TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

*Research Article*

Transportation Research Record
1–9

DOI: 10.1177/03611981211018474
journals.sagepub.com/home/trr
SAGE

# Hybrid-Data Approach for Estimating Trip Purposes

**Xiaoling Luo[1], Adrian Cottam[2], Yao-Jan Wu[2], and Yangsheng Jiang[3,4]**

**Abstract**

Trip purpose information plays a significant role in transportation systems. Existing trip purpose information is traditionally collected through human observation. This manual process requires many personnel and a large amount of resources. Because of this high cost, automated trip purpose estimation is more attractive from a data-driven perspective, as it could improve the efficiency of processes and save time. Therefore, a hybrid-data approach using taxi operations data and point-of-interest (POI) data to estimate trip purposes was developed in this research. POI data, an emerging data source, was incorporated because it provides a wealth of additional information for trip purpose estimation. POI data, an open dataset, has the added benefit of being readily accessible from online platforms. Several techniques were developed and compared to incorporate this POI data into the hybrid-data approach to achieve a high level of accuracy. To evaluate the performance of the approach, data from Chengdu, China, were used. The results show that the incorporation of POI information increases the average accuracy of trip purpose estimation by 28% compared with trip purpose estimation not using the POI data. These results indicate that the additional trip attributes provided by POI data can increase the accuracy of trip purpose estimation.

Trip purpose information can assist transportation planners in making data-driven decisions (*1–3*). Through understanding how and why people make trips, transportation planners can more accurately design transportation networks to plan ahead for future demand. Trip purpose information is traditionally collected through surveys such as emails, phone interviews, or even door-to-door household surveys. The drawbacks of these methods are that the process is time-consuming and the quality of the data depends on the survey respondents. It is also challenging to obtain a representative view of the area being observed, as commonly the number of survey responses are few.

However, because of recent advancements in information technology and data acquisition, large amounts of data are available that can be used in place of surveys. Several hot topics have developed in transportation planning research around using automatically collected data that can replace surveys in areas such as transit operations, taxi operations, and railway operations. Some of the transit issues that have been researched include estimating travel demand (*4–9*), fare structure analysis (*7*, *10*), and evaluating the reliability of transit services (*11*). Taxi operations research has been directed toward travel time and travel speed analysis (*12*, *13*) and visual analytics of taxi trips (*14*). Railway operations research has focused on demand estimation (*15*) and travel behavior analysis (*4*, *15*).

One of the data sources these studies found that can be used in place of surveys is GPS data. Several studies use GPS trajectory data to evaluate various transportation planning metrics (*2*, *3*, *5*, *16*) One area that has been studied using this GPS data is trip purpose estimation, usually using GPS data collected from a group of volunteers. One such study by Bohte and Maat (*5*) measured trip modes as well as trip purposes based on predefined principles using the GPS trajectory data of more than 1,000 participants. A similar study by Lu and Zhang (*16*)

[1]Chongqing Key Laboratory of Traffic & Transportation, College of Traffic and Transportation, Chongqing Jiaotong University, Chongqing, People's Republic of China
[2]Department of Civil and Architectural Engineering and Mechanics, University of Arizona, Tucson, AZ
[3]School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, Sichuan, People's Republic of China
[4]National United Engineering Laboratory of Integrated and Intelligent Transportation, Chengdu, Sichuan, People's Republic of China

**Corresponding Author:**
Yangsheng Jiang, jiangyangsheng@swjtu.edu.cn

concentrated on estimating trip purposes by applying GPS data to machine learning methods for long-distance passenger travel. Shen and Stopher (*2*) pointed out that the quality of the survey data used to validate the trip purpose estimation with GPS data is significant for the accuracy of the estimation. Based on this, Xiao et al. (*3*) used smartphone-based trip surveys as well as GPS trajectory data. Each time a trip was made, the user could answer a survey question verifying whether or not the destination was correct. This allowed for verification of the GPS trip purpose estimation, so that the model's accuracy could be improved.

GPS data are very useful for estimating trip purposes, but there are several drawbacks to using GPS data. Collecting GPS data generally requires many drivers, each of whom have to carry a GPS device or install a GPS tracking application on their mobile phone to collect real-time trajectory data. Furthermore, it is often necessary to perform the time-consuming task of manually evaluating the GPS data to verify its quality and accuracy. Another drawback of using GPS data is that it can be difficult to obtain a representative sample or a precise trip purpose estimation, as the sample size is generally small (*5*, *16*). If it was possible to collect the GPS data of every traveler in the study area, these methods would be very useful. However, until it is possible to achieve that, GPS data cannot represent the study area comprehensively.

To overcome these drawbacks of using GPS data, some studies attempted to estimate specific trip purposes, such as commuting (*17–20*) or education-related trips (*21*). However, these methods are limited because they can only obtain specific trip purposes. It is more useful to get several different categories of trip purposes, because it provides more information about the study area. This extra information allows the study area to be more thoroughly understood (*17*). Several studies focused on detecting trip purposes from automatically recorded data for transit trips Kumar et al., 2018 (*1, 22–24*). First, trip attributes are defined and extracted from transit data. A classification model is then trained and validated with survey data. Finally, trip purposes are estimated for transit passengers using the obtained attributes and the trained Bayesian classification model. However, the estimation of trip purposes with similar attributes is difficult when using temporal data alone (*1*). For example, the "commute to school" and "commute to work" trips can be easily misclassified. To address this issue, trip purposes that have similar attributes with different trip purposes were clustered into a single trip purpose category to improve the accuracy (e.g., "commute to school" and "commute to work" are clustered as a single "commuter" category). This approach improves the accuracy of the estimation, but information completeness is lost when the categories are broadened. To overcome the drawbacks of using temporal data alone, other studies tried to use point-of-interest (POI) data to detect trip purposes (*25*, *26*). Bao et al. (*25*) combined bikesharing data and POI data to estimate trip purposes according to the arrival type of POI. However, this study did not consider some influential attributes such as time of day or travel date. The absence of these attributes could contribute to low accuracy. Gong et al. (*26*) used a single temporal attribute with POI data to estimate trip purposes, but other studies found that multiple attributes can have an impact on trip purposes (*1*).

