Identifying and Segmenting Commuting Behavior Patterns Based on Smart Card Data and Travel Survey Data

Pengfei Lin, Jiancheng Weng, Dimitrios Alivanistos, Siyong Ma and Baocai Yin

1 The Key Laboratory of Transportation Engineering, Beijing University of Technology, No. 100 Ping Le Yuan, Chaoyang District, Beijing 100124, China; linpengfei@outlook.com (P.L.); siyongma0818@163.com (S.M.); ybc@bjut.edu.cn (B.Y.)
2 Elsevier B.V., Radarweg 29a, 1043 NX Amsterdam, The Netherlands; dimitrisalivas@gmail.com

Received: 14 May 2020; Accepted: 15 June 2020; Published: 19 June 2020

Abstract: Understanding commuting patterns could provide effective support for the planning and operation of public transport systems. One-month smart card data and travel behavior survey data in Beijing were integrated to complement the socioeconomic attributes of cardholders. The light gradient boosting machine (LightGBM) was introduced to identify the commuting patterns considering the spatiotemporal regularity of travel behavior. Commuters were further divided into fine-grained clusters according to their departure time using the latent Dirichlet allocation model. To enhance the interpretation of the behavior patterns in each cluster, we investigated the relationship between the socioeconomic characteristics of the residence locations and commuter cluster distributions. Approximately 3.1 million cardholders were identified as commuters, accounting for 67.39% of daily passenger volume. Their commuting routes indicated the existence of job–house imbalance and excess commuting in Beijing. We further segmented commuters into six clusters with different temporal patterns, including two-peak, staggered shifts, flexible departure time, and single-peak. The residences of commuters are mainly concentrated in the low housing price and high or medium population density areas; subway facilities will promote people to commute using public transport. This study will help stakeholders optimize the public transport networks, scheduling scheme, and policy accordingly, thus ameliorating commuting within cities.

Keywords: commuting; public transport; travel behavior; pattern clustering; smart card data

1. Introduction

The public transport system plays an important role in addressing urban traffic problems, such as traffic congestion, air pollution, and traffic accidents [1,2]. Numerous city administrators around the world have given priority to the development of public transport infrastructure, and encourage residents to use it by optimizing supply and policy. Taking the Beijing public transport network as an example, at the end of 2019, there were 23 subway lines with operating mileage of about 700 km and approximately 890 bus lines covering more than 40 thousand bus stops in the city of Beijing. The usage of public transport in Beijing is quite high, carrying 20 million riderships on average every day, accounting for 45.6 percent of all daily travel demand [3]. Public transport commuters are an important part of travel demand, especially during peak hours of weekdays. Understanding the travel demand of different groups of passengers, and especially that of commuters, will help urban planners and managers to improve the service level of public transport for a more sustainable public transport system.
Smart card data are collected during fare collection, thus recording the time and location of boarding and alighting for each passenger. Due to the advantage of continuous observation over a long period of time, it has been widely used to unveil travel behavior [4–7], forecast the travel demand [8–10], and investigate job–housing relationships [11–13]. Nevertheless, smart card data have several inherent limitations. Since smart cards are anonymous due to privacy concerns, the socioeconomic attributes of the cardholders, such as gender, age, education, and trip purpose, are not available. Furthermore, the public transport system is an open system, and cardholders may switch to other travel modes, such as private cars or taxis, at any given time, so their travel demands are also incomplete. Therefore, the identification of commuters using smart card data is a challenging task. Several studies have addressed this problem utilizing the spatiotemporal regularities of repetitive trips on public transit networks [4–7, 14–19]. However, these studies always used threshold or rule-based methods to classify regular passengers. For example, Briand et al. [4] proposed that the regular passengers are public transit riders who needed to travel at least ten days during a period of a month. Confirming a reasonable threshold value for the identification according to the statistics proved to be quite challenging, as the accuracy of the identification results was difficult to verify.

Although many studies measure spatiotemporal regularity of individual passengers based on smart card data, only a few studies further interpret the patterns combined with other source data, such as traditional household travel survey data. Travel survey data have the disadvantage of being time-consuming and expensive to gather, with small sample sizes and low recovery rates, but they record the individual socioeconomic attributes, travel purpose, travel mode, place of residence, place of employment, and other information, which are not available in the smart card data. Combining the advantages of smart card data with those of individual travel behavior survey data would improve both the accuracy and the interpretability of commuter identification and segmenting models.

To fill these gaps, our study developed a two-stage model to assess the rhythm of daily commuting by integrating smart card data and travel behavior survey data from a one-month period in Beijing. Firstly, we employed the light gradient boosting machine (LightGBM) model, a supervised learning algorithm, to identify the commuter considering the spatial and temporal regularity of travel behavior, and analyzed the residence and workplace distributions of commuters in Beijing. Secondly, we further divided the commuters into fine-grained groups according to their departure time using a latent Dirichlet allocation model, and analyzed the travel behavior patterns of each group. Finally, to enhance the interpretation of the patterns, we associated the socioeconomic characteristics of the identified residence locations with each commuter cluster in order to inspect the relationship between them. The contribution of our study is to provide deep insights into the spatial and temporal characteristics of commuting patterns, based on multi-source datasets, in Beijing, and interpret the behavior features that are challenging to obtain from each individual dataset. It will support the improvement of the existing transportation policies and service optimization while adapting to variations in different kinds of travel demand.

The remainder of this paper is organized as follows. In Section 2, we synthesize the related studies regarding public transport passengers’ travel patterns. In Section 3, the multi-source data for the study are presented. In Section 4, the method framework of the two-stage model is introduced. In Section 5, the results of the model are delineated and discussed. In Section 6, we summarize the key findings and propose suggestions for future work.

2. Related Studies

2.1. Spatial and Temporal Regularities

Public transit commuters among the cardholders can be identified based on their spatiotemporal regularities of repetitive trips on the public transit networks. Some studies employed the unsupervised clustering algorithm, i.e., the latent Dirichlet allocation (LDA) model and Gaussian mixture model to classify passengers into groups according to their temporal behavior [4–7]. Briand et al. [4] applied
a two-level Gaussian mixture model to identify several groups of passengers based on their timestamps of each trip, which innovatively represents the time in a continuous way rather than in discrete time bins. Similarly, Liu and Cheng [6] regarded passengers and their boarding time as “document” and “word”, respectively, and adopted the LDA model, a natural language processing algorithm, to assess their travel regularity over long periods. Mohamed et al. [7] constructed temporal profiles based on boarding information extracted from smart card data and employed a mixture of unigrams models to uncover the day-to-day travel habits of passengers.

