

Urban Mobility Analytics: A Deep Spatial-Temporal Product Neural Network for Traveler Attributes Inference

Can Li^a, Lei Bai^a, Wei Liu^{a,b,*}, Lina Yao^a and S Travis Waller^b

^aSchool of Computer Science and Engineering, University of New South Wales, Sydney, NSW 2052, Australia

^bResearch Centre for Integrated Transport Innovation, School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW 2052, Australia

ARTICLE INFO

Keywords:

Travel Pattern Recognition
Traveler Attributes Inference
Public Transport
Deep Learning


ABSTRACT

This study examines the potential of using smart card data in public transit systems to infer attributes of travelers, thereby facilitating a more user-centered public transport service design while reducing the use of expensive and time-consuming travel surveys. This is challenging since travel behaviors vary significantly over the population, space, and time and developing meaningful links between them and traveler attributes are not trivial. To achieve this, we conduct an extensive analysis of spatio-temporal travel behavior patterns using smart card data from the Greater Sydney area (Opal card), and then develop a Hybrid Neural Network to utilize spatial and temporal dependencies in the dataset. In particular, we first empirically analyze passengers' movements and mobility patterns from both spatial and temporal perspectives and design a set of discriminative features to characterize the patterns. We then propose a deep-learning-based framework to investigate spatial and temporal features in order to infer traveler attributes. The proposed modeling framework consists of two components, i.e., a Product-based Spatial-Temporal Module (**PSTM**) and an Auto-Encoder-based Compression Module (**AECM**). **PSTM** encodes the relationships across a variety of features while **AECM** derives useful spatial information from a transit stop matrix. The proposed model is tested and evaluated using a large-scale public transport dataset in the Greater Sydney area to infer two attributes of passengers, i.e., the age group and residential area. The experimental results demonstrate the effectiveness of the proposed method against a number of established tools in the literature. The developed techniques can be potentially adapted to other domains where spatio-temporal features are critical, such as commercial/entertainment site selection and urban service planning.

1. Introduction

With the accelerating urbanization, around 70% of the world's population is expected to live in cities by 2050. Urban public transport systems (buses, trains, and ferries, etc.) serve a large number of passengers on a daily basis and play an important role in metropolitan areas. However, current public transport system/service designs are often capacity-maximizing while individual preference is only considered to a limited extent. There is a growing trend to allow a more user-centered public transport system, which better accommodates, e.g., different age groups and the disabled. This requires improvements from at least the following aspects: infrastructures/equipment (vehicles, stations, access facilities, etc.); operation (line planning and scheduling); added services (connection information, entertainment TV programs, and advertisements). To provide such a user-centered public transport system, where individual preference is well accommodated, a critical input is the attributes of travelers. A conventional way is to conduct travel surveys to identify individual travel patterns and preferences, which can be costly and time-consuming. For example, Shifan et al. (2008) proposed to identify travelers' behaviors with the data collected from the Utah Transit Authority household survey. Differently, this study develops methods to infer traveler attributes (e.g., age groups, residential areas) based on individual travel trajectories while avoiding costly and time-consuming large-scale surveys. In particular, we test the effectiveness of the proposed model, a new hybrid spatial-temporal correlation model, for inferring age groups and residential areas of passengers. This has the potential to be utilized to improve transit services to accommodate requirements and preferences of different passenger groups (e.g., the elderly may be less demanding in terms of travel time, while they prefer quiet buses or trains; and young commuters/workers with work trips may be more demanding in terms of travel time and reliability). Moreover, the proposed method for inference may also be used to recover missing information/labels associated with individuals in a dataset. Besides, the generated

*Corresponding author

 wei.liu@unsw.edu.au (W. Liu)

insights from this study on how to link personal attributes/information with observable travel trajectories and mobility patterns may also be incorporated in other application domains with spatio-temporal complexity.

Location information with different levels of time resolution can be regarded as human trajectories with different sampling rates. The digital human trajectories collected by different types of data (e.g., GPS positions of mobile phones, smart card data, etc.) can be utilized to analyze human travel behaviors and mobility patterns (Song et al., 2010) and further infer the attributes of travelers. For example, Gonzalez et al. (2008) indicated that human trajectories exhibit a high degree of temporal-spatial regularity. Therefore, some statistical models and traditional machine learning methods were utilized to uncover travel patterns of passengers based on individual trajectories. Recently, the travel pattern regularity of public transport was analyzed by using the rough set theory and **K-Means++** in Ma et al. (2013) without considering the spatial patterns of transit riders. Sun and Axhausen (2016) applied the probabilistic factorization framework to reveal spatial-temporal patterns of urban mobility, which repeats the estimation process to improve solution quality. Further studies indicated that the mobility patterns extracted from human trajectories are related to passenger attributes (Zhong et al., 2015; Luo et al., 2016; Olmos et al., 2018; Li et al., 2019). Although some researchers studied the spatial and/or temporal travel behaviours based on human trajectories, and analyzed the relations between these behaviours and traveler attributes by traditional clustering methods, e.g., Density-Based Spatial Clustering of Applications with Noise (**DBSCAN**) and **K-means** (Mohamed et al., 2016; Kieu et al., 2014), the complex spatio-temporal inter-correlation among features of mobility patterns have not yet been fully studied and uncovered to infer traveler attributes.