This paper seeks to improve the existing methods of trip purpose estimation by proposing a hybrid-data approach that incorporates data collected automatically from taxi GPS devices and POI data information obtained from the AMAP online map (*27*), an open-data platform in China. Taxi GPS sensors report data in real time to provide a network-wide taxi operation dataset. The difference between taxi data and probe vehicle data is that the taxi data also include information about the boarding and alighting of passengers. Taxi operation data has been used widely for traffic status estimation (*28*, *29*), environment status estimation (*30*), and city logistics route planning (*30*). POI data are a collection of geographic points that may be useful or interesting to users. In this paper, trip purposes of passengers are determined from trip attributes and the type of POI at passengers' final destinations. Different strategies are used and compared to fuse the result from trip attributes and the result from POI. By using both the trip attributes and the POI information, accuracy for trip purposes with similar attributes can be improved.

The remainder of the paper is organized as follows. First, the data are described and the pre-processing procedures are detailed. Next, the methodology is presented. Subsequently, we implement the methodology using three data sources. Finally, the conclusions are drawn and opportunities for further research are outlined.

## Data Description

Three data sources were used for developing and testing the model. The data sources include taxi operation data, POI data, and survey data. The taxi operation data were used to determine the spatial and temporal attributes of passenger trips. The taxi data had to be pre-processed before it could be used because it was in a serial format. The pre-processed taxi data provided the arrival locations of passengers; the POI data provided possible destinations for each passenger around their arrival location. The survey data were used as ground truth data to train and validate the results of the models evaluated in this study.

### Taxi Operations Data

The taxi operation data were provided by the Taxi Company of Chengdu. The taxi operation data were collected between July 2, 2018, and July 8, 2018. The collected sample operation data include 219,942 trips and 20,198 taxis in Chengdu city. The taxi operation data consist of the vehicle identification number, GPS tracking data, and the taxi service state. Each vehicle has a unique vehicle identification number. The GPS tracking data record the duration that the on-vehicle-device is triggered and the real-time latitude and longitude. The service state field reported whether the taxi was in service or not. When the taxi was occupied by a passenger, the service state was set to "in service." When the passenger departed the taxi, the service state was set to "out of service." The boarding and arrival times were obtained when the service state changed. The arrival location was determined based on the latitude and longitude at the arrival time. Trip attributes have proven to be key inputs for trip purpose estimation models according to previous studies (*1*, *9*, *16*, *31*). In this paper, two types of attributes are considered: temporal attributes and spatial attributes. The temporal attributes consist of the day of the week, boarding time, arrival time, and travel time, whereas the spatial attribute is the arrival location.

The trip attributes are determined by the timestamps of the boarding and arrival nodes. Therefore, we need to determine the boarding and arrival timestamps for each passenger. The taxis report time in a serial format. Therefore, the timestamps were converted from this serial format into standard time-of-day format. The service statue values in the taxi database depict whether the taxi is occupied or vacant. More specifically, a value of one means the taxi is occupied, and zero otherwise. Thus, the boarding and arrival timestamps were estimated from the variation of the service state value.

Let a vector $V = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be a set of $n$ taxis and $M = \{\sigma_1, \sigma_2, \ldots, \sigma_m\}$ be a set of timestamps of the taxi operation data of a vehicle. Each element in set $V$ represents a unique vehicle in the taxi operation data, and each element in set $M$ represents a unique timestamp of the collected taxi operation data for each vehicle. The passenger boarding and arrival timestamps, as formulated in Equations 1 and 2, were obtained according to the value change of the service state.

$$S_{board} = \{\omega_\tau \epsilon V, \sigma_\vartheta \in M | p^{\omega_\tau}_{\sigma_\vartheta} - p^{\omega_\tau}_{\sigma_\vartheta - 1} = 1\} \quad \tau \in \{1, 2, 3, \ldots, n\}, \vartheta \in \{2, 3, \ldots, m\} \tag{1}$$

$$S_{alight} = \left\{\omega_\tau \epsilon V, \sigma_\vartheta \in M | p^{\omega_\tau}_{\sigma_\vartheta - 1} - p^{\omega_\tau}_{\sigma_\vartheta} = 1\right\} \quad \tau \in \{1, 2, 3, \ldots, n\}, \vartheta \in \{2, 3, \ldots, m\} \tag{2}$$

where $S_{board}$ is the set of timestamps of boarding passengers and $S_{alight}$ is the set of timestamps of arriving passengers. $p^{\omega_\tau}_{\sigma_\vartheta}$ is the value of service state for vehicle $\omega_\tau$ at timestamps $\sigma_\vartheta$.

Equations 1 and 2 determine the timestamps of boarding and arriving passengers from the dataset, respectively.