Based on the temporal pattern, the spatial pattern is introduced and incorporated to determine regular travel behavior [14–19]. Bhaskar et al. [14] used the density-based spatial clustering of application with noise (DBSCAN) algorithm to uncover the temporal and spatial travel features of public transport passengers, and classified them into transit commuters, regular origin-destination (OD) passengers, habitual time passengers, and irregular passengers. Ma et al. [15] extracted the spatiotemporal regularity characteristics of travel for public transit passengers based on one-week smart card data in Beijing, including the number of travel days, number of similar first boarding times, number of similar route sequences, and number of similar stop ID sequences, and then combined the K-Means++ clustering algorithm and rough set theory to classify the passengers into five categories according to their spatiotemporal regularity. Ma et al. [16] measured the repeatability of spatiotemporal patterns of commuting for each individual using one-month smart card data, including residence, workplace, and departure time, and then applied the technique for order of preference by similarity to ideal solution (TOPSIS) and iterative self-organizing data (ISODATA) analysis technique to identify transit commuters. The results indicated that commuters accounted for 65.14% of all transit ridership during weekdays in Beijing in June 2015. Goulet-Langlois et al. [18] constructed each passenger’s four-week activity sequence using the identified activity places extracted from smart card data, and then employed the principal component analysis (PCA) method to unveil the regular activity sequence structure of each user.

2.2. Public Transit Demand Distribution

Smart card data were also used to better understand the spatial and temporal distribution of the public transit demand from the collective viewpoint [20–23]. Yu et al. [20] constructed heat maps based on smart card data in order to quantitatively estimate the travel demand at the bus stop level in Guangzhou, China, and applied PCA and a Gaussian mixed model to extract and cluster the features of the heat maps. Tang et al. [21] utilized the nonnegative canonical polyadic (CP) decomposition to extract several stable and basic travel patterns from the departure and arrival mobility of passengers during weekdays and weekends. Cats et al. [22] implemented the agglomerative hierarchical clustering approach to uncover the urban structures dynamics and public transport activity centers based on passenger flow spatiotemporal distribution in Stockholm, Sweden. Qi et al. [9] proposed a multi-step model to investigate and predict the regional travel pattern of bus travelers, including dividing the urban space into several regions considering the service scope of the bus stops, extracting the regional mobility patterns based on the nonnegative tensor factorization, and predicting the regional boarding and alighting patterns using artificial neural network.

Some studies focused on the job–housing distribution and revealed the urban spatial structure based on smart card data [11–13,24,25]. Zheng et al. [24] identified the distribution of commuting trips with Nanjing smart card data, and then calculated the commuting time and job–housing ratio to quantify the level of job–housing balance in each district. Using smart card data from 2011 to 2017 in Beijing, Huang et al. [25] revealed the relationship between travel behavior and residence and workplace dynamics of the subway passengers; they found that if the commute time is less than 45 min, residents tended to commute longer and get better jobs or living environments. Moreover, some research tried to estimate the travel purpose of the public transit riders that are not available from the smart card data to better understand different types of demand [26–28]. Zou et al. [26] applied a centerpoint-based algorithm to detect the home location, then employed a rule-based approach to
identify the four types of trip purposes for subway passengers based on the spatial and temporal pattern within multi-day smart card data. Bao et al. [27] identified the types of bike sharing stations based on the distribution of point of interest (POI) within the service areas, and implemented the LDA model to discover the trip purposes of bike sharing trips using the identified station types and smart card data. Medina et al. [28] inferred three activity types according to the start time and duration of an activity extracted from one-week smart card data in Singapore, including staying at home, going to work or study, or other types.

2.3. Multi-Source Data Fusion

To address the problem of incomplete personal attributes in smart card data, some studies have combined the smart card data with data from other sources, such as travel survey data or socioeconomic data, to investigate the travel patterns of public transit passengers [29–34]. Kusakabe et al. [30] extracted travel behavior information from person trip survey data and then developed a naïve Bayes-based model to estimate the trip purpose of public transit passengers’ trips observed in the smart card data. Long et al. [31] integrated traditional household survey data and smart card data to understand spatial and temporal patterns for the four types of extreme public transit commuters in Beijing, including (1) traveling earlier than other commuters, (2) traveling later in the day, (3) traveling very long distances, and (4) traveling exceptionally many times a day; household survey data were used to supplement the socio-economic background of extreme commuters. Sun and Yang [32] used the naïve Bayesian classifier to identify public transit commuters based on smart card data and travel survey data in Xiamen, China. The identification model was validated with an accuracy of 92%. However, this study failed to further analyze the travel pattern of commuting activities, since the data used in their study only covered five consecutive workdays. Ji et al. [33] identified and clustered subway commuters into three groups—a classic group, an off-peak group, and a long-distance group—and applied a mixed logit regression model to investigate the factors influencing commuting patterns (i.e., individual socioeconomic attributes, bus station density, and transfer distance) by combining smart card data and travel survey data in Nanjing, China. Long et al. [34] used one-week smart card data to determine the distribution of work or housing land and commuter travel routes in Beijing, and verified the identified results with household travel survey data and land use data.

Generally speaking, the majority of the related works usually adopted clustering methods, a hypothetical threshold, or rule-based methods to segment the regular passengers and irregular passengers; for example, Briand et al. [4] and Ji et al. [33] proposed that passengers traveling at least ten days and twelve days of one month, respectively, can be identified as commuters. Bhaskar et al. [14] presented that transit commuters needed more than 50% of their trips to be made within habitual times and between regular OD without giving reasons. These threshold values may be varied across the different cases of cities, and it was difficult to select a reasonable threshold according to the statistics and verify the accuracy of the identification results. Furthermore, many studies have classified the travel patterns but failed to interpret the different behavior patterns, because the individual socioeconomic attributes of the smart card data were not available. To address these gaps in the existing literature, our study combined the smart card data and travel behavior survey data to complement the missing information of travel purpose in the smart card data, and then presented a two-stage model to identify and explain the different types of commuting patterns in Beijing.

3. Data

3.1. Smart Card Data

Our study area was within the Sixth Ring Road in Beijing, because most of the residents live and work in this area [16]. One-month period smart card data during May 2018 were provided by the Beijing Municipal Transportation Operations Coordination Center, covering both subway and bus modes in Beijing. The data record the transaction details when passengers swipe their smart card,
including card ID, card type, boarding time, alighting time, boarding station ID, alighting station ID, etc. The smart card types mainly contain regular card, student card, and senior card; during our study period, about 15 million passengers traveled at least one time, with 400 thousand student cards, 1.1 million senior cards, and the remainder being regular cards. Since our study focuses on commuting behavior, we only retained the trips of regular cards and student cards. The passenger might have transferred between subway and bus between origin to destination, so we generated individual travel chains based on smart card data, according to the methods conducted by Weng [35]. Finally, we obtained 10 million trip chain records, covering 4.6 million passengers per day.