In this study, we propose to uncover mobility patterns of passengers by inferring passenger attributes in public transport systems with the help of large-scale smart card usage data and land use data. We focus on identifying the passengers into three age groups (i.e., adults, seniors, and children) and inferring the residential areas of passengers. In particular, we will take the age group as an example to present the critical features of mobility patterns including both the spatial and temporal information together with an analysis of their relationships. Based on the extracted mobility features and analysis, a hybrid neural network consisting of two components is developed for traveler attributes inference. The first component, the Product-based Spatial-Temporal Module (**PSTM**), is used to analyze and capture the spatial-temporal correlations from the extracted features, where we tested two specific modules, i.e., Inner-Product-based Module (**Inner-PNN**) and Outer-Product-based Module (**Outer-PNN**). The second component, Compression Module (**CM**), is utilized for compressing the transit stop sparse matrix (reflecting spatial patterns of trips), and further extracting useful information and learning the embedding vectors from this matrix, where we also tested two specific models, i.e., the fully connected layers (**FCLs**) and Auto-Encoder-based Compression Module (**AECM**).

The main contributions of this paper are summarized in the following. **(i)** We uncover and extract representative spatial and temporal passenger behavior patterns from a large-scale real-world dataset collected in the largest metropolitan area in Australia (Greater Sydney area). To provide empirical insights regarding mobility patterns associated with different attributes of travelers, we use age group information as an example and quantify the correlations/mapping between the mobility patterns and age groups. The travel pattern analysis is further enhanced by utilizing land use information (Point of Interest or PoI), which enriches the analysis to emphasize both the temporal and spatial dimensions. **(ii)** To the best of our knowledge, this paper is among the earliest to illustrate the potential of inferring individual attributes from observable trajectories based on smart card data of public transport with deep-learning-based methods. In this context, we propose a hybrid Neural Network, which combines **PSTM** and **CM** for the age group and residential area inference. The developed approach can also be utilized to either infer or recover unknown or missing attributes or labels in a dataset, as will be further discussed in Section 6.1. **(iii)** We evaluate the proposed method on a large-scale real-world dataset collected in the largest metropolitan area in Australia (Greater Sydney area) and demonstrate the effectiveness of the proposed method against several baselines and state-of-the-art methods.

The rest of this paper is organized as follows. First, we review some related works in Section 2. Then, we provide detailed introduction to the dataset used, the descriptive data statistics, mobility feature representation, and travel pattern analysis in Section 3. Section 4 presents the proposed inference model and related techniques. The test and evaluation of the proposed method and comparison to other methods are presented in Section 5. Section 6 discusses the potential applications and implications from this study and future research directions, and then concludes the paper.

2. Related Work

In this section, we review works related to this study from two aspects: travel behavior and individual attributes studies with different data sources; and the inference and mining/learning strategies of travel patterns and/or traveler

72 attributes.

73 **(Travel behavior and individual attribute studies)** A branch of studies on travel behaviors are mainly based on
74 surveys. For instance, Axhausen et al. (2002) collected 6-week continuous data from a travel diary survey to show the
75 dynamic changes of travel choices including taking public transport, cycling, walking, and driving. Similarly, based
76 on a household survey, Shiftan et al. (2008) grouped passengers who had similar attitudes toward travel by attitudinal
77 factor score calculated from structural equation modeling.

78 In recent years, with the rapid expansion of data in the digital era, the massive transport data provides new op-
79 portunities to explore behaviours and attributes of travelers. For example, the location check-ins from online social
80 networks became one of the popular sources to record the human mobility patterns for passenger attributes analysis
81 (Zhong et al., 2015). However, it is not always appropriate to use such datasets as representative samples for the whole
82 targeted population. GPS-based data were also utilized to learn travel trajectories for inferring demographic attributes
83 (Wu et al., 2019). However, GPS taggers may not work well in urban areas due to signal multi-path and urban canyon
84 obstructions (Calabrese et al., 2011).

85 Electronic smart cards, as a widely used tool for accessing public transport services (Pelletier et al., 2011), provide
86 valuable ready-to-use passenger transit data and a potentially efficient and effective way to identify travel patterns and
87 traveler attributes. It should be noted that the socioeconomic information of users is often not included during the
88 usage of smart cards for protecting privacy (Cottrill, 2009). Numerous works (Lathia et al., 2010; Kieu et al., 2014;
89 Mohamed et al., 2016; Bai et al., 2019a; Zhao et al., 2020; Wang et al., 2020) used the smart card data to examine travel
90 mode choices, travel time, travel demand, trip purposes of the users, or mobility regularity. Moreover, smart card data
91 and a household survey were also combined to examine the spatio-temporal travel patterns of commuters (Long and
92 Thill, 2015). However, inferring traveler attributes based on smart card data in public transit systems has been rarely
93 studied, which is the main focus of this paper.