Consider $U^{\omega_\tau} = \left\{u_1^{\omega_\tau}, u_2^{\omega_\tau}, \ldots, u_i^{\omega_\tau}\right\}$ to be the set of passengers for vehicle $\omega_\tau$. Based on the timestamps of boarding and arriving passengers, the trip attributes were obtained by

$$a^{bt}_\gamma \in S_{board} \quad \gamma \in U^{\omega_\tau} \tag{3}$$

$$a^{at}_\gamma \in S_{alight} \quad \gamma \in U^{\omega_\tau} \tag{4}$$

It is worth noting that the $a^{at}_\gamma$ attribute consists of two parts: the arrival time of the trip and the date of the trip. The trip date can then be converted to the day of the week using a calendar.

$$a^{at}_\gamma \overset{calendar}{\rightarrow} a^{w}_\gamma \quad \gamma \in U^{\omega_\tau} \tag{5}$$

where $a^{bt}_\gamma$ is the boarding time and $a^{at}_\gamma$ is the arrival time. The attribute $a^{w}_\gamma$ depicts the day of the week of the trip.

The travel time can be obtained according to the boarding time and arrival time.

$$t_\gamma = a^{at}_\gamma - a^{bt}_\gamma \quad \gamma \in U^{\omega_\tau} \tag{6}$$

where $t_\gamma$ is the travel time.

Constraints 3–6 depict the temporal attribute extraction process based on a determined series number of timestamps of boarding and arrival.

Based on the arrival time for each passenger, it is easy to determine the spatial attribute, that is, the latitude $a^{la}_\gamma$ and longitude $a^{lo}_\gamma$ for traveler $\gamma$. We can get the longitude and latitude from the taxi operation data according to the arrival timestamps. With all of the attributes extracted, the data could then be used as a direct input into the model.

### POI Data

POI data are a collection of geographic points that may be useful or interesting to users. The POI data used in this study were obtained from the AMAP online platform (*27*). The POI data are an open dataset, because it is made readily available to the public on an online platform. Online platforms contain a wealth of POI information. Users only need to apply for a free key online from the company to receive open-access privileges so that they can go through data information from the platform. If a traveler is travelling to a specific destination (i.e., not wandering), then the traveler's destination can be assumed to be a POI (*3*, *32*). POI data include four main fields: name of the destination,

POI category (e.g., health care services), longitude, and latitude. With details of the POI category, trip purpose can be estimated. For example, the POI category of "health care services" indicates that the traveler's trip purpose is likely to be "medical service." Thirteen different POI categories are used in this study, including dining, leisure and entertainment, business, shopping, transportation facilities, education, residence, domestic service, recreation, health care services, government, and accommodation service.

### Survey Data

The survey data were collected from a questionnaire survey conducted in Chengdu, China, from July 2, 2018 to July 8, 2018. A total of 1,083 survey responses were recorded. Boarding time, arrival time, travel time, day of the week, arrival location, destination, and trip purpose for each survey respondent were collected through the survey. Each taxi's vehicle identification number was also collected to match the survey data with the taxi operation data. During the survey process, boarding time, arrival time, and travel time were obtained from the taximeter. The destinations and trip purposes were collected by consulting the passengers. The arrival location was recorded by a portable GPS device.

## Methodology

Figure 1 shows the framework of the trip purpose estimation hybrid-data approach. First, the temporal and spatial attributes were extracted from taxi operations data. Two methods, a temporal attribute–based method and a spatial attribute and POI–based method, were used to estimate trip purposes. The first method (shown in blue on the left in Figure 1) used the temporal attributes extracted from taxi operations data (detailed in the section on data description of taxi operations data) to estimate trip purposes using the Bayes model (*1*). The second method (shown in red on the right in Figure 1) merged the spatial attributes extracted from the taxi operations data (as detailed in the section on data description of taxi operations data) with the POI data (as detailed in the section on data description of POI data) to estimate trip purposes. This merging process used either a distance-based technique or a category-based technique to merge spatial attributes with the POI data. Finally, the results of these two methods of trip purpose estimation were fused together to obtain the final estimated results. The fusion process used either a product of probabilities technique or a set operation technique to combine the estimated trip purposes.

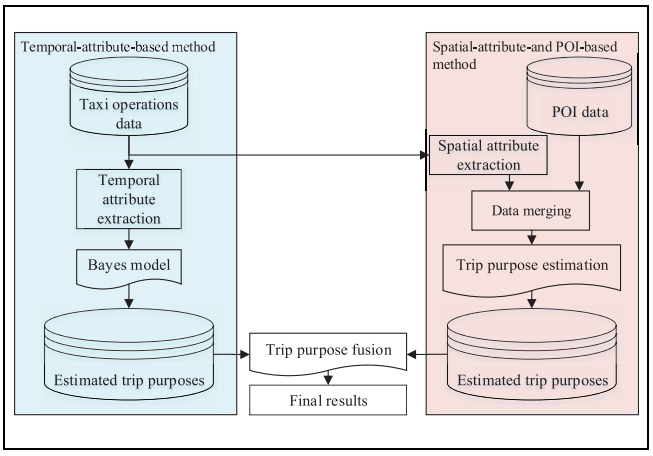


**Figure 1.** The trip purpose estimation framework.

### Bayes Model

The Bayes model is a critical component used in the temporal attribute–based method to estimate trip purposes using the temporal attributes extracted from the taxi operations data (as detailed in the section on data description of taxi operations data). According to a previous study by Kusakabe and Asakura (*1*), the Bayes model can be used to estimate trip purposes using only the trip's temporal attributes. The following is a mathematical expression of the Bayes model used for trip purpose estimation. Let a vector $\boldsymbol{A}_\gamma = \left\{a_\gamma^1, a_\gamma^2, \ldots, a_\gamma^k\right\}$ be a set of temporal trip attributes. Each element $a_\gamma^\vartheta$ ($\vartheta \in \{1, 2, \ldots, k\}$) in set $\boldsymbol{A}_\gamma$ represents a unique temporal trip attribute derived from taxi operation data. The Bayes model was used and can be mathematically expressed as

$$P\left(pur_\gamma^\omega \middle| \boldsymbol{A}_\gamma\right) = \frac{P\left(\boldsymbol{A}_\gamma \middle| pur_\gamma^\omega\right) P\left(pur_\gamma^\omega\right)}{P\left(\boldsymbol{A}_\gamma\right)} \tag{7}$$

where $P\left(pur_\gamma^\omega \middle| \boldsymbol{A}_\gamma\right)$ is the posterior probability and $P(pur_i)$ is the prior probability

$$P\left(\boldsymbol{A}_\gamma \middle| pur_\gamma^\omega\right) P\left(pur_\gamma^\omega\right) = P\left(pur_\gamma^\omega\right) \prod_{\vartheta=1}^{k} P\left(a_\gamma^\vartheta \middle| pur_\gamma^\omega\right) \tag{8}$$