3.2. Travel Behavior Survey Data

The anonymous travel behavior survey was carried out by the offline survey and the online survey (via WeChat) in May 2018 to obtain public transit passengers’ socioeconomic attributes (e.g., gender, age, occupation, monthly income, and vehicle ownership), travel behavior, and smart card ID. The offline survey was completed by the investigators through face-to-face inquiry in the stations. Combining offline and online surveys can ensure that the survey scope covers some typical residential areas and workplaces within the sixth ring road in Beijing, such as Huilongguan, Tiantongyuan, and Tongzhou District. The participants included civil servants, enterprise employees, students, and others. Specifically, each respondent was requested to fill in their smart card ID and the main purpose (commuting or non-commuting) of their trip by taking public transport. The questions about the travel behavior information in the survey were as shown in Table 1. We collected a total of 1197 questionnaires; to verify whether each respondent is indeed a commuter, we extract the travel chain data over one month according to the smart card ID provided by the respondents, and then compared the information in the smart card data and the questionnaire data. We finally got 1042 valid questionnaires, including 515 commuters and 527 non-commuters.

Table 1. Some important questions about the travel behavior information in the survey.

| Question | Question Type |
|----------|---------------|
| 1. What is your smart card ID? | Mandatory |
| 2. Do you use public transit mainly for commuting (i.e., for work or study)? | Mandatory |
| 3. How many days do you take public transit during the week? | Mandatory |
| 4. How many times do you take public transit during weekdays? | Mandatory |
| 5. What is the name of the station closest to your home? | Optional |
| 6. What is the name of the station closest to your workplace? | Optional |
| 7. What is the frequent departure time for your first trip? | Mandatory |
| 8. What is the frequent departure time of your last trip? | Mandatory |

3.3. Socioeconomic Data

To further investigate how factors influence the commuting behavior patterns, we introduced housing price data, population distribution data, and subway station density to represent the socioeconomic attributes of each traffic analysis zone (TAZ). The TAZ data were provided by the Beijing Urban Planning Bureau. The housing price data were collected from the Lianjia website (http://www.ljia.com/), which is a famous trading platform for second-hand housing in China. The dataset contains about 500 thousand transaction records from 2015 to 2017 in Beijing, including the house location and the price; it was applied to calculate the average housing price of each TAZ. People who live in areas with higher housing prices usually have higher personal incomes. The population density in each TAZ during May 2016 was provided by the Beijing Municipal Transportation Operations Coordination Center based on mobile cellular signaling data. We also calculated the subway station density in each TAZ to reflect the public transport accessibility of the region.
4. Methods

4.1. Commuting Behavior Feature Extraction

In this study, a commuter is defined as a regular passenger who periodically travels between home and a destination, including workplaces and schools. In general, the majority of passengers travel twice a day for commuting, starting their trips from their homes to their workplaces (or school), and then returning to their homes. Commuting behavior is a long-term activity and subject to departure time and travel distance, and therefore commuters have formed a regular pattern during the travel, including the departure time homogeneity, travel route choice stickiness, activity place homogeneity, and symmetry of origin–destination, etc. As shown in Figure 1, a sample passenger usually takes the bus line 807 and transfers to subway line 1 and 2 for work at around 8 a.m., and returns home at around 6 p.m. by taking subway line 2 and 1 then transferring to bus line 316 or 317, which indicates departure time homogeneity and travel route choice stickiness. The activity place of this passenger is often concentrated near station 63, which suggests activity place homogeneity. Commuters usually travel two times every weekday, and the trip chains of the first trip and last trip of an individual can form a closed loop with a symmetry pattern; as shown in Figure 1, this passenger travels from station 1276 to station 63 in the morning, and then returns from station 63 to station 1276 in the afternoon. To measure the spatial and temporal regularity of travel behavior, six indicators were introduced, including travel days ($N_{day}$), travel times ($N_{trip}$), frequency of the most used routes ($N_{route}$), the number of days with symmetrical travel ($N_{symmetry}$), the entropy of departure time ($E_{time}$), and the entropy of activity place ($E_{place}$).

\[
E_{time} = \frac{1}{2} \left[ E_{time, first} + E_{time, last} \right] = \frac{1}{2} \left[ (-1) \times \frac{P_i \times \log_2(P_i)}{\log_2(I)} + (-1) \times \frac{P_j \times \log_2(P_j)}{\log_2(J)} \right]
\]  

where $P_i$ is the proportion of $i$th departure time period label of the first trip, and $P_j$ is the proportion of $j$th departure time period label of the last trip. $I$ and $J$ are the number of departure time period labels of the first trip and last trip, respectively. $E_{time}$ varies between 0 and 1, and a smaller value means that an individual’s departure time distribution is more similar.

| Card ID  | Travel mode | Boarding time | Alighting time | Boarding Station | Alighting Station | Route               |
|---------|-------------|---------------|----------------|------------------|-------------------|---------------------|
| ***15709600*** | B-R         | 2018/5/7 8:30:00 | 2018/5/7 9:26:09 | 1276            | 63                | BB07-R1-R2          |
| ***15709600*** | R-B         | 2018/5/7 18:46:00 | 2018/5/7 19:43:45 | 63              | 1276              | R2-R1-B316          |
| ***15709600*** | B-R         | 2018/5/8 8:07:01 | 2018/5/8 9:01:38 | 1276            | 63                | BB07-R1-R2          |
| ***15709600*** | R-B         | 2018/5/8 18:28:00 | 2018/5/8 19:32:36 | 63              | 1276              | R2-R1-B317          |
| ***15709600*** | B-R         | 2018/5/9 8:01:01 | 2018/5/9 9:04:05 | 1276            | 63                | BB07-R1-R2          |
| ***15709600*** | R-B         | 2018/5/9 18:33:00 | 2018/5/9 19:29:56 | 63              | 1276              | R2-R1-B317          |

Figure 1. Example characteristics of passenger commuting trips.
The frequency of the most-used routes \( (N_{\text{route}}) \), number of days with symmetrical travel \( (N_{\text{symmetry}}) \), and entropy of activity place \( (E_{\text{place}}) \) measure the spatial regularity at the route level and station level. Commuters prefer to choose the fixed travel mode from origin to destination; in city areas with the high public transit service level, passengers can choose multiple alternative route sequences to commute. Therefore, although commuters choose a fixed mode, they do not necessarily choose a fixed route every time. We calculated the numbers of most frequently used route sequences for traveling between home and workplace for each individual based on one-month travel chain data and regarded them as \( N_{\text{route,home}} \) and \( N_{\text{route,work}} \), respectively; we also calculated numbers of the corresponding most commonly used alternative route sequences and regarded them as \( N'_{\text{route,home}} \) and \( N'_{\text{route,work}} \), respectively. \( N_{\text{route}} \) can be defined as:

\[
N_{\text{route}} = N_{\text{route,home}} + N_{\text{route,work}} + N'_{\text{route,home}} + N'_{\text{route,work}}
\]

\( N_{\text{symmetry}} \) could be defined as the number of days in which the OD of the first trip and the OD of the last trip are symmetrical. Due to the alternative routes between home and workplaces, passengers may choose different stations near their origins and destinations, which will lead to bias when calculating \( N_{\text{symmetry}} \); for example, the boarding station ID of the first trip may not be inconsistent with the alighting station ID of the last trip. To solve this problem, we applied an improved DBSCAN algorithm proposed by Ma et al. [16] in order to cluster the spatially adjacent stations into several common stations and relabel these common stations with new station IDs. The 40 thousand bus stops and subway stations were clustered into four thousand groups, and the new group labels were used to represent the stations.

