Trip purpose inference and spatio-temporal characterization based on anonymized trip data - an empirical study from dockless shared bicycle dataset in Xi'an, China

Jianwei Wang1,2, Mengyan Cui1, Hao Yang3, Sen Fu4, Xuxu Liu1, Xin Fu1,2,*

1. College of transport Engineering, Chang'an University, Xi'an 710064, China
2. Engineering Research Center of Highway Infrastructure Digitalization, Ministry of Education, Chang'an University, Xi'an 710064, China
3. Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195-2700, USA
4. Shanghai Radio Equipment Research Institute, No.1555 Zhongchun Road, Minhang District, Shanghai, 201109, China

*Corresponding author: Xin Fu, fuxin@chd.edu.cn

Abstract

The development of IoT and big data technology has generated massive personal travel data. While most of the data is anonymized, it still provides more possibilities for analyzing individual travel behavior. It is worthwhile to keep exploring how to more accurately infer travel purposes based on trajectory data, smart card fare data, or shared bicycle data. In this paper, an improved research framework is proposed for travel purpose inference by applying the gravity model, Bayesian criterion and spatial clustering method. The gravity model and Bayesian rule are used to calculate the probability of users traveling to nearby POIs, and the clustering algorithm is used to identify the locations regularly visited by users. Through the identification of maximum probability POI and regular trips, different travel purposes represented by POI can be distinguished. The results showed that the identification of regular trips could verify and complement the recognition of trip purposes by POIs. Using Xi'an city of China as an example for the study, the results show that regular trips accounted for 32% of the week's trips. Of these regular trips, 30% of the trips are most likely to
go to malls, restaurants, hospitals, or recreational areas. While these POIs often represent trip purposes such as shopping, eating, medical care, and entertainment, the purpose of these trips is more likely to be commuting due to the regularity of the trip. Ultimately, there are 10% of the trips reflect different trip purposes for the same POI. And there is a distance worth noting. 70% of the users have a parking error of fewer than 100 meters when going to the exact location. This distance can provide a reference for the study of bicycle commuting. Finally, we analyzed the different trip purposes' spatial and temporal distribution characteristics. The analysis of the spatial and temporal distribution can provide suggestions for the operation of bike-sharing enterprises and the management of regulators.

Keywords: Trip activity inference, Gravity model, Bayesian rules, DBSCAN, Dockless shared bicycle; Anonymized trip data

1. Introduction

Travel purpose inference is one of the hot topics in the field of urban transport, and this research is continuing with the emergence of a huge amount of travel data, especially the collection of shared travel data. Bike-sharing has significantly impacted urban travel as an emerging transport technology that integrates the Internet with traditional transport. At the same time, the travel record data generated in shared travel has been noticed by many research scholars. Scholars have used these data to conduct research on many aspects, including demand and dispatch (Yang et al, 2020; Zhang et al, 2021), role and impact (Li et al, 2020; Saberi et al, 2018), behavior and patterns (Si et al, 2020; Jia et al, 2019), Spatio-temporal characteristics (Maas et al, 2020), and so on. These studies affirm the importance of bike-sharing and offer many suggestions for its development.

In recent years, considerable studies have been using bike-sharing data for trip purpose inference (Li et al, 2020; Xing et al, 2020; Li et al, 2021). Previously, most of the data used for travel purpose inference were smartphone-based travel surveys (Ermagun et al, 2017; Gx et al, 2016), smart card fare data (Kusakabe et al, 2014; Alsger et al, 2018), taxi trajectory data (Furletti et al, 2013; Gong et al, 2015), and so on. However, it should be noted that most of these travel record data are anonymized and
have been removed the private information such as personal identity before being used in the study. Therefore, these anonymous data have limitations when studying travel purposes and behavior. Compared with smart card fare data, the travel location recorded by shared bicycle data is more accurate. Compared with taxi trajectory data, the shared bicycle data can record the travel chain of users by user ID. Therefore, the bike-sharing dataset is well suited for studies on the purpose of travel. However, all these datasets share a common drawback: they do not contain land use information. Although these data accurately record the location of people or vehicles, these locations often contain only latitude and longitude information. Therefore, offline datasets like land use or POI are often used as a supporting data source for travel purpose inference. POI data can be crawled through online maps. For instance, the Google Places application programming interface (API) (Ermagun et al., 2017) searches POIs according to input place names or locations and a search radius. The API returns detailed information about that place (category, location, opening hours, and so on) or all detailed information about POIs within the search area of that location.