94 **(Inference, mining or learning strategies)** Statistical tools and conventional machine learning algorithms have
95 been applied to analyze behaviours and attributes of travelers such as travel choices, travel time, commuting patterns,
96 and personal information with respect to both the spatial and temporal dimensions. For instance, Liu et al. (2014)
97 provided a mobility pattern analysis by integrating the taxi traces and bus transactions to model travelers' mode choices.
98 Kieu et al. (2014) segmented passengers into a number of identifiable types of similar behaviors and needs. Kusakabe
99 and Asakura (2014) estimated the arrival time and duration of stay at the station by naive Bayes Classifier to observe
100 the continuous long-term changes of passengers' trip purposes. The results were validated by the person trip survey
101 data, which was often hard (if not impossible) to achieve in other studies. Furthermore, Zhong et al. (2015) collected
102 a large real-world check-in dataset from an online social network to infer users' demographic attributes by exploiting
103 human mobility patterns through the Location to Profile framework (**L2P**). More recently, Luo et al. (2016) constructed
104 space-time trajectories from the location-based social media data and further found that urban human mobility patterns
105 are affected by the demographic attributes of travelers. The long-term spatio-temporal travel behaviours of commuters
106 were visualized leveraging modified **DBSCAN** algorithm and multi-criteria detection analysis approaches in Ma et al.
107 (2017b). To better understand human mobility patterns and their relationships with socioeconomic status of travelers,
108 Xu et al. (2018) adopted a social stratum model to group the users into various classes based on the mobility indicators
109 (e.g., the number of activity locations and travel diversity).

110 With the widespread use of deep learning in recent years, a growing number of neural-network-based models (e.g.,
111 Recurrent Neural Network, Convolutional Neural Network, and Graph Convolutional Network) have been utilized in
112 spatial-temporal data mining (Wang et al., 2019). In particular, Karlaftis and Vlahogianni (2011) compared statistical
113 methods and Neural Network (**NN**) for modeling and analyzing transport data. They indicated that **NN** had potential
114 and stronger learning ability to capture nonlinear and complex relations among features. Many aspects of transport or
115 traffic issues dealing with spatial-temporal features were explored recently, including those for time-series prediction
116 (e.g., traffic speed prediction (Ma et al., 2015, 2017a; Guo et al., 2020), traffic flow forecasting (Van Lint, 2008;
117 Van Hinsbergen et al., 2009; Li et al., 2017; Chu et al., 2018; Ma and Qian, 2018; Bai et al., 2019b; Liu et al., 2019;
118 Guo et al., 2020; Li et al., 2020), and parking occupancy prediction (Yang et al., 2019)), traffic incident detection (Ren
119 et al., 2018), trajectory representations (Li et al., 2018; Yao et al., 2017), and PoI recommendation (Chang et al., 2018).
120 Although there are quite many deep-learning-based studies for spatial-temporal analysis of transport systems, few of
121 them looked at inferring individual attributes from observed travel behaviors.

122 In general, previous works often explored travel patterns and passenger attributes by traditional machine learning
123 or statistical tools, where the complex and non-linear spatial-temporal features and relations in the datasets are not
124 necessarily well captured. The deep-learning-based methods were mostly used for prediction, but not for passenger

Table 1

A summary of related attributes in the smart transit card dataset

Attribute	Definition
Hashed Card Identification Number	A unique identification number for the Opal card
Passenger Type Code	Description of card type
Journey Segment Start Date Foreign Key	Date of the journey
Journey Segment Start Time	The time the journey segment commenced
Journey Segment Duration Seconds	The length of time between the journey segment start time and end time
Transit Stop Type Code	A description of the type of transit stop
Inter-modal Transfer Type Code	A code representing the inter-modal transfer type related to the journey segment
Transfer Indicator	Indication of whether the tag of the journey segment represents a transfer
Transit Stop Name	A description of the transit stop where the journey segment occurred
Latitude and Longitude Value	The latitude and longitude coordinate indicating the position of the journey location

125 travel patterns analysis. Moreover, little has been examined in relation to inferring passenger attributes based on
 126 observed travel trajectories. Different from the literature, deep learning models are utilized in this paper to explore the
 127 relevance among diverse spatial and temporal mobility features for traveler attributes inference.

128 3. Data Description and Behavioural Features

129 This section describes the real-world dataset from Sydney used in this study and mobility feature extraction details.
 130 Then, spatial and temporal mobility features are analyzed based on age groups as an example to illustrate the mapping
 131 between mobility patterns and individual attributes.

132 3.1. Dataset Description

133 **Smart Card Dataset** is collected from Opal (<https://www.opal.com.au/en/about-opal/>), the electronic smart card
 134 ticket system in Sydney covering main public transport services (buses, trains, ferries, and light rails). The dataset is
 135 collected from 01/Apr/2017 to 30/Jun/2017 and records 171.77 million trips covering 6.374 million users. The dataset
 136 does not involve personal information that can be used to identify individuals.