Therefore

$$P\left(pur_\gamma^\omega \middle| \boldsymbol{A}_\gamma\right) = \frac{P\left(pur_\gamma^\omega\right) \prod_{\vartheta=1}^{k} P\left(a_\gamma^\vartheta \mid pur_\gamma^\omega\right)}{P\left(\boldsymbol{A}_\gamma\right)} \tag{9}$$

where $pur_\gamma^\omega$ indicates the trip purpose of the passenger $\gamma$ is trip purpose $\omega$. $P\left(pur_\gamma^\omega \boldsymbol{A}_\gamma\right)$ is the probability that the

trip purpose of the passenger $\gamma$ is estimated as $\omega$ according to the trip temporal attributes. In the study by Kusakabe and Asakura (*1*), $\widehat{pur}_{\gamma}^{\omega}(\boldsymbol{A}_{\gamma}) = \arg maxP\left(pur_{\gamma}^{\omega}\boldsymbol{A}_{\gamma}\right)$ is used as the trip purpose when only the temporal attributes are used to estimate the trip purpose.

## Data Merging and Trip Purpose Estimation

For the spatial attribute and POI–based method, two data sources were used to estimate trip purposes. However, before trip purposes could be estimated, a technique to merge the two datasets had to be developed. A distance-based technique and a category-based technique were proposed to calculate the trip purpose probability distribution according to POI and arrival information. In the proposed distance-based technique, we assumed that the passenger chose the POI as their destination according to the distance between the arrival location and the POI location; in the category-based technique, we assumed that the passenger chose the destination according to the number of same category POIs. A threshold value is defined as $r$ to depict the maximal distance between the destination and arrival locations. The threshold value $r$ is set to determine the possible final destinations according to the survey data. In this paper, we define $r = 50$ m according to our survey. In the survey, the alighting location and the destination of each passenger are collected, which can be used to calculate the walking distance. These walking distances were evaluated, and it was determined that 50 m best matched measured conditions for the radius $r$.

Figure 2 presents five POIs, in which four of the POIs are out of the circle plotted by the arrival location and threshold value of $r$. Only the second POI is in the circle. Therefore, we determine this node as the destination, and the trip purpose can be estimated according to the category of this POI.

However, there could be more than one possible destination. Therefore, two techniques were used to fuse the trip purposes derived from the temporal attribute–based method with the trip purposes derived from the spatial attribute and POI–based method: the first fusion strategy used was the product of probabilities and the second strategy used was the set operation method. Therefore, trip purposes derived from the spatial attribute and POI–based method needed to be formatted to be input into one of these two fusion methods, as seen in Figure 1 in the "Trip purpose fusion" process.

*Product of Probabilities Formatting.* All POIs in the circle plotted by the arrival location and threshold value $r$ are to be considered as possible destinations. Let $O_{\gamma} = \{o_1, o_2, \ldots, o_h\}$ be the set of POIs of passenger $\gamma$ in the circle and $\partial \in O_{\gamma}$ be a specific type of POI. The two trip purpose fusion techniques described, distance-based and category-based, are developed to estimate trip purposes using POI data.

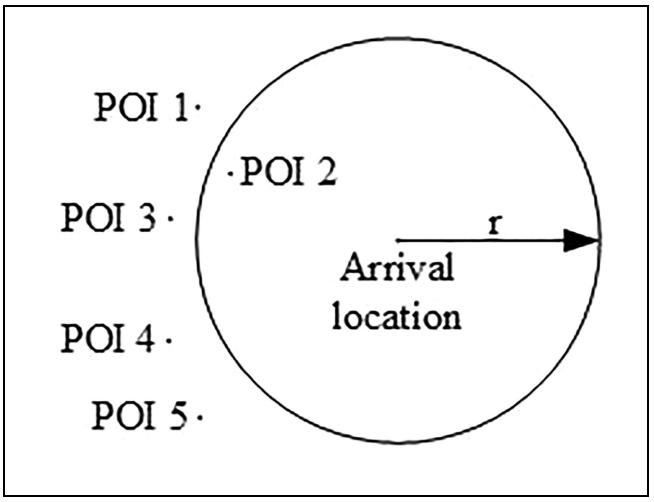


**Figure 2.** An illustration of the threshold value of *r*.
*Note*: POI = point of interest.

$$P(pur_{\gamma}^{\omega}|poi) = \frac{\sum_{\partial \in o^{\omega}} \frac{1}{dis_{\gamma}^{\partial}}}{\sum_{\partial \in O_{\gamma}} \frac{1}{dis_{\gamma}^{\partial}}} \quad \gamma \in U^{\omega_{\tau}} \tag{10}$$

$$P(pur_{\gamma}^{\omega}|poi) = \frac{num_{poi}^{\omega}}{num_{O_{\gamma}}} \quad \gamma \in U^{\omega_{\tau}} \tag{11}$$

where $dis_{\gamma}^{\partial}$ is the distance between arrival location and POI $\partial$ for passenger $\gamma$. $num_{poi}^{\omega}$ is the number of POIs that belong to category $\omega$ and $num_{O_{\gamma}}$ is the number of elements in set $O_{\gamma}$. $P(pur_{\gamma}^{\omega}|poi)$ is the probability that the trip purpose of passenger $\gamma$ is estimated as $\omega$ according to the spatial trip attributes and POI information.