Commuters are characterized by frequent traveling between residence and the workplace. This is moreover evident from the low number of trips for purposes outside of commuting, such as shopping and leisure activities. We recorded the alighting stations of the first trips and boarding stations of the last trips for each individual in the one-month period, regarded the station set as a home-based activity place set, and calculated the frequency of each place. We introduced the entropy of activity place to measure the behavior similarity at the station level; the entropy of activity place \( E_{\text{place}} \) is defined as:

\[
E_{\text{place}} = (-1) \times \sum_{a} P_a \log_2(P_a)
\]

where \( P_a \) is the proportion of \( a \) th home-based activity places. \( A \) is the number of home-based activity places. \( E_{\text{place}} \) varies between 0 and 1; a smaller value means the individual’s activity places are concentrated on workplaces.

We calculated the six travel behavior indicators for each passenger based on smart card data; particularly, the six travel behavior indicators of 1042 survey participants were combined with travel purpose (commuting or non-commuting) by their smart card ID. The sample of fusion results of smart card data and travel survey data are shown in Table 2.

| Smart Card ID | \( N_{\text{day}} \) | \( N_{\text{trip}} \) | \( N_{\text{route}} \) | \( N_{\text{symmetry}} \) | \( E_{\text{time}} \) | \( E_{\text{place}} \) | Label      |
|--------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|------------|
| ***891999087*** | 7               | 11              | 6               | 2               | 0.91            | 0.65            | Non-commuter |
| ***89186404***  | 17              | 31              | 21              | 11              | 0.51            | 0.31            | Commuter    |
| ***89125720***  | 18              | 28              | 28              | 10              | 0.41            | 0              | Commuter    |
| ***89123750***  | 20              | 42              | 37              | 3               | 0.42            | 0.51            | Commuter    |
| ***89034363***  | 5               | 9               | 9               | 3               | 0.83            | 0.41            | Non-commuter |

4.2. Commuter Identification

Based on the fusion dataset, the LightGBM algorithm was applied to identify the commuters. LightGBM is a novel GBDT (gradient boosting decision tree) algorithm proposed by Ke et al. [38]. LightGBM has faster training speed, lower memory consumption, and better model accuracy; it also
where I represents the travel behavior indicators of passengers and y represents the travel purpose labels (commuter or non-commuter), the objective of LightGBM is to find a function \( f(x) \) to minimize the loss function \( L(y, f(x)) \):

\[
\hat{f} = \arg\min_{f} E_{x,y} [L(y, f(x))] 
\]  

The LightGBM will integrate all decision trees to get the final model as follows:

\[
\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) 
\]  

where \( K \) is the number of trees and \( f_k \) denotes the \( k \)th tree with the leaf score. For the \( k \)th tree, the objective function can be approximated using Taylor expansion, and after removing the constant term it becomes:

\[
f_k \approx \arg\min_{f_k} \sum_{i=1}^{n} \left[ g_i f_k(x_i) + \frac{1}{2} h_i f_k^2(x_i) \right] + \Omega(f_k) 
\]  

where \( g_i \) and \( h_i \) denote the first-order derivative and second-order derivative of the loss function, respectively. \( \Omega(f_k) \) is the regularization term to avoid overfitting; it can be defined as:

\[
\Omega(f_k) = \gamma T_r + \frac{1}{2} \lambda \sum_{j=1}^{T_r} w_j^2 
\]  

where \( T_r \) denotes the number of leaves in a tree, \( \gamma \) denotes the complexity of each leaf, \( w_j \) is the score on the \( j \)th leaf, and \( \lambda \) is the penalty coefficient. For a specific tree \( q(x) \), the optimal leaf weight scores of each leaf node \( w_j^* \) is defined as:

\[
w_j^* = -\frac{\sum_{i \in I_l} g_i}{\sum_{i \in I_l} h_i + \lambda} 
\]  

Then, \( w_j^* \) is substituted into the loss function, and the objective function becomes:

\[
f^*_r(q) = -\frac{1}{2} \sum_{j=1}^{T_r} \frac{(\sum_{i \in I_l} g_i)^2}{\sum_{i \in I_l} h_i + \lambda} + \gamma T 
\]  

The objective function after splitting is defined as follows:

\[
G = \frac{1}{2} \left[ \frac{(\sum_{i \in I_l} g_i)^2}{\sum_{i \in I_l} h_i + \lambda} + \frac{(\sum_{i \in I_r} g_i)^2}{\sum_{i \in I_r} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] 
\]  

where \( I_L \) and \( I_R \) are the left and right leaf node sets after the splitting, respectively, and \( I = I_L \cup I_R \).

LightGBM employs the gradient-based one-side sampling (GOSS) algorithm and the exclusive feature bundling (EFB) algorithm to reduce the computational complexity without losing accuracy. The gain of the objective function mainly depends on the samples with a larger value of gradient, and thus the GOSS algorithm retains the samples with a large gradient and randomly samples those instances with a small gradient so as to reduce the size of the dataset. In large-scale datasets, there are usually a lot of sparse features; the EFB algorithm bundles multiple sparse and mutually exclusive features into new features, successfully reducing the dimensions of the dataset. Additionally, the model innovatively uses the leaf-wise strategy and histogram algorithm to grow trees, so as to avoid overfitting and improve the model performance.
The fusion dataset was divided into the training dataset and the test dataset; the training dataset accounts for 80% of the fusion dataset. The LightGBM model was developed using the LightGBM package in Python (https://github.com/microsoft/LightGBM). Some hyper-parameters, i.e., the number of trees, learning rate, max depth, and the minimum number of data needed in each leaf will significantly affect the performance of the model. After calibrating these four hyper-parameters based on the grid search and five-fold cross-validation, the LightGBM achieved its best performance with the learning rate of 0.05, the number of the trees of 120, the max depth of 3, and the minimum number of data needed in each leaf of 30. Based on the test dataset, we used the F1 score [39] and the area under the curve of receiver operating characteristic (AUC ROC) score [40] to compare the accuracy of LightGBM with other classification models, including K-nearest neighbors (KNN), support vector machines (SVM), decision tree (DT), artificial neural networks (ANN), and naïve Bayes (NB). As shown in Table 3, the LightGBM outperformed the other methods.