137 There are ten attributes in the dataset that are particularly relevant to this study, which are summarized in Table 1.
 138 Other less relevant attributes in the original dataset are not listed here. These redundant attributes are mainly about
 139 the operating agency of the system, types of devices used for the tag of journeys, and the information of drivers. We
 140 focus on utilizing travel pattern related data to infer the attributes of travelers, where those excluded attributes are not
 141 directly related to travel patterns of travelers.

142 Traveler attributes can include information such as age, gender, income level, residential areas. Specifically, in
 143 this study, the age group and residential areas of travelers, are studied and inferred. The age groups covered in this
 144 dataset are adults (16 - 60 years old), children (4 - 15 years old), and seniors (60 years old or older) according to
 145 the card types of Opal (<https://transportnsw.info/tickets-opal/opal/opal-card-types>). Children aged under 4 years travel
 146 free on public transport in Sydney (the data does not involve their trips). The distribution of passengers with respect
 147 to the age group is highly imbalanced (with 85.87% adults, 12.29% seniors, and only 1.84% children in the dataset).
 148 Therefore, Synthetic Minority Oversampling Technique (SMOTE) algorithm (Chawla et al., 2002) is utilized as the
 149 data augmentation method to up-sample and generate synthetic data for classes with fewer samples. SMOTE does not
 150 simply copy the types with a few samples. It helps avoid/limit over-fitting of the classifier to a certain extent. The main
 151 steps of SMOTE are summarized in the following:

152 (i) For each sample x in class C with fewer samples, the Euclidean distances between x and other samples in C are
 153 calculated to obtain its k nearest neighbors.

154 (ii) According to the sample imbalance ratio, the sampling magnification is determined. For each minority sample
 155 x , several samples are randomly selected from their k nearest neighbors x_n .

156 (iii) For x_n , a new sample x_{new} is constructed with the original sample according to the formula:

$$157 \quad x_{new} = x + rand(0, 1) \times |x - x_n|.$$

158 For the inference of residential areas, we do not have true information of the residential areas of each passenger.
 159 Instead, we assume that the administrative area (defined by postcodes in Sydney) contains the station associated with

160 the first trip of the passenger in a day is the residential area of the passenger. For those with different first-stations on
161 different days, we take the station with the most occurrences as the passenger’s residential area. In particular, in order
162 to avoid utilizing the departure place information that may directly reflect the passenger’ residential areas, we exclude
163 latitude and longitude values of first-stations (or assumed origins) for inference. That is to say, the information of the
164 assumed origins is not directly utilized in the inference framework (for residential area inference). The inference is still
165 based on the observed travel trajectories of passengers. In the experiments, we excluded administrative areas with a
166 relatively small number of residents (i.e., less than 3000 passengers associated with the area in the dataset) and finally
167 used 11 areas including 49.1 thousand passengers for residential areas inference.

168 **PoI Dataset** is also collected under the consideration that PoI (Point of Interest) information is closely related to
169 a region’s functionalities (Bai et al., 2019c), as well as travel patterns. One may infer the trip purposes of passengers,
170 along with their attributes, through analyzing the PoI information of frequently visited places. In this study, we collect
171 the PoI data of six different categories (shopping mall, church, school, hospital, club, and gym) in Sydney from Google
172 Maps Platform (<https://developers.google.com/maps/documentation/>). Each PoI item contains the name, category, and
173 location (latitude and longitude). We pre-process this dataset to map PoI data onto related transit stops (within a 500-
174 meter radius) with the help of latitude and longitude information.

175 3.2. Feature Extraction

176 As mentioned in Section 1, travel patterns of each passenger follow a certain regular distribution in spatial and
177 temporal dimensions. Thus, we extract and analyze spatial and temporal features from the original data to provide
178 intuitive understanding of the mobility patterns of passengers. In the process of feature extraction, we focus on the
179 five most frequent routes that each passenger took as representatives. Note that trip records for the five most frequent
180 routes contribute to 74.33% of the total number of trips. If we increase the scope to six most frequent routes for each
181 passenger, this percentage increases by 2.91% (from 74.33% to 77.24%), which means that the trips associated with
182 the sixth route of each passenger rarely occur (only 2.91%) and trips associated with other less used routes rarely occur
183 as well (even less than 2.91% for each route). The routes rarely used by travelers are negligible since they can hardly
184 be linked to travelers’ recurrent behaviors. For each of the five routes, we extract five features, i.e., departure/arrival
185 times, number of trips (or travel frequency), travel location, travel distance, and PoI of the destination:

- 186 • **Departure/Arrival Times:** We divide a day into six time intervals, i.e., morning rush hour (6 am to 10 am),
187 afternoon rush hour (2 pm to 4 pm, 4 pm to 8 pm), and non-busy hour (0 am to 6 am, 10 am to 2 pm, 8 pm to 10
188 am). We identify time intervals that departure/arrival times of passengers lie in.
- 189 • **Number of trips or Travel Frequency:** We count both the number of trips for the five selected (frequently used)
190 routes and the total number of trips per passenger (all routes included). The number of trips or travel frequency
191 may differ significantly over different groups.
- 192 • **Travel Location:** The latitude and longitude values of origins and destinations for the five selected routes are
193 listed to explore whether the passenger only visits a few regular destinations and whether the passenger with
194 same attributes prefer to visit nearby destinations or not.
- 195 • **Travel Distance:** Euclidean distance of each trip is calculated based on origin and destination locations. These
196 can facilitate the analysis in the sense that different groups may have diverse preferences for public transport and
197 alternative modes when the distances vary. We did not use the distance defined by the bus/train/light rail/ferry
198 lines since they are governed by the services provided.
- 199 • **PoI of Destination:** PoI information includes six categories of the destinations (shopping mall, church, school,
200 hospital, club, and gym) for each route. Destination information is often related to trip purpose, which can
201 partially reflect passenger attributes.