This study used a one-week dataset of Haro Shared Bicycle (First market share in China) in Xi'an. The POI data from AMAP (One of the most comprehensive maps in China) was used to combine the land use information. First, the POIs in the ride-end buffer are defined as candidate POIs. The probability of the user traveling to candidate POIs was calculated using the Gravity model and Bayesian criterion. Gong et al. (2015) and Li et al. (2021) have used the Gravity model and Bayesian criterion to infer the trip purpose. The factors they consider when calculating the probability of traveling to a POI include 1) the distance between the parking point and the POI point, 2) the time at which this trip occurs, and 3) the type of land use near the parking point. They usually assume that the POI corresponds to a single travel purpose, such as the purpose of the mall for shopping, etc. Ross et al. (2022) offer a different perspective. He thinks that the travel purpose corresponding to POIs varies with time. Based on this, we further introduce regularity to identify POIs corresponding to different trip purposes. For example, when a user arrives at a mall every morning at 8:00 a.m., he
may be working here instead of shopping, and when he arrives near a mall every afternoon at 6:00 p.m., his home address may be located nearby. We identify such regularities by spatial clustering and indicator filtering. Finally, the maximum probability POI, regularity identification results, and arrival time are inferred through specific rules for the final travel purpose. This paper has two main contributions: 1) a rule to distinguish different travel purposes corresponding to POIs is proposed; 2) a spatial distance threshold suitable for bicycle commuting is proposed. We believe that such improvements will enhance the potential and accuracy of anonymous data such as shared bicycle order data in the field of travel behavior.

The rest of the paper is organized as follows. Section 2 presents a literature review on the inference of sharing bicycles purpose. The third section shows the datasets used in this paper, the bicycle sharing order dataset and the POI dataset. Section 4 describes the overall framework of trip purpose inference. The fifth part analyzes the results, and the sixth part concludes the paper with an outlook.

2. Literature Review

Purpose speculation and analysis in travel behavior studies have been critical of urban transport research, especially when increasingly rich individual travel data. The massive amount of data brought by the Internet of Things era has significantly boosted the research on travel behavior. Studies such as travel mode choice (Liu et al., 2015; Zhang et al., 2021; Pieroni et al., 2021), trip chain (Kim et al., 2022; Huang et al., 2021; Bautista-Hernández et al., 2022), and travel purpose recognition (Ross-Perez et al., 2022; Xie et al., 2018; Li et al., 2020) all belong to the field of travel behavior. In the past, people could only obtain information on travel modes and travel purposes through traditional surveys (Minh et al., 2020). Traditional travel surveys are labor-intensive and time-consuming (Minh et al., 2020). The popularity of location-aware devices has made easy access to this information possible. Researchers have started to study how to obtain information related to travel behavior through GPS data.

Travel mode choice and travel chain recognition can provide ideas for travel purpose recognition. Pieroni et al. (2021) used the DBSCAN algorithm to infer the
residence location of bus card users. He found that low-income residents tend to start their first trips from 5-7 a.m., and those with higher incomes start from 7-9 a.m. Zhang et al. (2021) used smart card data to examine the travel frequency, travel distance, the characteristics of activity spaces for disadvantaged groups using transit cards, and the regularity/randomness of human activities. He divided the city into grids and calculated the number of arrivals of each passenger in each grid to determine the anchor point (the location of the main activity). Liu et al. (2015) used the cab trajectory data, built spatially embedded networks to model intra-city spatial interactions, and analyzed the city's spatial structure in conjunction with land use.

Most of the current studies on travel purpose inference use real data such as taxi trajectory data (Gong et al, 2015) and smart card data (Alsger et al, 2018). Gong et al. (2015) constructed a two-layer inferential model using taxi data. The first layer uses K-means to cluster drop-off time, walking distance, and trip distance respectively and selects the intersection term of the clustering results as the final destination. The second layer identifies possible pairs of return trips by rules. Alsger et al. (2018) uses smart card data. A rule-based modeling approach is deployed to infer passengers' trip purposes based on spatial, temporal, and frequency contributions. Still, some scholars have also conducted travel purpose inference through new traffic survey data. Xiao et al. (2016) developed a cell phone application where respondents' latitude and longitude location information, speed, etc., are recorded every few seconds. They also record respondents' contact information, personal characteristics, and family characteristics. After the inference is completed, the respondent is contacted to verify the inference results.

Methodologically, research on travel purposes can be broadly classified into three categories, machine learning-based (Krause et al, 2019), rule-based (Furletti et al, 2013), and probability-based (Gong et al, 2015). Among them, probability-based approaches are the most common. Probability-based methods aim to calculate the probability of users going to different POIs. The methods for calculating probabilities can be divided into two main categories: polynomial models and Bayesian criteria. Chen (2010) introduced a polynomial logit model to detect travel modes and infer trip
purposes in the complex urban environment of New York City and divided the purposes into home-based and non-home-based groups. Similar to the study by Chen (2010), Usyukov (2017) developed a polynomial model that uses the start time and duration of activities to distinguish work activities from other activities. Wang (2017) used probability distributions (polynomial and Dirichlet) first to classify POIs based on functional similarity into 10 themes (e.g., shopping, university, office) and then assigned themes to the starting point and destination of each trip.

Li (2021), Furletti (2013), Gong (2015), and Antonio (2022) combined the gravity model with Bayes' rule to infer travel purposes. Furletti (2013) proposed temporal and spatial rules to select a set of cars corresponding to potential locations for each stopping point. Only locations within a buffer zone with a radius of 1000 m and open at the time of parking were retained. Subsequently, the probability of each location is estimated based on a gravity model. The probability of the target location is maximized. Gong (2016) develops Spatio-temporal rules combined with distance decay effects and Bayes' rules for POI attractiveness (i.e., service capacity, size, and reputation) to create a probability function for accessing each POI. Li (2020) considers the service capacity of POIs based on previous work. The service capacity of POIs was calculated from TUD and AOI data. Antonio (2022) considered the different purposes represented by the POI. He distinguishes between different purposes through the time of the day.