Therefore, Equations 10 and 11 describe the probability of estimated trip purposes according to POI information by using either the distance-based technique or the category-based technique, respectively.

*Set Operation Formatting.* To determine the set to be input, all the categories in the circle were considered to be potential trip purposes, as detailed in Equation 12.

$$pur_{\gamma}^{\omega}(poi) = O_{\gamma} \quad \gamma \in U^{\omega_{\tau}} \tag{12}$$

where $pur_{\gamma}^{\omega}(poi)$ is the estimated trip purposes set.

## Trip Purpose Fusion

By fusing the estimated trip purpose results of both the temporal attribute–based method and the spatial attribute and POI–based method, the approach can achieve a higher accuracy because both the temporal and the spatial factors are considered. Two different techniques, the

product of the probabilities technique and the set operation technique, were used to fuse the trip purpose estimation results.

*Product of Probabilities.* The product of probabilities can be expressed by Equation 13.

$$P\left(pur_{\gamma}^{\omega}|poi, \boldsymbol{A}_{\gamma}\right) = P\left(pur_{\gamma}^{\omega}\boldsymbol{A}_{\gamma}\right) \cdot P(pur_{\gamma}^{\omega}|poi) \quad (13)$$

Equation 13 describes how to calculate the probability that the trip purpose of passenger γ is estimated as trip purpose ω according to the spatio-temporal trip attributes and POI information. The $\widehat{pur_{\gamma}^{\omega}}(poi, \boldsymbol{A}_{\gamma}) = \arg maxP\left(pur_{\gamma}^{\omega}|poi, \boldsymbol{A}_{\gamma}\right)$ is used as the estimated trip purpose.

*Set Operation.* When using the set operation technique (i.e., union set operation as the fusion method), all the POIs surrounding the passenger arrival location are considered as potential trip purposes, which means the POI data can have several estimated results. In this manner, a set of estimated trip purposes is generated for each passenger. Equations 14 through 17 detail the four possible cases to be considered when this set operation technique is used.

$$\textbf{Case1}\ \widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma}) \cap pur_{\gamma}^{\omega}(poi) = \emptyset \quad pur_{\gamma}^{\omega}(poi) = \emptyset \quad (14)$$

$$\textbf{Case2}\ \widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma}) \cap pur_{\gamma}^{\omega}(poi) = \emptyset \quad pur_{\gamma}^{\omega}(poi) \neq \emptyset \quad (15)$$

$$\textbf{Case3}\ \widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma}) \cap pur_{\gamma}^{\omega}(poi) = \widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma}) \quad \widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma}) = pur_{\gamma}^{\omega}(poi) \quad (16)$$

$$\textbf{Case4}\ \widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma}) \cap pur_{\gamma}^{\omega}(poi) = \widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma}) \quad \widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma}) \neq pur_{\gamma}^{\omega}(poi) \quad (17)$$

As the estimated results according to the POI information, $pur_{\gamma}^{\omega}(poi)$ can produce a multiple elements set or an empty set. The union set operation of $pur_{\gamma}^{\omega}(poi)$ and $\widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma})$ can yield two results: an empty set and $\widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma})$. When the $pur_{\gamma}^{\omega}(poi) = \emptyset$, we can obtain Equation 14 because the union of any empty set is an empty set. In Equation 15 when $pur_{\gamma}^{\omega}(poi) \neq \emptyset$ but no elements in $pur_{\gamma}^{\omega}(poi)$ are the same with the $\widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma})$, the union is still an empty set. Equation 16 is a perfect result showing that the estimation results by POI information only have one element and it is the same with $\widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma})$. In Equation 17, the estimation results by POI information have several elements and one that is the same as $\widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma})$. Obviously, the union set results in Equations 16 and 17 are $\widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma})$.

For Equations 16 and 17 we can determine the set operation results as $\widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma})$ according to union set operation. For Equation 14, as no spatial information is available, we can only estimate trip purposes according to the temporal information. Consequently, the set operation results for Equation 14 is also $\widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma})$.

However, for Equation 15, trip purposes estimated using temporal attributes do not overlap with the trip purposes estimated using POI. In this case, there are contradictory results from different sources, based on the assumption that the passenger arrival location is close to a unique POI. We can determine that the estimated trip purpose using temporal trip attributes is incorrect, because no POI with the corresponding category is close to the arrival location. Therefore, we need to determine a method to correct the estimated result by using POI information. In summary, the set operation technique can be expressed as detailed in Equation 18.

$$pur_{\gamma}(\boldsymbol{A}_{\gamma}, poi_{\gamma}) = \begin{cases} \widehat{pur_{\gamma}^{\omega}}(\boldsymbol{A}_{\gamma}) & \textbf{Case}1 \cup \textbf{Case}3 \cup \textbf{Case}4 \\ pur_{\gamma}^{\omega}(poi) & \textbf{Case}2 \end{cases} \quad (18)$$

In Case 2, the set could be more than one trip purpose. If this is the case, the trip purpose can be determined according to either the distance-based or the category-based technique; in this paper, the distance-based technique was used.