Table 3. Model performance comparison. KNN: K-nearest neighbors; SVM: support vector machines; DT: decision tree; ANN: artificial neural networks; NB: naïve Bayes; LightGBM: light gradient boosting machine; AUC ROC: area under the curve of receiver operating characteristic.

| Model         | F1 Score | AUC ROC |
|---------------|----------|---------|
| KNN           | 0.9257   | 0.9256  |
| SVM           | 0.9300   | 0.9301  |
| DT            | 0.9094   | 0.9097  |
| ANN           | 0.9290   | 0.9289  |
| NB            | 0.9271   | 0.9269  |
| LightGBM      | 0.9343   | 0.9343  |

4.3. Commuter Segmentation

Based on the commuter identification results, the LDA topic model was employed to cluster the commuters into fine-grained groups based on their departure time of every weekday. LDA is a three-level Bayesian mixture model, originally proposed by Blei et al. [41]. The LDA model is usually used to find the topics of a corpus of documents in the field of natural language processing, where the documents contain multiple words, and each topic is modeled as a distribution of words. In our study, every commuter can be regarded as a document comprised of words, where words represent the departure date and time label of each trip on every weekday; for example, Monday_9. The topics represent different temporal pattern distributions.

1. For each temporal pattern \( k \in 1, 2, \ldots, K \) draw \( \beta_k \sim \text{Dirichlet}(\beta) \);
2. For each commuter \( d \in 1, 2, \ldots, D \) in the dataset, choose a temporal pattern \( \theta_d \sim \text{Dirichlet}(\alpha) \);
3. For each departure time label \( i \) in commuter \( d \):
   1. Draw a temporal pattern \( Z_{di} \sim \text{Multinomial}(\theta_d) \);
   2. Draw the departure time label \( w_i \sim \text{Multinomial}(\beta_{Z_{di}}) \).

The parameters of the LDA model were estimated by the expectation-maximization (EM) algorithm, which requires the posterior distribution of the latent variable in advance; the posterior distribution is:

\[
p(\theta, \beta, z|w, \alpha, \eta) = \frac{p(\theta, \beta, z|w, \alpha, \eta)}{p(w|\alpha, \eta)} \quad (11)
\]

However, this posterior distribution is difficult to estimate due to the coupling between \( \theta, \beta, \) and \( z \). To address this problem, Hoffman et al. [42] introduced the variational Bayesian method to use a simpler distribution \( q(z, \theta, \beta|\lambda, \phi, \gamma) \) to estimate \( p(\theta, \beta, z|w, \alpha, \eta) \), supposing \( \lambda, \phi \) and \( \gamma \) are formed
by independent distributions of $w$, $\alpha$, and $\eta$, respectively. Those variational parameters $\lambda$, $\phi$, and $\gamma$ are optimized to maximize the evidence lower bound (ELBO):

$$\log P(w|\alpha, \eta) \geq L(w, \phi, \gamma, \lambda) = E_q[\log p(w, z, \theta, \beta|\alpha, \eta)] - E_q[\log q(z, \theta, \beta)]$$ (12)

The optimal value of the ELBO can be obtained by minimizing the Kullback–Leibler (KL) divergence between $q(z, \theta, \beta)$ and the true posterior $p(\theta, \beta, z|w, \alpha, \eta)$. More details can be found in the study constructed by Hoffman et al. [42].

The LDA model was developed using the Scikit-learn package (https://scikit-learn.org/stable/). Perplexity is usually applied to determine the number of topics of the LDA model [41]. We tested the different number of topics from 2 to 10; when the number of topics is 6, the perplexity achieves the minimum value. Therefore, in this study, the commuters were further classified into six groups with different departure time patterns.

5. Results

5.1. Commuter Identification Results

We employed the LightGBM model to identify the commuters and non-commuters in the smart card data of May 2018. The numbers of commuters and non-commuters were 3,109,655 and 12,128,971, respectively. Commuters accounted for 67.39% of daily passenger volume during weekdays in May 2018. This finding is similar to Ma et al. [16], which indicated that commuters accounted for 65.14% of passenger volume in Beijing in June 2015. Table 4 shows the average values of the six travel behavior indicators between non-commuters and commuters. The values of travel days ($N_{day}$), travel times ($N_{trip}$), frequency of the most used routes ($N_{route}$), and the number of days with symmetrical travel ($N_{symmetry}$) of commuters were larger than non-commuters, which indicates that commuters have higher travel intensity and more regular spatiotemporal behavior. The values of the entropy of departure time ($E_{time}$) and the entropy of activity place ($E_{place}$) of non-commuters were higher than those of the commuters, suggesting that the trips generated by non-commuters have a more discretionary character.

| Type       | $N_{day}$ | $N_{trip}$ | $N_{route}$ | $N_{symmetry}$ | $E_{time}$ | $E_{place}$ |
|------------|-----------|-----------|-------------|----------------|------------|-------------|
| Non-commuters | 3.77      | 5.95      | 3.82        | 1.06           | 0.57       | 0.52        |
| Commuters   | 17.98     | 33.76     | 24.18       | 9.42           | 0.45       | 0.30        |

The home and workplace distributions of commuters in the TAZ level are shown in Figure 2. The residence was mainly distributed in the urban periphery; some typical residential areas include Huilongguan, Tiantongyuan, Shahe, Wangjing, Shuangjing, Fangzhuang, and Tongzhou District. For example, more than 260 thousand passengers lived in the Tongzhou District and took the bus or subway to commute every weekday. The workplaces were mainly distributed in the central area of the city; some typical workplaces include the Beijing central business district (CBD), Beijing Financial Street, Wangjing, Zhongguancun, and Zhongguancun Software Park. The Beijing Financial Street and CBD areas have gathered a large number of eminent listed companies and government agencies; commuters who work there must take public transport for commuting due to the high parking costs and limited parking resources. Many high-tech companies, Internet companies, and universities are located in the Zhongguancun area, and thus it appeals to a large number of residents and students without private cars who commute by public transport. Furthermore, some famous primary and secondary schools in Beijing are also located in the Zhongguancun area, drawing many students to come here by bus or subway in order to enjoy better educational resources. Our finding of job–housing distribution in Beijing is in accordance with the case study constructed by Ma et al. [16].
located in the Zhongguancun area, and thus it appeals to a large number of residents and students without private cars who commute by public transport. Furthermore, some famous primary and secondary schools in Beijing are also located in the Zhongguancun area, drawing many students to come here by bus or subway in order to enjoy better educational resources. Our finding of job–housing distribution in Beijing is in accordance with the case study constructed by Ma et al. [16].

Figure 2. (a) The residence distribution of commuters; (b) The workplace distribution of commuters.

The study area was divided into five regions according to the spatial location of the Beijing Ring Road. The numbers of residences and workplaces were counted for the different regions as shown in Table 5. The results reveal the existence of a job–housing imbalance in Beijing. More than 60 percent of commuters live in areas outside the Fourth Ring Road in Beijing, while about 60 percent of workplaces are located within the central urban area within the Fourth Ring Road. This finding is largely consistent with previous studies [12,16,43]; the job–housing imbalance phenomenon could be ascribed to the single-centered urban structure and the centripetal public transport network of Beijing.