Few studies have been conducted specifically on the trip purpose of shared bicycles. Xing et al. (2020) classified the trip origins and destinations of shared bicycles into five categories by K-means clustering. Li et al. (2020) applied the Dirichlet multiple regression topic models (DMR model) to identify the trip purpose of bicycle trajectories, considering both arrival time and drop-off location. Based on the obtained trip purposes, they analyzed the bicycle accessibility to restaurants and hospitals in downtown Shanghai. Li et al. (2021) inferred the trip purposes by a probability-based approach using a gravity model and Bayesian criterion. They considered the distance to the POI, travel time, type of land use near the destination, and different trip purposes corresponding to the POI.
In summary, probability-based inference methods for trip purposes are the most commonly used and interpretable. Therefore, we make further improvements to the study of Li (2021) and Ross-Perez (2022). We further refine the different travel purposes corresponding to POIs. This idea is inspired by Pieroni (2021), who identifies anchor points by methods such as DBSCAN and grid arrival frequency. We will identify orders with regularity in time and space by DBSCAN and metric filtering. We consider orders with Spatio-temporal regularity to travel for commuting. If regular orders have the highest probability of traveling to the mall, the user probably works or lives nearby, rather than shopping. Next, the travel time is combined to distinguish whether the regular order will work or go home.

3、Study area and data description

In this paper, the city of Xi'an (Located in Shaanxi Province, China, which is the only national central city in western China) is chosen as a case study for empirical research. As of October 2021, there are three bike-sharing companies in Xi'an city, namely Haro, Meituan, and Qingcheng. And Haro is the company with the highest market share. In September 2020, Haro's average daily usage reached 600,000 in Xi'an. This paper use Haro's bike-sharing order data to analyze travel regularity and travel purpose. This dataset contains all orders from September 14 to September 20, 2020, totaling 2.85 million orders. Each order contains the following fields: user ID, trip start and end time, trip start and end latitude and longitude, ride time and ride distance, as shown in Table 1. The data pre-processing eliminates some missing data and abnormal data, where missing data refers to orders that are missing specific fields, and abnormal data refers to ride time less than 1 minute, ride distance less than 50 meters, etc.

<table>
<thead>
<tr>
<th>User ID</th>
<th>Start time</th>
<th>End time</th>
<th>Start latitude</th>
<th>Start longitude</th>
<th>End latitude</th>
<th>End longitude</th>
<th>Ride time(h)</th>
<th>Ride distance (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>e7ac39a16d93041129e</td>
<td>2020/9/1 4 7:25:05</td>
<td>2020/9/1 4 7:31:07</td>
<td>108.94</td>
<td>34.316</td>
<td>108.937</td>
<td>34.3148</td>
<td>0.10027</td>
<td>714</td>
</tr>
<tr>
<td>b2419a184ff343</td>
<td>2020/9/1 4 7:25:05</td>
<td>2020/9/1 4 7:31:07</td>
<td>7328</td>
<td>0379</td>
<td>6201</td>
<td>0157</td>
<td>7778</td>
<td>714</td>
</tr>
</tbody>
</table>
In addition, we use POI data to associate trip purpose with activity location attributes. Each point of interest is a row of records containing fields such as latitude and longitude, address name, primary, secondary, tertiary, and quaternary classification. In previous studies on the travel purposes of taxis and shared bicycles, the travel purposes are roughly divided into Home, Work, Transfer, Dining, Shopping, Recreation, Schooling, and Medical, as shown in Table 2, so these eight categories are also used as travel purposes in this paper.

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Method</th>
<th>Travel purpose classification</th>
<th>Number of categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shuhui Gong</td>
<td>2019</td>
<td>Clustering</td>
<td>Work, Home, School, Restaurant, Shopping, Entertainment, Medical</td>
<td>7</td>
</tr>
<tr>
<td>Barbara Furletti</td>
<td>2013</td>
<td>Bayes' Rule</td>
<td>Food, Daily shopping, Shopping, Education, Leisure, Service</td>
<td>6</td>
</tr>
<tr>
<td>Chao Chen</td>
<td>2017</td>
<td>Bayes' Rule</td>
<td>Work, Schooling, Recreation, Shopping, Others</td>
<td>5</td>
</tr>
<tr>
<td>Li Gong</td>
<td>2015</td>
<td>Bayes' Rule</td>
<td>Work, Home, Schooling, Transfer, Dining, Shopping, Recreation, Schooling, Medical, Lodging</td>
<td>9</td>
</tr>
<tr>
<td>Yingying Xing</td>
<td>2020</td>
<td>Clustering</td>
<td>Company, House, Food, Transport, Shopping</td>
<td>5</td>
</tr>
<tr>
<td>Sharing Li</td>
<td>2020</td>
<td>Bayes' Rule</td>
<td>Work, Home, Schooling, Transfer, Dining, Shopping, Recreation, Schooling</td>
<td>9</td>
</tr>
<tr>
<td>Aoyong Li</td>
<td>2020</td>
<td>Theme Model</td>
<td>Work, Home, Dining, Shopping, Schooling, Recreation</td>
<td>6</td>
</tr>
</tbody>
</table>

The pre-processing of POI data includes eliminating missing data and associating POI data with the purpose of the trip according to the tertiary classification. The correspondence table between POI data and travel purposes is shown in Table 3. The main POI types are the tertiary classification of POI data. For some POIs, they represent three types of travel purposes. We will distinguish these three travel purposes according to the regularity of travel and travel time. Antonio (2022) distinguishes the different purposes represented by POIs according to different times of the day. Based on this, we introduce travel regularity. For food, entertainment, mall,
and hospital type POIs, when users regularly travel to these POIs in the morning and evening peak, users may work or live near these POIs.