**Table 1.** Estimation Accuracy for Different Data Splits

| Data split (%) | 10–90 | 20–80 | 30–70 | 33–67 |
|---|---|---|---|---|
| Average estimation accuracy (%) | 80.8 | 79.3 | 80.1 | 80.6 |

## Implementations

### Test-Train Split Impacts

Determining the test-train data splits for model training and model validation is important. To investigate the impacts of varying training and validation data splits on the estimation results, the proposed method is evaluated using several different data splits for training and validation. Table 1 summarizes the average estimation accuracy for each data split. The estimation accuracy is measured by $\frac{\sum_{\omega=1}^{I} C_{\omega}}{I}$, where $C_{\omega}$ is the estimated accuracy for purpose ω, and $I$ is the number of trip purposes. The estimation accuracy is evaluated by $\frac{N_c^{\omega}}{N_t^{\omega}} \times 100\%$, where $N_c^{\omega}$ and $N_t^{\omega}$ are the correct estimation trips for total number of

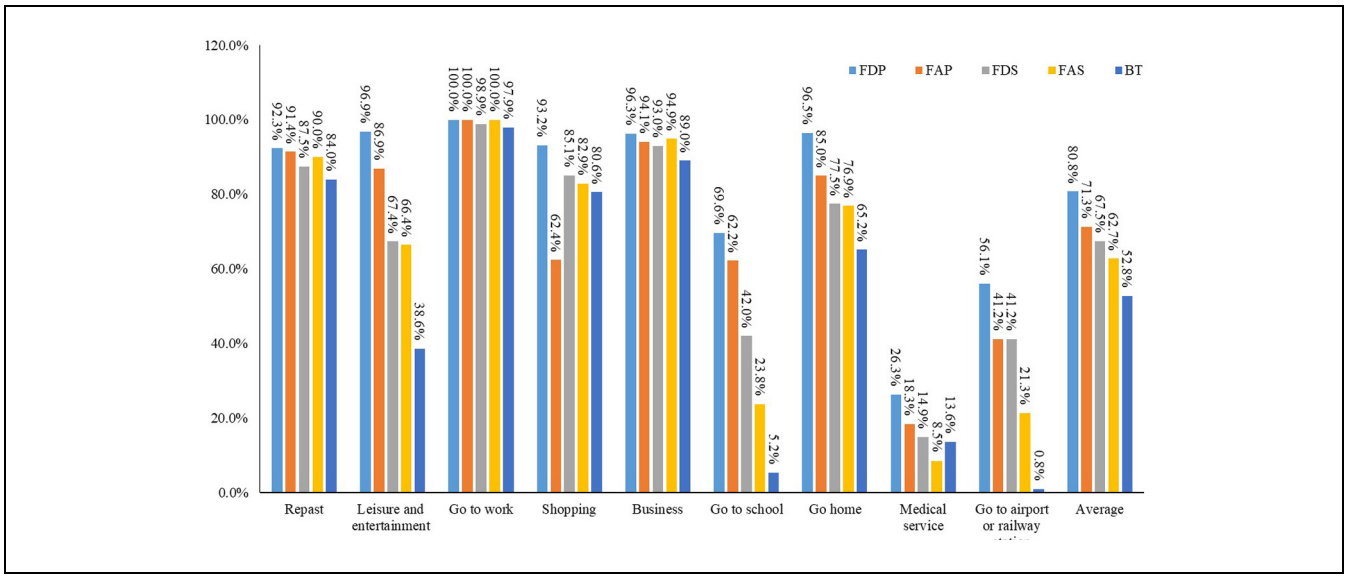


**Figure 3.** Estimation accuracy of each estimation strategy.
*Note*: F = fusion; D = distance-based; A = category-based; P = product of probabilities; S = set operation; BT = Bayes temporal. The product of probabilities strategy is used to fuse the results for strategies FDP and FAP and the set operation is used to fuse the results for strategies FDS and FAS. The distance-based technique is used for FDP and FDS. The category-based technique is applied in FAP and FAS.

trips for purpose ω. The different data splits were all evaluated using the product of probabilities strategy combined with the distance-based technique for consistency.

### *Validation*

Survey data were used to validate the results. The survey data were randomly divided into training and testing sets using K-fold cross-validation. For each fold, 90% of the survey data was used to train the estimation model and the remaining 10% of the survey data was used to validate the trained estimation model. The validation process was repeated 10 times using the same operation with the cross-validation process. Figure 3 illustrates the estimation accuracy of each different estimation strategy for each estimated trip purpose. Bayes temporal is the estimated trip purposes when using only the temporal attribute–based method. All the remaining strategies use the trip purpose fusion of the temporal attribute–based method and the spatial attribute and POI–based method.

Figure 3 shows that higher estimation accuracy is achieved when using the fused trip purpose estimations instead of using only the temporal attributes–based method. This makes sense intuitively, because more information is made available when making the estimation. Furthermore, because the POI data are open source, the addition of POI data increases the estimation's accuracy without accruing any added cost. Figure 3 shows that the product of probabilities and distance-based fusion strategy has the highest accuracy of all the strategies analyzed, which is mostly because of the addition of POI data. By considering both the spatial and temporal attributes, different trip purposes with similar trip attributes can be more accurately distinguished. For, example, the "go to work" and "go to school" trip purposes are easily confused when considering only the temporal attributes. Figure 3 also shows that the product of probabilities strategy performs better than the set operation strategy when the POI information is incorporated. Furthermore, the distance-based technique performs better than the category-based strategy.