202 Besides, in order to fully use the travel destinations information and identify the possible relations among all transit
203 stops for travel pattern analysis and personal attributes inference, we have included all transit stop names and counted
204 travel frequency of passengers to each place to form a sparse matrix. This transit stop sparse matrix will be utilized in
205 the second component of the proposed deep learning model (refer to Section 4.2 for more details).

206 This study infers individual attributes based on the observed features of mobility patterns summarized in the above
207 rather than raw data, i.e., we extract the above features from raw data for further inference. The proposed approach is

208 based on the reasoning that the extracted features are indicators of mobility patterns and individual attributes are related
 209 to regular mobility patterns. As will be seen in the feature analysis in Section 3.3, we use the age group as an example
 210 and show that different groups indeed have distinguished travel patterns. This means that the features/patterns we
 211 extracted can help to identify the groups with different attributes. Also, different users can have very different behavior
 212 patterns, which leads to generating a sparse dataset (e.g., each person only visits a very small set of places). The
 213 repetitive patterns of the raw data for each passenger are indeed redundant. Moreover, as will be shown in Section 5,
 214 considering the five most frequent routes and the extracted features for each passenger is sufficient to produce a high-
 215 quality inference, which illustrates the great potential of linking the selected features of mobility patterns to individual
 216 attributes.

217 3.3. Behaviour Features Analysis

218 We now analyze spatial and temporal mobility characteristics against age group and their structural correlations
 219 based on the extracted features introduced in Section 3.2, which exhibit distinguished patterns for different groups.

220 3.3.1. Temporal Behaviour

221 Temporal Behaviour analysis in this subsection consists of three parts, i.e., the average numbers of trips in a week
 222 and a day, respectively, and the number of transit stops visited for three age groups.

223 Figure 1 shows the travel frequency (number of trips) of passengers on weekends and weekdays, where the x-axis
 224 is the average number of trips (per day) and the y-axis is the proportion of passengers. If a passenger does not travel on
 225 a certain day, he or she has zero trip on this day. The average number of trips for most people is less than 'three' times
 226 a day (for both weekends and weekdays). Specifically, on the weekend, the peak of adults exist at 'one' time a day
 227 while seniors have a larger peak value, 'two' times. Children do not have any obvious peak with a relatively uniform
 228 distribution from '0.5' to '1.5' times. Moreover, the average number of trips in a day for children is higher than adults
 229 since they are not able to drive. On weekdays, the number of trips with the highest proportion increases with age since
 230 children and adults need to study or work and they are less likely to visit other places.