Table 3 Comparison of travel purpose and POI

<table>
<thead>
<tr>
<th>Main POI types</th>
<th>Regular trips in the Morning peak</th>
<th>Regular trips in Evening Peak</th>
<th>Rest trips</th>
</tr>
</thead>
<tbody>
<tr>
<td>Companies, factories, agriculture, forestry, fishery, and animal husbandry bases</td>
<td>Work</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subdivision, village, gate number, business residence</td>
<td>Home</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subway, Bus</td>
<td>Transfer</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Schools, training institutions, research institutions</td>
<td>School</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese food, Western food, dessert, coffee shop</td>
<td>Work</td>
<td>Home</td>
<td>Food</td>
</tr>
<tr>
<td>Gymnasiums, theatres, recreational areas, parks, scenic spots</td>
<td>Work</td>
<td>Home</td>
<td>Recreation</td>
</tr>
<tr>
<td>Malls, stores, markets</td>
<td>Work</td>
<td>Home</td>
<td>Shopping</td>
</tr>
<tr>
<td>Hospitals, Clinics</td>
<td>Work</td>
<td>Home</td>
<td>Medical</td>
</tr>
</tbody>
</table>

4. Methodology

4.1 Overview of the framework

The overall research framework is shown in Figure 1. The framework consists of two main parts. The first part is the inference of trip purpose based on the Bayesian criterion. The second part is the identification of regular trips based on spatial clustering. The gravity model combined with the Bayesian criterion has been widely used in inferring travel purposes. Li (2021), Furletti (2013), and Gong (2015) have used this method to infer the travel purpose, and the principle is to infer the POI where the cyclist is most likely to go. However, the above studies usually ignore the different purposes that POIs represent. So we identify the regularity of travel to distinguish different purposes of the same POI. The detailed distinction rules are shown in Table 3. If some users are regularly going somewhere in the morning peak or evening peak, the purpose of their trip is then defined as work or home. The regular trips were recognized by spatial clustering and indicator screening.
4.2. Models for inferring the trip purpose

The first part of the model is a probability-based inference of travel purpose, using the gravity model and Bayes' rules. The gravity model is used to calculate the spatial attraction between candidate POIs and drop-off points. Many studies have applied the gravity model to infer the attraction between two places. The first part can be divided into four steps: determining the maximum walking distance, identifying candidate POIs, calculating the spatial attractiveness of candidate POIs, and calculating the probability of candidate POIs being selected by combining Bayesian criteria.

4.2.1 Maximum walking distance and candidate POIs

For different modes of transport, the acceptable walking distance for users is influenced by various factors. In studies on public transport, the maximum walking distance often uses the minimum interval between bus stops. Zhao (2011) used 500 m as the maximum walking distance for public transport, Gong (2015) used 200 m as the maximum walking distance for taxis, and Li (2021) also used 200 m as the maximum walking distance for shared bicycles. Gong (2015) and Li (2021) obtained the maximum walking distance by calculating the cumulative percentage of the number of POIs within 10 m to 500 m of the drop-off point. The results show that the cumulative percentage gradually decreases from 200 m, indicating that the drop-off...
point's activity range is likely to be within this 200 m. Since this study uses a shared bicycle dataset, 200 meters was chosen as the maximum walking distance. The candidate POIs are all the POIs within the buffer constructed by taking the maximum walking distance as the radius and the drop-off point as the circle's center. This study uses the Generate Near Table tool of ArcGIS to obtain the candidate POIs. The Near Table is shown in Table 4. IN_FID indicates the parking point number, and NEAR_FID indicates the number of POI points in the buffer. Longitude, Latitude, Purpose, and NEAR_DIST indicate the latitude and longitude of the POI point, the corresponding travel purpose, and the distance from the parking point.