## Conclusions

Trip purposes are critical information for transportation planners when making data-driven decisions such as designing transportation networks. However, the existing approach using single sourced data to estimate trip purpose can be further improved. The objectives of this paper are to incorporate an open-source data, POI data, and to develop a hybrid-data approach to fuse the estimated trip purposes from different data sources to improve the final estimation accuracy. The results show the estimation accuracy can be improved, because different trip purposes with similar temporal attributes can be more accurately distinguished when POI data are used. We also analyzed several different strategies to incorporate POI data. The distance-based strategy performed the best at determining the potential destination of each passenger, and the product of probabilities strategy performed better than set operations when fusing the estimation results. The hybrid-data approach proposed in

this paper is transferrable to other locations by using data sources such as Google Maps or Yelp to obtain POI data.

The hybrid-data approach developed in this paper can increase the accuracy of trip purpose estimation, but still has a few limitations. First, the estimation accuracy is dependent on the quality of the data collected. In other words, both temporal and spatial data have to be of acceptable quality to provide accurate estimated trip purposes. A second limitation is that GPS data are still needed to estimate trip purposes, and so it may be difficult to collect a sample large enough to be representative if the study area is large. Furthermore, in locations with mixed land use containing several POI categories, it could be difficult for the model to accurately predict the correct trip purpose.

Based on the results of this study, several recommendations for future work can be made. One possible future study could be implementing other data sources to extract additional attributes that are related to trip purposes and further improving the estimation accuracy. By using other data sources, it may be possible to increase the estimation accuracy in mixed land-use scenarios. Another future study could be to apply the open data source, POI data, to other topics such as travel behavior analysis.

## Acknowledgments

The authors thank the Taxi Company of Chengdu for providing data for the study.

## Author Contributions

The authors confirm contributions to the paper as follows: study conception and design: X. Luo, Y. Jiang; data collection: Y. Jiang; analysis and interpretation of results: X. Luo, Y.-J. Wu; draft manuscript preparation: A. Cottam, Y.-J. Wu. All authors reviewed the results and approved the final version of the manuscript.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the Chinese National Natural Science Fund (Serial No.51578465 and 71402149), Doctoral Innovation Fund Program of Southwest Jiaotong University (DCX201826), and Chongqing Municipal Transportation Engineering Key Laboratory Open Project (2018TE04).

## ORCID iDs

Xiaoling Luo https://orcid.org/0000-0002-2113-7650
Adrian Cottam https://orcid.org/0000-0001-5654-4347
Yao-Jan Wu https://orcid.org/0000-0002-0456-7915

## References

1. Kusakabe, T., and Y. Asakura. Behavioural Data Mining of Transit Smart Card Data: A Data Fusion Approach. *Transportation Research Part C: Emerging Technologies*, Vol. 46, 2014, pp. 179–191. https://doi.org/10.1016/j.trc.2014.05.012.
2. Shen, L., and P. R. Stopher. A Process for Trip Purpose Imputation from Global Positioning System Data. *Transportation Research Part C: Emerging Technologies*, Vol. 36, 2013, pp. 261–267. https://doi.org/10.1016/j.trc.2013.09.004.
3. Xiao, G., Z. Juan, and C. Zhang. Detecting Trip Purposes from Smartphone-Based Travel Surveys with Artificial Neural Networks and Particle Swarm Optimization. *Transportation Research Part C: Emerging Technologies*, Vol. 71, 2016, pp. 447–463. https://doi.org/10.1016/j.trc.2016.08.008.
4. Boerma, J. T., R. E. Black, A. E. Sommerfelt, S. O. Rutstein, and G. T. Bicego. Accuracy and Completeness of Mothers' Recall of Diarrhoea Occurrence in Pre-School Children in Demographic and Health Surveys. *International Journal of Epidemiology*, Vol. 20, No. 4, 1991, pp. 1073–1080. https://doi.org/10.1007/s11116-010-9290-0.
5. Bohte, W., and K. Maat. Deriving and Validating Trip Purposes and Travel Modes for Multi-Day GPS-Based Travel Surveys: A Large-Scale Application in the Netherlands. *Transportation Research Part C: Emerging Technologies*, Vol. 17, 2009, pp. 285–297. https://doi.org/10.1016/j.trc.2008.11.004.
6. Chu, K. K. A., and R. Chapleau. Enriching Archived Smart Card Transaction Data for Transit Demand Modeling. *Transportation Research Record: Journal of the Transportation Research Board*, 2008. 2063: 63–72.
7. Furth, P., J. Strathman, and B. Hemily. Part 4: Marketing and Fare Policy: Making Automatic Passenger Counts Mainstream: Accuracy, Balancing Algorithms, and Data Structures. *Transportation Research Record Journal of the Transportation Research Board*, 1927, 2005: 205–216.
8. Munizaga, M. A., and C. Palma. Estimation of a Disaggregate Multimodal Public Transport Origin–Destination Matrix from Passive Smartcard Data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies*, Vol. 24, 2012, pp. 9–18. https://doi.org/10.1016/j.trc.2012.01.007.
9. Trépanier, M., N. Tranchant, and R. Chapleau. Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, Vol. 11, No. 1, 2007, pp. 1–14. https://doi.org/10.1080/15472450601122256.
10. Matas, A. Demand and Revenue Implications of an Integrated Public Transport Policy: The Case of Madrid. *Transport Reviews*, Vol. 24, No. 2, 2004, pp. 195–217. https://doi.org/10.1080/0144164032000107223.