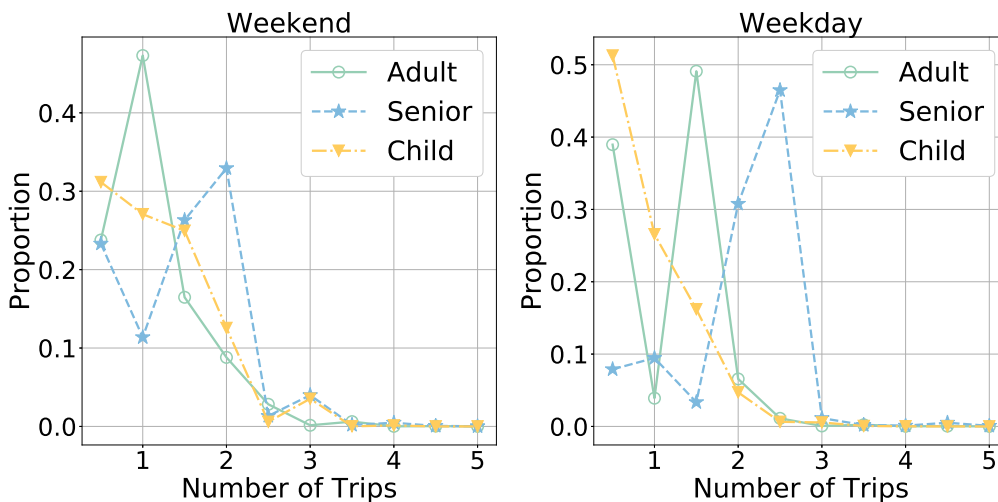
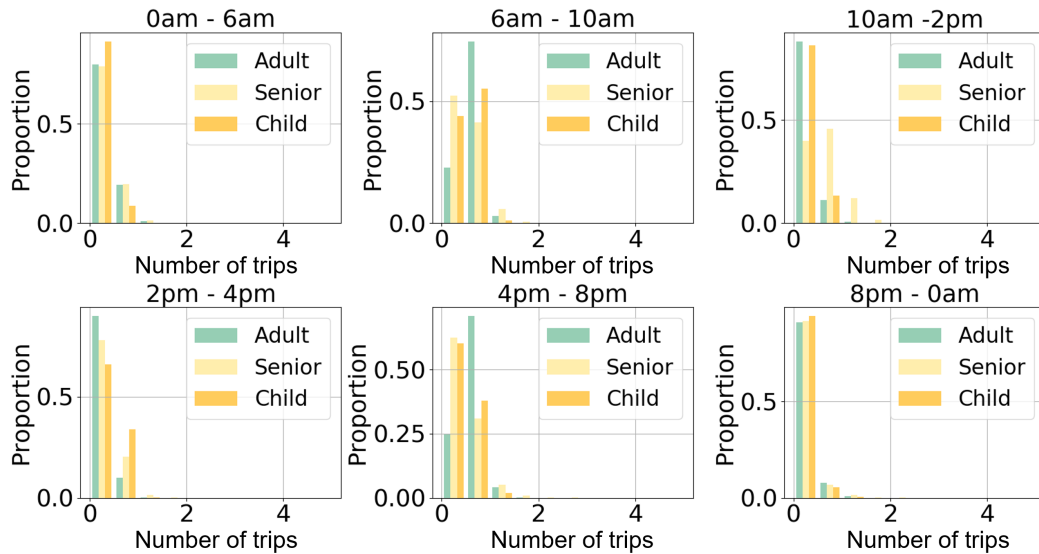
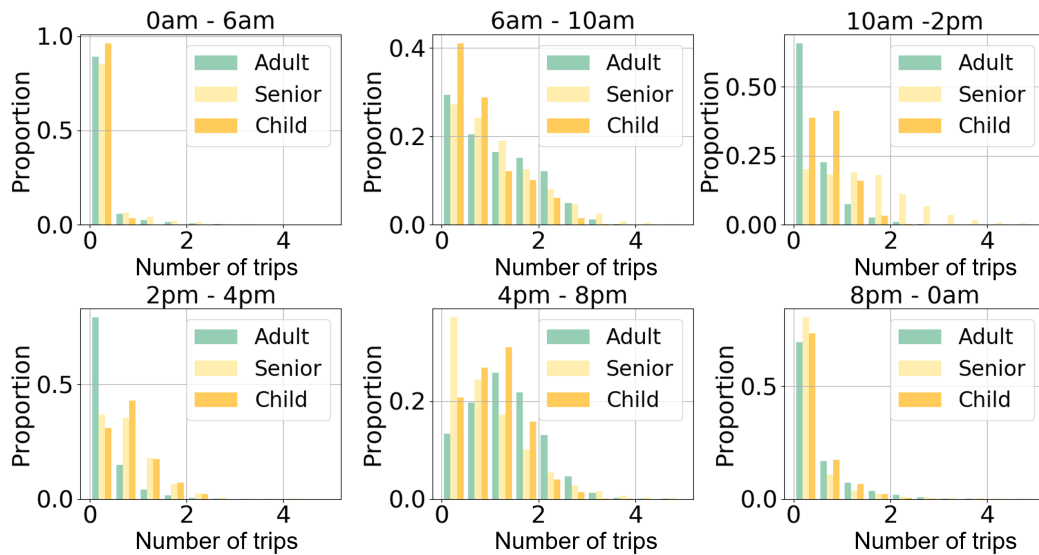


Figure 1: Numbers of trips in a day for different age groups

231 Figure 2 shows the probability distributions of the average numbers of trips in six intervals of a day for different
 232 age groups (x-axis is the number of trips while y-axis is the proportion of the passengers). The six time intervals are
 233 listed in Section 3.2, i.e., Interval 0 (0 am - 6 am), Interval 1 (6 am - 10 am), Interval 2 (10 am - 2 pm), Interval 3 (2
 234 pm - 4 pm), Interval 4 (4 pm - 8 pm), and Interval 5 (8 pm - 0 am). On weekdays, from 6 am to 10 am, the proportion
 235 of adults and children who travel is larger than the elderly since they have to travel to work or study. From 10 am to 2
 236 pm, the proportion of the elderly taking public transport is more than others. This is because, adults and kids usually
 237 stay in or around school or workplace for lunch while seniors may choose to visit other places or have lunch outside.
 238 The proportion of adults with trips between 4 pm and 8 pm is higher than the other two groups. This is likely since



(a) Weekday



(b) Weekend

Figure 2: The distribution of travel frequency in a day

239 their office hour ends in this period while the school closure time is between 2 pm and 4 pm. On the weekend, the
 240 elderly have more trips than the other two groups.