<table>
<thead>
<tr>
<th>IN_FID</th>
<th>NEAR_FID</th>
<th>Longitude</th>
<th>Latitude</th>
<th>Purpose</th>
<th>NEAR_DIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>352925</td>
<td>108.93354</td>
<td>34.273605</td>
<td>transfer</td>
<td>134.35154</td>
</tr>
<tr>
<td>0</td>
<td>348452</td>
<td>108.93378</td>
<td>34.273501</td>
<td>shopping</td>
<td>121.817751</td>
</tr>
<tr>
<td>0</td>
<td>348081</td>
<td>108.932968</td>
<td>34.274731</td>
<td>transfer</td>
<td>182.916414</td>
</tr>
<tr>
<td>0</td>
<td>347580</td>
<td>108.933052</td>
<td>34.275085</td>
<td>transfer</td>
<td>193.220099</td>
</tr>
<tr>
<td>0</td>
<td>345072</td>
<td>108.934873</td>
<td>34.273485</td>
<td>food</td>
<td>75.526651</td>
</tr>
<tr>
<td>0</td>
<td>344508</td>
<td>108.93535</td>
<td>34.275452</td>
<td>food</td>
<td>150.490705</td>
</tr>
<tr>
<td>0</td>
<td>342496</td>
<td>108.934861</td>
<td>34.273487</td>
<td>food</td>
<td>75.259725</td>
</tr>
<tr>
<td>0</td>
<td>342464</td>
<td>108.933181</td>
<td>34.273235</td>
<td>food</td>
<td>183.863136</td>
</tr>
</tbody>
</table>

4.2.2 Spatial attraction measurement

Spatial attractiveness refers to the strength of attraction of POI to users. Furletti (2013) and Gong (2015) define this strength of attraction as spatial attractiveness and calculate spatial attractiveness based on the gravity model. Spatial attractiveness varies depending on distance, time, and type of land use. Furletti (2013) argues that spatial attractiveness is inversely proportional to the square of the distance between two points. However, Gong (2015) chose 1.5 times as the distance coefficient. He fitted the results of travel purpose with the questionnaire results, and the best fit was achieved when the coefficient of distance was 1.5 times. So we chose a distance coefficient of 1.5 times. Distance, time, and land use type are considered to calculate spatial attractiveness. The set of candidate POI points was defined as \{D1, D2...Di\} and the parking point is P. The following formula can express the spatial
attractiveness of the candidate POI to point P:

$$P(P | (Di,t)) \propto \frac{\rho_{\text{destination}(j)}}{d(P,Di)^{1.5}}$$  \hspace{1cm} (1)

Where $\rho_{\text{destination}(j)}$ is the ratio of the number of jth type of destinations in the buffer to the number of all POIs in the buffer, see formula 2.

$$\rho_{\text{destination}(j)} = \frac{\text{number of destination}(j)}{\text{number of all destinations}}$$  \hspace{1cm} (2)

4.2.3 Visit probability of candidate POIs

Next, the spatial attractiveness calculated by the formula (2) is used to calculate the probability of the user going to each POI within the buffer. Assume that the rider is not likely to go to a POI outside this buffer. According to the Bayes’ rule, the probability that the user ends at time t at point P and goes to POI (Di) can be expressed as formula (3).

$$P(Di | P,t) = \frac{P(P,Di,t)}{P(P,t)} = \frac{P(P | (Di,t)) \cdot P(Di | t) \cdot P(t)}{P(P,t)}$$  \hspace{1cm} (3)

Where $P(P | (Di,t))$ denotes the probability when a user heads to point Di, the probability of stopping at point P. It is proportional to the attraction between the parking spot and the POI. The final formula was obtained by substituting formula 1 into formula 3. It is shown in formula 4.

$$P(Di | P,t) = \frac{\rho_{\text{destination}(j)}}{d(P,Di)^{1.5}} \cdot P(Di | t) \cdot P(t)$$  \hspace{1cm} (4)

4.3 Models for inferring regularity

Most studies have defined regularity as temporal and spatial regularity. Time regularity is usually identified through metric screening, and standard metrics contain trip frequency (Zhou et al, 2021), departure time difference, or arrival time difference (Long et al, 2012). The methods to identify spatial regularity are K-means or DBSCAN clustering (Ma et al, 2017; Zhou et al, 2021). DBSCAN is very suitable for identifying spatial regularity because of its internal parameters. It can set the
maximum distance within a cluster and the minimum number of points within a
cluster. DBSCAN can discover the places where users go most frequently with these
two parameters. We will cluster the start and endpoints of each user's morning and
evening peaks separately. By clustering all the start and endpoints of each user's
morning and evening peaks separately, it is possible to obtain a cluster that the user
visited frequently. The model for regularity identification can be divided into three
steps: extracting morning and evening rush hour orders, clustering individual users'
starting and ending points (identifying spatial regularity), and eliminating data with
significant time differences within clusters (identifying temporal regularity).

4.3.1 Morning and evening peaks extraction

The morning and evening rush hours are essential for identifying the regular trips.
In Sun's (2015) study, he approximated the peak hours of public transport commuting
by analyzing the time of people's first and last bus trip. In addition to that, many
scholars identify the morning and evening peaks through travel quantity distribution
per day. Therefore, we combine the two methods to determine the morning and
evening peaks.

Through these two methods, Xi'an's morning and evening rush hours are defined
as 6:00-10:00 and 17:00-20:00. The data was divided into fourteen categories based
on the number of morning and evening peak rides by users to make the analysis more
detailed. These fourteen categories are the groups of users who have ridden for 7 days,
6 days...1 day in the morning peak or evening peak, as shown in Figure 2. The
following analyses were performed based on the classification.
4.3.2 Recognition of spatial regularity

There have been several attempts to use spatial clustering for destination extraction for GPS data. Pieroni (2021) used DBSCAN clustering for residence identification. In this paper, DBSCAN clustering is used to identify locations where users arrive more frequently during morning and evening peak periods, thus marking some trips with regularity. The clustering mechanism is shown in Figure 3. The order is considered to be spatially regular when both the starting and ending points are within a cluster respectively. As shown in Figure 3, trips 2, 3, and 4 are considered to be regular trips, while trips 1 and 5 are not regular trips.