11. El-Geneidy, A. M., Y. Krizek, and J. J. K. Horning. Analyzing Transit Service Reliability Using Detailed Data from Automatic Vehicular Locator Systems. *Journal of Advanced Transportation*, Vol. 45, No. 1, 2011, pp. 66–79. https://doi.org/10.1002/atr.
12. Zhan, X., S. Hasan, S. V. Ukkusuri, and C. Kamga. Urban Link Travel Time Estimation Using Large-Scale Taxi Data with Partial Information. *Transportation Research Part C: Emerging Technologies*, Vol. 33, 2013, pp. 37–49. https://doi.org/10.1016/j.trc.2013.04.001.
13. Zheng, F., and H. Van Zuylen. Urban Link Travel Time Estimation Based on Sparse Probe Vehicle Data. *Transportation Research Part C: Emerging Technologies*, Vol. 31, 2013, pp. 145–157. https://doi.org/10.1016/j.trc.2012.04.007.
14. Ferreira, N., J. Poco, H. T. Vo, J. Freire, and C. T. Silva. Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 19, 2013, pp. 2149–2158. https://doi.org/10.1109/TVCG.2013.226.
15. Tsai, T. H., C. K. Lee, and C. H. Wei. Neural Network Based Temporal Feature Models for Short-Term Railway Passenger Demand Forecasting. *Expert Systems with Applications*, Vol. 36, No. 2, 2009, pp. 3728–3736. https://doi.org/10.1016/j.eswa.2008.02.071.
16. Lu, Y., and L. Zhang. Imputing Trip Purposes for Long-Distance Travel. *Transportation*, Vol. 42, No. 4, 2015, pp. 581–595. https://doi.org/10.1007/s11116-015-9595-0.
17. Devillaine, F., M. Munizaga, and M. Trépanier. Detection of Activities of Public Transport Users by Analyzing Smart Card Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2012. 2276: 48–55.
18. Jun, C., and Y. Dongyuan. Estimating Smart Card Commuters Origin–Destination Distribution Based on APTS Data. *Journal of Transportation Systems Engineering and Information Technology*, Vol. 13, No. 4, 2013, pp. 47–53. https://doi.org/10.1016/S1570-6672(13)60116-6.
19. Lee, S. G., and M. Hickman. Trip Purpose Inference Using Automated Fare Collection Data. *Public Transport*, Vol. 6, No. 1–2, 2014, pp. 1–20. https://doi.org/10.1007/s12469-013-0077-5.
20. Zhou, J., E. Murphy, and Y. Long. Commuting Efficiency in the Beijing Metropolitan Area: An Exploration Combining Smartcard and Travel Survey Data. *Journal of Transport Geography*, Vol. 41, 2014, pp. 175–183. https://doi.org/10.1016/j.jtrangeo.2014.09.006.
21. Chu, K., and R. Chapleau. Augmenting Transit Trip Characterization and Travel Behavior Comprehension. *Transportation Research Record: Journal of the Transportation Research Board*, 2010. 2183: 29–40.
22. Alsger, A., A. Tavassoli, M. Mesbah, L. Ferreira, and M. Hickman. Public Transport Trip Purpose Inference Using Smart Card Fare Data. *Transportation Research Part C: Emerging Technologies*, Vol. 87, 2018, pp. 123–137. https://doi.org/10.1016/j.trc.2017.12.016.
23. Kumar, P., A. Khani, and Q. He. A Robust Method for Estimating Transit Passenger Trajectories using Automated Data. *Transportation Research Part C: Emerging Technologies*, Vol. 95, 2018, pp. 731–747.
24. He, L., B. Agard, and M. Trépanier. A Classification of Public Transit Users with Smart Card Data Based on Time Series Distance Metrics and a Hierarchical Clustering Method. *Transportmetrica A: Transport Science*, Vol. 16, No. 1, 2020, pp. 56–75. https://doi.org/10.1080/23249935.2018.1479722.
25. Bao, J., C. Xu, P. Liu, and W. Wang. Exploring Bikesharing Travel Patterns and Trip Purposes Using Smart Card Data and Online Point of Interests. *Networks and Spatial Economics*, Vol. 17, No. 4, 2017, pp. 1231–1253. https://doi.org/10.1007/s11067-017-9366-x.
26. Gong, L., X. Liu, L. Wu, and Y. Liu. Inferring Trip Purposes and Uncovering Travel Patterns from Taxi Trajectory Data. *Cartography and Geographic Information Science*, Vol. 43, No. 2, 2016, pp. 103–114. https://doi.org/10.1080/15230406.2015.1014424.
27. AutoNavi Software Co., Ltd. Chinese Web Mapping. 2018. https://www.amap.com/. Accessed May 5, 2018.
28. Cheu, R. L., C. Xie, and D. Lee. Probe Vehicle Population and Sample Size for Arterial Speed Estimation Measurement Techniques. *Simulation*, Vol. 17, 2002, pp. 53–60.
29. Wang, X., H. Liu, R. Yu, B. Deng, X. Chen, and B. Wu. Exploring Operating Speeds on Urban Arterials Using Floating Car Data: Case Study in Shanghai. *Journal of Transportation Engineering*, Vol. 140, No. 9, 2014, p. 04014044. https://doi.org/10.1061/(asce)te.1943-5436.0000685.
30. Ehmke, J. F., S. Meisel, and D. C. Mattfeld. Floating Car Based Travel Times for City Logistics. *Transportation Research Part C: Emerging Technologies*, Vol. 21, 2012, pp. 338–352. https://doi.org/10.1016/j.trc.2011.11.004.
31. Barry, J., R. Newhouser, A. Rahbee, and S. Sayeda. Origin and Destination Estimation in New York City with Automated Fare System Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2002. 1817: 183–187.
32. Feng, T., and H. J. P. Timmermans. Detecting Activity Type from GPS Traces Using Spatial and Temporal Information. *European Journal of Transport and Infrastructure Research*, Vol. 15, No. 4, 2015, pp. 662–674. https://doi.org/10.1007/978-3-319-11463-7_5.