241 Figure 3 further shows the average numbers of trips over the clock time in a day for the three different age groups,
 242 which are consistent with the “Circadian Rhythm” of human activities (Liu et al., 2020). During weekdays, peaks
 243 occur around 8 am and 5 pm for adults while peaks occur around 8 am and 3 pm for children since they have to work or
 244 attend class around 8 am and the students’ school closure time is around 3 pm, which is earlier than the work closure
 245 time for adults. The departure/arrival times are more flexible for seniors since most of them are not governed by study
 246 or work schedules. Therefore, there is no sharp peak or trough within the distribution.

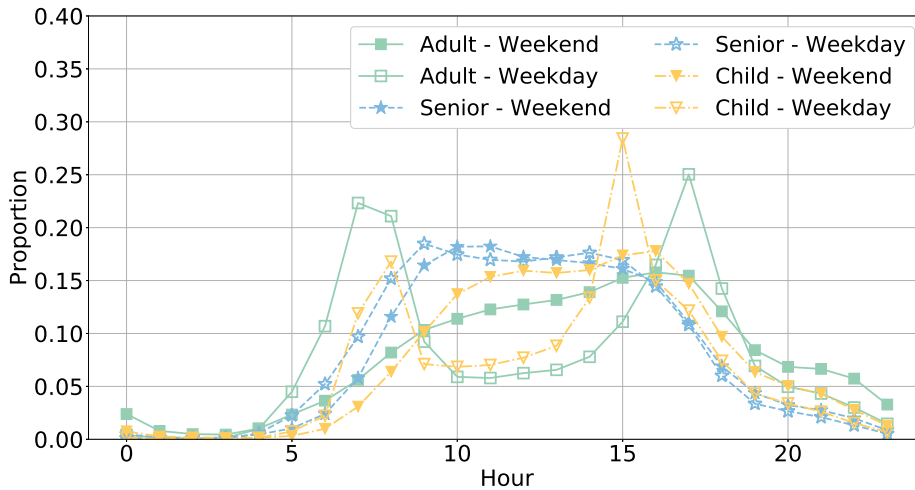


Figure 3: Number of trips in a day

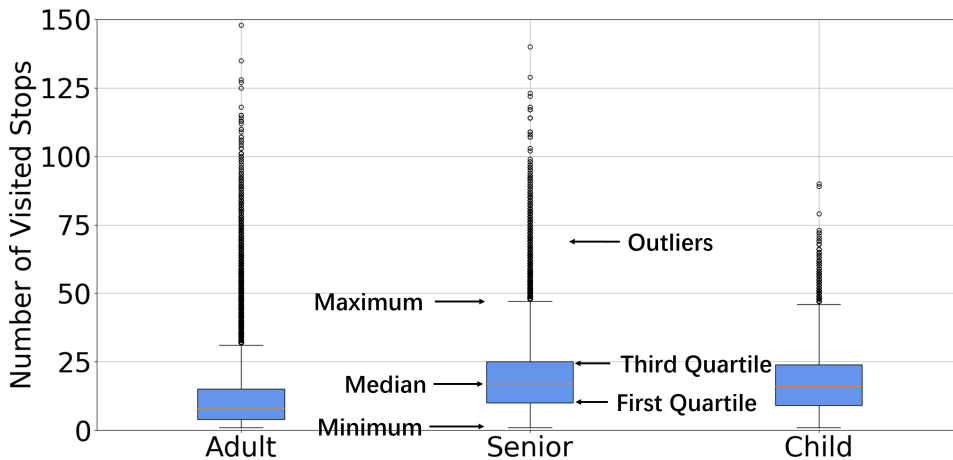


Figure 4: Number of visited stops

247 Furthermore, Figure 4 shows the distribution for the number of visited transit stops (as destinations). It indicates
 248 that the average number of stops that adults visited is 10.73, which is relatively small. The average values of seniors
 249 and children are 18.94 and 17.40, respectively. Then, for each passenger, let α_1 be the total number of trips associated
 250 with the selected five regular routes, and let α_2 be the total number of trips (includes all routes). We can define the
 251 ratio $r = \alpha_1/\alpha_2$, which is plotted in Figure 5 for each age group. For adults, as r increases, the proportion increases.
 252 Differently, values of r for the other two groups reach a peak at around 50% and then decrease. These results imply that
 253 adults travel more frequently on their regular routes than the other two groups. Also, we can calculate the proportion
 254 of passengers with a value of r greater than 90%, i.e., 23.74% for adults, 8.16% for the elderly and 6.50% for children.
 255 This is consistent with the previous results as adults need to work on weekdays and they repeatedly use a few routes to
 256 visit certain places. Also, many of the adults may choose to drive by themselves to destinations for leisure time (not
 257 regular places), which cannot be observed in the smart card data. Differently, some of the elderly and all children have
 258 to take public transport when they visit different places on their own.

259 3.3.2. Spatial Behaviour

260 In this subsection, we examine travel patterns from the spatial perspective. In particular, we consider two aspects:
 261 travel distances and PoI categories of the destinations.