The DBSCAN algorithm needs to determine two parameters in advance, the distance threshold (ε) and the size threshold (MinPts.) The DBSCAN algorithm first selects an arbitrary point to start visiting and identifies all other data points within the distance threshold (ε). If the number of neighboring points is at least MinPts (size threshold), then the data point under consideration becomes a new cluster. From previous studies, the distance threshold varies depending on the mode of transport. The study by Li (2021) used 1200 m as the distance threshold, while Long (2012) used 500 m as the distance threshold. However, their study was about public transport.
So we tested different distance thresholds and selected the best distance threshold by observing the recognition results. The size threshold is the number of points included in the category. It is similar to the frequency of travel. Most researchers set it to 2-4 trips per week. So we set the size threshold to 3.

![Figure 3 DBSCAN clustering schematic](image)

### 4.3.3 Recognition of temporal regularity

We defined a metric to filter temporal regularity based on spatial regularity identification results. This metric is named the intra-category arrival time extreme difference. This metric is for individual users. It is the result of subtracting the earliest and latest arrival times, shown in table 5. The user arrives at the destination on September 19 at the earliest and September 20 at the latest, and the difference between the earliest arrival time and the latest arrival time is 17 mins.

<table>
<thead>
<tr>
<th>Arrival time</th>
<th>Intra-class arrival time extremes difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020/9/14 6:49:31</td>
<td>31</td>
</tr>
<tr>
<td>2020/9/15 6:50:27</td>
<td>27</td>
</tr>
<tr>
<td>2020/9/16 6:49:32</td>
<td>32</td>
</tr>
<tr>
<td>2020/9/17 6:52:53</td>
<td>53</td>
</tr>
<tr>
<td>2020/9/18 6:53:06</td>
<td>06</td>
</tr>
<tr>
<td>2020/9/19 6:40:35</td>
<td>The earliest arrival day</td>
</tr>
<tr>
<td>2020/9/20 6:57:50</td>
<td>The latest day of arrival</td>
</tr>
</tbody>
</table>
5. Results and discussion

In this part, we first verify and analyze the parameters for identifying Spatio-temporal regularity, including the DBSCAN distance threshold and critical value of indicator screening. Then the recognition results with and without regular trips are compared and analyzed. By comparison, we find that regular trips can make the identification of travel purposes more accurate. Finally, the statistical and spatial-temporal characteristics of different travel purposes are analyzed based on the identification results of regular trips.

5.1 Regularity travel recognition

5.1.1 Spatial characteristics of regularity travel

This section focuses on the validation of the DBSCAN parameter distance threshold (ε). As the distance threshold decreases, the number of regular travelers obtained by clustering gradually decreases. When the distance threshold decreases to 100 meters, the decrease speed becomes slow. This suggests that more regular travelers park within 100 meters. Therefore, we believe that the appropriate distance threshold is 100 meters when applying DBSCAN to identify the regularity of bike-sharing trips, commuting, or residential workplace identification. The clustering results of DBSCAN at different distances are shown in Figure 5, with the horizontal axis indicating the distance threshold (ε) and the vertical axis indicating the proportion of regular trips to all trips. Different color dashes indicate different categories, which is the same as the classification in Section 4.3.1. Since the travel distance of bike-sharing is usually between 500 meters and 2000 meters, our distance threshold is set below 500 meters. In this paper, the distance thresholds are set to 500 m, 400 m, 300 m, 200 m, 100 m, 75 m, and 50 m.
Figure 4 Variation of recognition ratio under different clustering distances

From Figure 4, it can see that as both the number of days ridden decreases and the distance threshold decreases, the recognition ratio decreases. And the category with more riding days was less affected by the change in distance threshold. At whatever distance threshold, the proportion of regular trips to all trips can reach 70% for categories of 5, 6, and 7 days. And there is a significant drop in recognition ratio for 4 days and 3 days.

It can be noticed that at 100 m, an inflection point occurs in all folds. Before 100 meters, the recognition ratio decreases slowly; after 100 meters, the recognition ratio drops rapidly. Therefore, 100 meters is selected as the distance threshold in this study. The user's parking error will not exceed 100 meters when going to the exact location.

5.1.3 Temporal characteristics of regularity travel

This subsection mainly shows the results of the metrics screening. The distribution of the intra-category arrival time extreme difference before and after index screening is shown in Figure 5. The horizontal axis represents the different categories, and the vertical axis represents the time difference. It can be seen that 75% of the data in all categories were already less than 50 min before exclusion. So we excluded the data over 60 minutes. After exclusion, it is ensured that all the extreme differences are within 60 min and 75% of the extreme differences are within 20 min.
The time difference between regular trips of the same user within one week is controlled within one hour through index screening.