Table 5. Residence and workplace distribution in different regions.

| Region                              | Proportion of Residences | Proportion of Workplace |
|-------------------------------------|--------------------------|-------------------------|
| Within the 2nd Ring Road            | 6.31%                    | 15.9%                   |
| Between the 2nd and 3rd Ring Road   | 14.00%                   | 25.1%                   |
| Between the 3rd and 4th Ring Road   | 19.01%                   | 23.3%                   |
| Between the 4th and 5th Ring Road   | 18.91%                   | 18.2%                   |
| Between the 5th and 6th Ring Road   | 41.78%                   | 17.5%                   |

5.2. Commuter Segmentation Results

We further divided commuters into six clusters with different temporal patterns, such as two-peek, staggered shifts, flexible departure time, and single-peak. The temporal pattern profiles for all six clusters during weekdays are graphically depicted in Figure 3. Cluster 1, Cluster 2 and Cluster 3 have obvious two-peak patterns with some nuances; the three clusters contribute to 73% of all commuters, indicating that passengers of these three clusters are typical commuters in Beijing. Both the morning peak and evening peak of Cluster 1 are earlier and more concentrated than Cluster 2 and Cluster 3, and
such commuters usually take public transit at 7 a.m. and 7 p.m., respectively. Passengers in Cluster 2 usually travel during the morning peak hours (from 7 a.m. to 9 a.m.) and evening peak hours (from 5 p.m. to 7 p.m.). Cluster 3 shows staggered shift patterns; passengers of this group are more likely to commute after the morning peak hour (from 9 a.m. to 10 a.m.) and evening peak hour (from 7 p.m. to 10 p.m.), and the evening peak has a greater variance than Clusters 1 and 2. Cluster 4 shows a flexible departure time pattern and accounts for 14% of all the commuters. Their departure time distribution is more diffuse than other groups, and the passengers of this group have a high possibility between 9 a.m. to 3 p.m.. Cluster 5 and Cluster 6 present only a one-peek pattern, and contribute to 9% and 5% of all commuters, respectively. Cluster 5 represents high regular travel in the morning and a sporadic manner in the evening peak. Cluster 6 shows regular afternoon activity from 3 p.m. to 5 p.m. with a relatively high variance of departure time, and its afternoon peak hour occurs earlier than the peak hour of other clusters.

Figure 3. The temporal pattern profiles for all six clusters during weekdays.

5.3. Commuting Behavior Interpretation

We further analyzed the commuting behavior between six commuter clusters in terms of the number of daily trips ($N_{daily}$), the entropy of departure time ($E_{time}$), the entropy of activity place ($E_{place}$), commuting distance ($D$), and relative proportion of student cards ($P_{student}$), as shown in Table 6. Passengers in both Cluster 1 and Cluster 5 have a larger commuting distance, which may result from the fact that such passengers travel earlier than other clusters, especially in the morning. Four indicators of Cluster 2 are relatively balanced compared with other groups. Passengers in Cluster 3 show a relatively shorter commuting distance and higher heterogeneity of departure time. Cluster 4 has the highest travel intensity and high heterogeneity of both departure time and activity place, which suggests that commuters in this group may have trips with other travel purposes, such as business outside of their commuting travel. Cluster 6 also has the shortest travel distance for passengers and a higher heterogeneity of departure time. Moreover, this group has the highest relative proportion of student cards, and their departure time is mainly concentrated at 3 p.m., which also coincides with the dismissal time of primary and secondary schools in Beijing.
Figure 4 illustrates the 100 most frequently used commuting routes for each commuter cluster. From Figure 4a–c, it is found that commuters who work in the CBD are mainly from Tongzhou District, and commuters whose workplaces are in Zhongguancun or Zhongguancun Software Park live in Shahe, Huilongguan or Tiantongyuan. Other commuting routes include from Pingguoyuan to Beijing Financial Street and from Changyang to Fengtai Science Park. The aforementioned commuting routes are an important part of commuting demand in Beijing; this also reflects the single-centered urban structure of Beijing. Compared with Clusters 1, 2, and 3, commuters who travel from Tongzhou District to the CBD are still an important part of Cluster 4; however, the proportion of commuting routes with a short distance in this cluster is higher than the aforementioned three clusters. In Figure 4e, a large number of commuters live in areas outside the Fifth Ring Road, such as Changyang and Shahe, whereas their workplace is located in the area around the Second Ring Road; this further explains the reason that Cluster 5 commuters have longer commute distances and earlier departure times than other types of commuters. Figure 4f shows that Cluster 6 has many short-distance commuting routes, and this proportion is significantly higher than in other groups.

Table 6. Travel patterns of the six clusters.

| Category | N_daily | E_time | E_place | D   | P_student |
|----------|---------|--------|---------|-----|-----------|
| Cluster 1 | 1.87    | 0.40   | 0.25    | 15.77 | 25.33%    |
| Cluster 2 | 1.86    | 0.41   | 0.24    | 13.72 | 5.26%     |
| Cluster 3 | 1.82    | 0.49   | 0.27    | 13.24 | 3.17%     |
| Cluster 4 | 2.11    | 0.62   | 0.52    | 12.84 | 4.29%     |
| Cluster 5 | 1.73    | 0.41   | 0.30    | 18.00 | 22.00%    |
| Cluster 6 | 1.72    | 0.50   | 0.32    | 12.46 | 39.94%    |
The commuting routes of cluster 1; (b) The commuting routes of cluster 2; (c) The commuting routes of cluster 3; (d) The commuting routes of cluster 4; (e) The commuting routes of cluster 5; (f) The commuting routes of cluster 6.

To interpret the patterns of different commuter clusters, we introduced the socioeconomic information of each TAZ assigned to each commuter through their home location. The socioeconomic conditions include three variables: the housing price, population density, and subway station density of each TAZ. We employed the K-means method to cluster the TAZ into seven categories according to the distribution level of the three variables in each TAZ. The mapping of the clustering result is shown in Figure 5. It is worth noting that we omitted the TAZ with low population density. From Figure 5, the distribution of the socioeconomic clusters has a relationship with the distance to the center of Beijing. Generally, TAZs located in the urban center have higher house pricing, population density, and subway station density.

The distributions of the socioeconomic clusters for traffic analysis zones (TAZs).