Figure 5 Intra-class arrival time extremes difference of data before and after exclusion

5.2 Inference of the trip purpose

The method using only the gravity model with Bayesian criterion inference was defined as Method 1, and the method combining the results of regular commuting identification is defined as Method 2. The percentages of weekday and non-weekday trips for each purpose resulting from the analysis of the two methods are shown in Table 6.

Table 6 Comparison of the results of the two methods

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Weekday</th>
<th>Weekday</th>
<th>Weekend</th>
<th>Weekend</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Method1</td>
<td>Method2</td>
<td>Method1</td>
<td>Method2</td>
</tr>
<tr>
<td>work</td>
<td>24.5%</td>
<td>28.8%</td>
<td>18.1%</td>
<td>21.0%</td>
</tr>
<tr>
<td>transfer</td>
<td>8.0%</td>
<td>8.0%</td>
<td>7.9%</td>
<td>7.9%</td>
</tr>
<tr>
<td>home</td>
<td>16.6%</td>
<td>22.6%</td>
<td>18.4%</td>
<td>23.0%</td>
</tr>
<tr>
<td>school</td>
<td>3.9%</td>
<td>3.9%</td>
<td>3.5%</td>
<td>3.5%</td>
</tr>
<tr>
<td>food</td>
<td>22.2%</td>
<td>16.0%</td>
<td>24.5%</td>
<td>19.9%</td>
</tr>
<tr>
<td>medical</td>
<td>6.0%</td>
<td>4.7%</td>
<td>5.6%</td>
<td>4.8%</td>
</tr>
<tr>
<td>recreation</td>
<td>1.7%</td>
<td>1.5%</td>
<td>2.0%</td>
<td>1.9%</td>
</tr>
<tr>
<td>shopping</td>
<td>17.0%</td>
<td>14.5%</td>
<td>20.1%</td>
<td>18.0%</td>
</tr>
</tbody>
</table>

With the improvement of method 2, the proportion of some purposes, such as work and home, has been increased, while the proportion of food, medical, recreation, and shopping has been decreased. Trips for work in malls, hospitals, and recreational facilities were identified by regularity. Different from Li(2020)'s study, the proportion of shopping we identified varied widely. Because our shopping places also include
some small stores. The following analysis is based on the results of method 2.

The travel distance and riding time for various travel purposes were analyzed in Table 7. Food has the shortest travel distance and riding time, while shopping has a slightly higher travel distance and riding time than other travel purposes, which is also consistent with Xing(2020)'s study.

<table>
<thead>
<tr>
<th>Trip purpose</th>
<th>Trip distance</th>
<th>Trip duration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>75%</td>
</tr>
<tr>
<td>work</td>
<td>1.49</td>
<td>1.92</td>
</tr>
<tr>
<td>transfer</td>
<td>1.52</td>
<td>1.97</td>
</tr>
<tr>
<td>home</td>
<td>1.44</td>
<td>1.85</td>
</tr>
<tr>
<td>food</td>
<td>1.43</td>
<td>1.85</td>
</tr>
<tr>
<td>school</td>
<td>1.00</td>
<td>1.20</td>
</tr>
<tr>
<td>shopping</td>
<td>1.52</td>
<td>1.96</td>
</tr>
<tr>
<td>medical</td>
<td>1.39</td>
<td>1.71</td>
</tr>
<tr>
<td>recreation</td>
<td>1.52</td>
<td>1.98</td>
</tr>
</tbody>
</table>

5.2.1 Time distribution characteristics

Regarding the characteristics of the time distribution of trips, two points are analyzed, the comparison of weekdays and weekends, and the distribution of peaks. The time distribution of different travel purposes is shown in Figure 6, and the 0-48 hours on the horizontal axis represent the 24 hours on Monday and Saturday. The characteristics of the time distribution for different travel purposes can be divided into two points as follows:

(1) From the comparison of weekdays and weekends: For Work, Transfer, Home, and School, the weekday trips are higher than the weekend trips. But for Food and Shopping, the weekday trips are lower than the weekend trips. And there is no significant difference between Medical and Recreation.

(2) From peak hours: For Work, Transfer, Home, School, and Food, have distinct peaks. Work's peak occurs at 8:00 on both weekdays and weekends. Home's peak occurs at 18:00 on weekdays, while the peak occurs at 20:00 on weekends. There are
two peaks for transfer at 8:00 and 18:00, both on weekdays and weekends. Transfer's peaks are similar to Work and Home. For school, there are four peaks, both on weekdays and weekends. The most prominent peak of school occurs at 8:00, the peak value at 12:00 and 14:00 is lower than 8:00, and the lowest peak appears at 18:00. And these four peaks are the same on weekdays and weekends. The peak of school at 18:00 may be due to training institutions. For food, the three peaks occur at 8:00, 12:00, and 18:00. It is the same as the time to dine. For Shopping, Medical, and Recreation, there are no distinct peaks. For shopping, the number of trips is steady. Medical has a low point at 12:00. The recreation trips begin to grow at 10:00.

Figure 6 Time distribution characteristics of different travel purposes

5.2.2 Spatial and temporal distribution characteristics

The spatial distribution of the purposes is represented using the kernel density
method, as shown in Figure 7. Figure 7(a) and (c) show the spatial distribution of Work and Home. The hotspot region of work is centralized in the Southwest. The distribution of the home's hotspot region is more decentralized. At the same time, there is still some similar region between the hotspot region of Work and Home, all distributed in the Southwest. This phenomenon is similar to Li's (2021) research for Shenzhen, China. A similar area in Shenzhen is located in the suburbs, with many industrial parks. The suburbs have lower housing prices, and industrial parks usually provide worker housing. The reason for Xian is similar to Shenzhen. There are many high-tech industrial parks in southwest Xian in southwest Xian, and the differences in rental prices in Xi'an are not apparent. This suggests that bike operators should pay more attention to the Southwest of Xian.

As shown in Figure(c), the hotspot regions of transfer are mainly distributed around the metro. This is similar to Li's (2021) research. However, there is a significant difference with Gong's study on taxis, which shows that transfers occur mainly at train stations and airports. As for schools, there are a few hotspot regions where some universities are. The distribution of food is similar to work. This may be a lunch-dinner demand generated by work. The hotspot region of shopping is concentrated. In the center of Xian, there is more big shopping mall than suburbs. The hotspot region of Medical and Recreation is similar to hospitals and places of interest. It can be seen that the distribution of hospitals is even.
Figure 7 The spatial distribution of different travel activities of dockless shared bike users

According to the time distribution pattern of work, home, and transfer, four peak hours on weekdays and non-workdays have been used for further analysis. A spatial representation of the distribution of work trips is shown in Figure 8. Firstly, there are differences in travel volume and hotspot distribution between weekends and weekdays. We can find that some morning peaks disappear on weekends, while some areas still have significant morning peaks. This indicates that there are some enterprises still working on weekends. From the comparison of lunch peak on
weekdays and weekends, we can find that a few enterprises have the benefit of a lunch break.

![Figure 8 The spatiotemporal distribution of work activities](image)

The spatiotemporal distribution of transfer is represented as shown in Figure 9. Compared with weekdays, the hotspot region of weekends is much reduced. It is common to sense. The spatiotemporal distribution of home is represented as shown in Figure 10. For the night peaks, home's distribution is decentralized. For the lunch peaks, the hotspot region of home is distributed in the Southwest. This indicates that people may go directly to work by bicycle in the Southwest.
7 Conclusion

In this study, the effect of user travel regularity on travel purpose inference was considered. The results showed that the introduction of user travel regularity was practical for travel purpose inference. The user's travel regularity could distinguish different travel purposes of the same POI. The primary performance was to identify trips to shopping malls, hospitals, restaurants, and recreational places for work. The regularity of the extracted trips accounted for 32% of the weekly trips and 52% of the morning and evening peak trips, which was similar to the findings of Long (2015) et al. 70% of the trips with regularity were to companies, schools, residences, and bus stops, which were commonly known as commuting. The remaining 30% of the trips were to
POIs such as leisure and entertainment venues, restaurants, hospitals, and shopping malls. We believe that the purpose of regular trips to these POIs is no longer for shopping, medical, or treatment but for work or going home. Eventually, 10% of the trip purposes changed due to the identification of regular trips.

Subsequently, we analyzed different trip purposes' statistical characteristics and Spatio-temporal distribution characteristics. They could be mainly divided into the following four points. First, the cycling time and cycling distance for dining were the shortest, and shopping was the longest. Second, the time distribution of travel purposes was consistent with people's habits, which indirectly proved the method's accuracy. Third, the travel purposes could be roughly divided into two categories from the spatial distribution pattern. The spatial distribution of travel purposes such as Work, family, food, and Shopping was more dispersed throughout the operation area; for travel purposes such as school, entertainment, medical, and transfer, the hotspots were more apparent; hotspots are located in the corresponding POIs. Fourth, two interesting phenomena could be found through the spatial and temporal distribution patterns of work, family, and transfer. First, the lunch peak hotspots of work and family were the same, indicating that these hotspot areas benefited from going home for a lunch break; second, the hotspot areas of interchange were close to the hotspot areas of work and family. This also reflected the vital role of bike-sharing in commuting interchanges.

It was worth noting that there were limitations regarding the data and model validation. Since the dataset contained only one bike company, the data did not represent the complete picture of bike-sharing trips in Xi'an, especially when regular trips were identified. The use of other brands of bikes by users was likely overlooked. For the results, the model's accuracy can only be indirectly verified by the spatio-temporal distribution, as there was no real trip purpose for comparison. In the future, the parameters in the model can be more accurately adjusted if actual travel purposes are available for comparison. Scholars should continue their in-depth research to improve the accuracy of travel purposes in the future. In this way, this data can help public transport operators improve their services.
Declarations

Ethical Approval
Not applicable.

Competing interests
Not applicable.

Authors' contributions
Jianwei Wang and Fu Xin determined the idea and outline of the manuscript; Mengyan Cui and Hao Yang were responsible for data processing and manuscript writing; Mengyan Cui, Sen Fu and Xuxu Liu created the images; and all authors participated in the manuscript review.

Funding
Not applicable.

Availability of data and materials
Not applicable.
Reference