We identified the relationship between the socioeconomic clusters of the TAZ and the commuter clusters by calculating the proportion of socioeconomic clusters for each commuter cluster, as shown in Figure 6. Generally, the socioeconomic clusters have similar distributions across all commuter clusters. The commuting demand is mainly concentrated in the low housing price and high or medium population density areas, and the subway facilities will also attract more people to travel by public transport. By contrast, people who live in high housing price areas in the urban center of Beijing may have a relatively high personal income, and they may use other travel modes for commuting rather than public transport, for example, private cars. As an important part of the commuters in Beijing, Cluster 1, Cluster 2, and Cluster 3 commuters are mainly distributed in areas with relatively low
housing prices and high population density. Job opportunities in Beijing are mainly distributed in the downtown with high housing price and limited parking resources, most of the commuters cannot afford the higher house rent in the residences in the downtown, and these commuters are compelled to live in urban periphery areas, which lead to excessive commuting. This finding is in accord with the existing research [25].

The Cluster 4 commuters mainly live in the areas with high and medium housing prices and subway facilities; their flexible departure time and shortest commuting distance characteristics could be attributed to the fact that their home locations usually have high accessibility to their destination. The Cluster 5 commuters are mainly located in areas with low housing prices at a relatively high proportion; these areas are located far from the center of Beijing and without subway facilities, which may explain the reason for their early departure time. The Cluster 6 commuters with a high proportion of student card type live in the urban center areas with high housing prices and subway stations; these areas are surrounded by relatively good educational resources in Beijing. It is worth noting that although our socioeconomic attribute data are not in the same period as the smart card data, this will not cause much bias, because socioeconomic attribute data reflect the urban structure, which will not change significantly in the short term.

![Figure 6. The proportion of socioeconomic clusters for each commuter cluster.](image-url)

6. Conclusions

This study proposed a two-stage model to identify and classify the commuters by incorporating the smart card data and travel behavior survey data. We employed the LightGBM to identify non-commuters and commuters using six travel behavior attributes by mining spatiotemporal travel regularities over a one-month period of weekdays. We identified the residence and workplace areas and analyzed the job–housing balance after extracting the home and workplace stations of each commuter. We used the LDA model to classify the commuters into fine-grained groups based on their departure time. We also introduced the socioeconomic attribute of each TAZ to enhance the interpretation of the patterns of different commuter groups.

Approximately 3.1 million cardholders were identified as public transit commuters in Beijing, accounting for 67.39% of the daily passenger volume; this finding is similar to Ma et al. [16].
The identification accuracy of the LightGBM model reaches as high as 93.43%, which outperforms other classification methods such as the KNN, SVM, DT, ANN, and NB algorithms. The home and workplace location distributions indicated a job–housing imbalance and excess commuting in Beijing, which are related to the single-centered urban structure of Beijing [12,16,43]. The commuters were further divided into six groups according to their temporal pattern during every weekday such as two-peak, staggered shifts, flexible departure time, and single-peak patterns. The spatiotemporal heterogeneity of travel behavior for the six groups was further presented by comparing their travel intensity, travel distances, activity entropy, and commuting route distributions, etc. The commuting demand in Beijing is mainly concentrated in the low housing price and high or medium population density areas, and the subway facilities also attract more people to travel by public transport.

Our findings could be useful for several relevant applications and scenarios. The identified excessive commuting phenomenon in Beijing will help planners to reshape a more balanced job and housing relationship in Beijing to further reduce reliance on private cars and relieve traffic congestion. It may help public transport operators better know the travel demand of passengers and propose targeted policies accordingly; for example, it is suggested to implement more elaborate discount fare strategy to attract more commuters with flexible departure times to travel during non-peak hours, so as to reduce the crowdedness of some commuter routes during the peak hours. Moreover, it may also help redesign and optimize the existing public transit networks. For example, planning a demand-response customized shuttle bus route between Tongzhou Distinct and the CBD will significantly reduce travel time for public transit commuters.

However, several aspects need to be further improved. Our study focused on the commuting behavior (for work or study) during the weekdays, but we plan to explore the characteristics of other travel activities, such as leisure, and the inherent difference between travel activities during the weekday and weekend. Furthermore, we studied the travel behavior of passengers with student type cards and regular type cards. Routine travel behavior of other passenger groups, such as older people, will also be investigated in the future. Finally, we identified that commuter travel behavior varies based on their place of residence and place of work. In the future, we plan on using long-term smart card data and on introducing other socioeconomic attributes in each TAZ to deeply analyze the longitudinal changes in public transport commuting behavior.

**Author Contributions:** Conceptualization, J.W. and B.Y.; methodology, P.L.; formal analysis, P.L. and S.M.; visualization, P.L.; writing—original draft preparation, P.L.; writing—review and editing, J.W. and D.A.; supervision, J.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China (Grant No.51578028, U1811463), the “Beijing Nova” Program by the Beijing Science and Technology Commission (Grant No. Z17110000171100), and the program of China Scholarships Council (No.201906540004).

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Litman, T. Evaluating Public Transit Benefits and Costs: Best Practices Guidbook. 2010. Available online: [https://www.vtpi.org/tranben.pdf](https://www.vtpi.org/tranben.pdf) (accessed on 3 June 2020).
2. Truong, L.T.; Currie, G. Macroscopic road safety impacts of public transport: A case study of Melbourne, Australia. *Accid. Anal. Prev.* 2019, 132, 105270. [CrossRef] [PubMed]
3. Beijing Municipal Committee of Transport. Beijing Public Transport Network Master Planning. 2020. Available online: [http://jtweb.beijing.gov.cn/xxgk/tpxw/202004/P020200410652921416550.pdf](http://jtweb.beijing.gov.cn/xxgk/tpxw/202004/P020200410652921416550.pdf) (accessed on 3 June 2020).
4. Briand, A.S.; Côme, E.; Mohamed, K.; Oukhellou, L. A mixture model clustering approach for temporal passenger pattern characterization in public transport. *Int. J. Data Sci. Anal.* 2016, 1, 37–50. [CrossRef]
5. Briand, A.S.; Côme, E.; Tripanier, M.; Oukhellou, L. Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transp. Res. Emerg. Technol.* 2017, 79, 274–289. [CrossRef]
6. Liu, Y.; Cheng, T. Understanding public transit patterns with open geodemographics to facilitate public transport planning. *Transportmetrica* 2020, 16, 76–103. [CrossRef]
7. Mohamed, K.; Côme, E.; Baro, J.; Oukhellou, L. Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data. In Proceedings of the 3rd International Workshop on Urban Computing, Seattle, WA, USA, 24 August 2014.
8. Liu, Y.; Liu, Z.; Jia, R. DeepPF: A deep learning based architecture for metro passenger flow prediction. *Transp. Res. Emerg. Technol.* 2019, 101, 18–34. [CrossRef]
9. Qi, G.; Huang, A.; Guan, W.; Fan, L. Analysis and prediction of regional mobility patterns of bus travellers using smart card data and points of interest data. *IEEE Trans. Intell. Transp. Syst.* 2018, 20, 1197–1214. [CrossRef]
10. Ding, C.; Wang, D.; Ma, X.; Li, H. Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. *Sustainability* 2016, 8, 1100. [CrossRef]
11. Marciniacz, S.; Bartosiewicz, B. Commuting patterns and urban form: Evidence from Poland. *J. Transp. Geogr.* 2018, 70, 31–39. [CrossRef]
12. Zhao, P. The impact of the built environment on individual workers’ commuting behavior in Beijing. *Int. J. Sustain. Transp.* 2013, 7, 389–415. [CrossRef]
13. Zhou, J.; Murphy, E. Day-to-day variation in excess commuting: An exploratory study of Brisbane, Australia. *J. Transp. Geogr.* 2019, 74, 223–232. [CrossRef]
14. Bhaskar, A.; Chung, E. Passenger segmentation using smart card data. *IEEE Trans. Intell. Transp. Syst.* 2014, 16, 1537–1548.
15. Ma, X.; Wu, Y.J.; Wang, Y.; Chen, F.; Liu, J. Mining smart card data for transit riders’ travel patterns. *Transp. Res. Emerg. Technol.* 2013, 36, 1–12. [CrossRef]
16. Ma, X.; Liu, C.; Wen, H.; Wang, Y.; Wu, Y.J. Understanding commuting patterns using transit smart card data. *J. Transp. Geogr.* 2017, 58, 135–145. [CrossRef]
17. Kieu, L.M.; Bhaskar, A.; Chung, E. A modified density-based scanning algorithm with noise for spatial travel pattern analysis from smart card AFC data. *Transp. Res. Emerg. Technol.* 2015, 58, 193–207. [CrossRef]
18. Goulet-Langlois, G.; Koutsopoulos, H.N.; Zhao, J. Inferring patterns in the multi-week activity sequences of public transport users. *Transp. Res. Emerg. Technol.* 2016, 64, 1–16. [CrossRef]
19. Manley, E.; Zhong, C.; Batty, M. Spatiotemporal variation in travel regularity through transit user profiling. *Transportation* 2018, 45, 703–732. [CrossRef] [PubMed]
20. Yu, C.; He, Z. Analysing the spatial-temporal characteristics of bus travel demand using the heat map. *J. Transp. Geogr.* 2017, 58, 247–255. [CrossRef]
21. Tang, J.; Wang, X.; Zong, F.; Hu, Z. Uncovering Spatio-temporal Travel Patterns Using a Tensor-based Model from Metro Smart Card Data in Shenzhen, China. *Sustainability* 2020, 12, 1475. [CrossRef]
22. Cats, O.; Wang, Q.; Zhao, Y. Identification and classification of public transport activity centres in Stockholm using passenger flows data. *J. Transp. Geogr.* 2015, 48, 10–22. [CrossRef]
23. Gong, Y.; Lin, Y.; Duan, Z. Exploring the spatiotemporal structure of dynamic urban space using metro smart card records. *Comput. Environ. Urban* 2017, 64, 169–183. [CrossRef]
24. Zheng, M.; Liu, F.; Guo, X.; Lei, X. Assessing the Distribution of Commuting Trips and Jobs-Housing Balance Using Smart Card Data: A Case Study of Nanjing, China. *Sustainability* 2019, 11, 5346. [CrossRef]
25. Huang, J.; Levinson, D.; Wang, J.; Zhou, J.; Wang, Z.J. Tracking job and housing dynamics with smartcard data. *Proc. Natl. Acad. Sci. USA* 2018, 115, 12710–12715. [CrossRef] [PubMed]
26. Zou, Q.; Yao, X.; Zhao, P.; Wei, H.; Ren, H. Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway. *Transportation* 2018, 45, 919–944. [CrossRef]
27. Bao, J.; Xu, C.; Liu, P.; Wang, W. Exploring bikesharing travel patterns and trip purposes using smart card data and online point of interests. *Netw. Spat. Econ.* 2017, 17, 1231–1253. [CrossRef]
28. Medina, S.A.O. Inferring weekly primary activity patterns using public transport smart card data and a household travel survey. *Travel Behav. Soc.* 2018, 12, 93–101. [CrossRef]
29. Zhou, J.; Murphy, E.; Long, Y. Commuting efficiency in the Beijing metropolitan area: An exploration combining smartcard and travel survey data. *J. Transp. Geogr.* 2014, 41, 175–183. [CrossRef]
30. Kusakabe, T.; Asakura, Y. Behavioural data mining of transit smart card data: A data fusion approach. *Transp. Res. Emerg. Technol.* 2014, 46, 179–191. [CrossRef]
31. Long, Y.; Liu, X.; Zhou, J.; Chai, Y. Early birds, night owls, and tireless/recurring itinerants: An exploratory analysis of extreme transit behaviors in Beijing, China. *Habitat Int.* **2016**, *57*, 223–232. [CrossRef]

32. Sun, S.; Yang, D. Identifying Public Transit Commuters Based on Both the Smartcard Data and Survey Data: A Case Study in Xiamen, China. *J. Adv. Transp.* **2018**, *2063*, 1–10. [CrossRef]

33. Ji, Y.; Cao, Y.; Liu, Y.; Guo, W.; Gao, L. Research on classification and influencing factors of metro commuting patterns by combining smart card data and household travel survey data. *IET Intell. Transp. Syst.* **2019**, *13*, 1525–1532. [CrossRef]

34. Long, Y.; Thill, J.C. Combining smart card data and household travel surveys to analyze jobs–housing relationships in Beijing. *Comput. Environ. Urban* **2015**, *53*, 19–35. [CrossRef]

35. Weng, J.C.; Wang, C.; Wang, Y.Y.; Chen, Z.H.; Peng, S. Extraction Method of Public Transit Trip Chains Based on the Individual Riders’ Data. *J. Transp. Syst. Eng. Inf. Technol.* **2017**, *17*, 67–73.

36. Mercorelli, P. Denoising and harmonic detection using nonorthogonal wavelet packets in industrial applications. *J. Syst. Sci. Complex.* **2007**, *20*, 325–343. [CrossRef]

37. Mercorelli, P. Biorthogonal wavelet trees in the classification of embedded signal classes for intelligent sensors using machine learning applications. *J. Franklin Inst.* **2007**, *344*, 813–829. [CrossRef]

38. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *3146–3154*.

39. Chinchor, N.; Sundheim, B. MUC-5 Evaluation Metrics. In *Proceedings of the 5th Conference on Message Understanding*, Maryland, MD, USA, 25–27 August 1993.

40. Hand, D.J.; Till, R.J. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach. Learn.* **2001**, *45*, 171–186. [CrossRef]

41. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.

42. Hoffman, M.D.; Blei, D.M.; Wang, C.; Paisley, J. Stochastic variational inference. *J. Mach. Learn. Res.* **2013**, *14*, 1303–1347.

43. Hu, L.; Sun, T.; Wang, L. Evolving urban spatial structure and commuting patterns: A case study of Beijing, China. *Transp. Res. Transp. Environ.* **2018**, *59*, 11–22. [CrossRef]