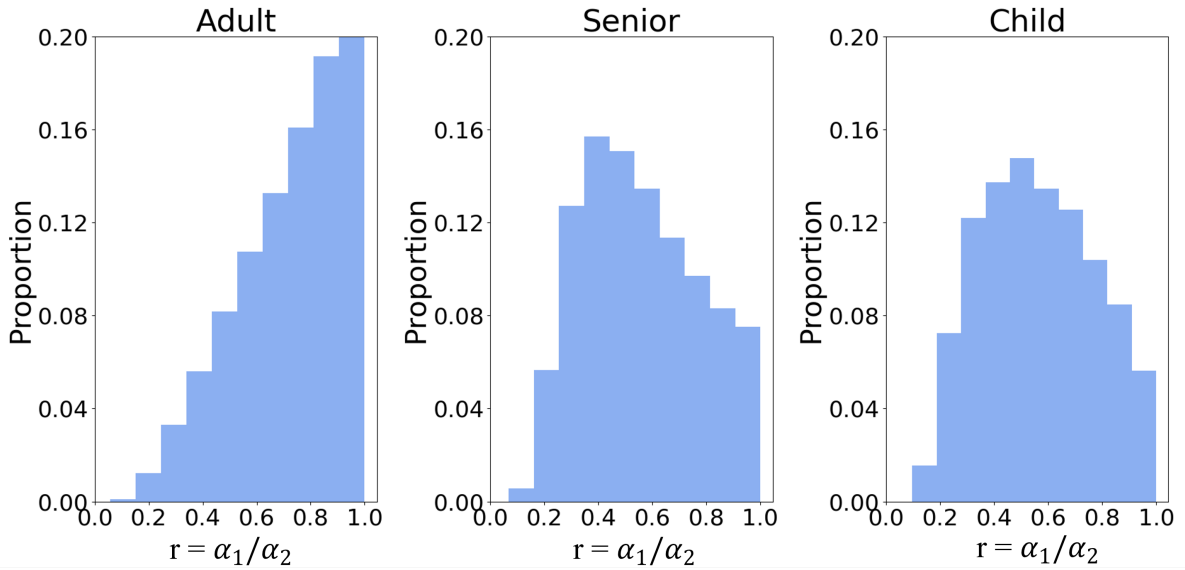


Figure 5: Distribution of the ratio of the number of trips for the five frequent routes to the total number of trips

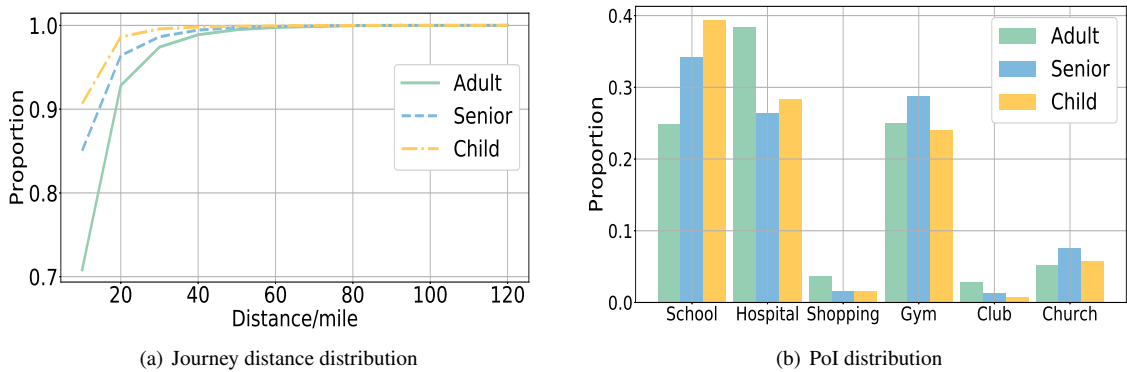


Figure 6: Spatial features based on age groups

262 The Euclidean distance between the origin and the destination of each route is calculated. The journey distance
 263 distributions of three age groups are displayed in Figure 6(a) where the x-axis represents the distances (mile) and
 264 y-axis represents the proportion of the passengers. It can be seen that 98.98% of the trips are within 40 miles. The
 265 overall distributions of travel distances for the three groups follow similar patterns. However, there also exist several
 266 differences. For example, the percentage of travel distances within 10 miles for children is 90.66%, which is the largest
 267 among the three groups while that of adults is only 70.82%. This is likely since in most cases children will not travel
 268 to places that are very far away from home on their own and their schools are usually close to home locations.

269 We further examine the distributions of the six categories collected from PoI of the destinations, which are related
 270 to possible trip purposes, and thus related to individual attributes, such as age group information. We indeed can
 271 visualize clear patterns for different age groups from Figure 6(b). For example, 30.93% of the places where children
 272 visit most often are schools and the ratio is higher than that of adults and the elderly. Adults hold 0.19% to visit clubs
 273 while the other two groups hold almost zero. And the elderly have the highest probability to visit churches, 6.82%.

274 **3.3.3. Structural Spatio-Temporal Dependencies**

275 We now move to analyze spatio-temporal patterns. The following analysis consists of two major parts, i.e., PoI
 276 categories distribution of destinations based on arrival time, and the correlation between travel destination and time.

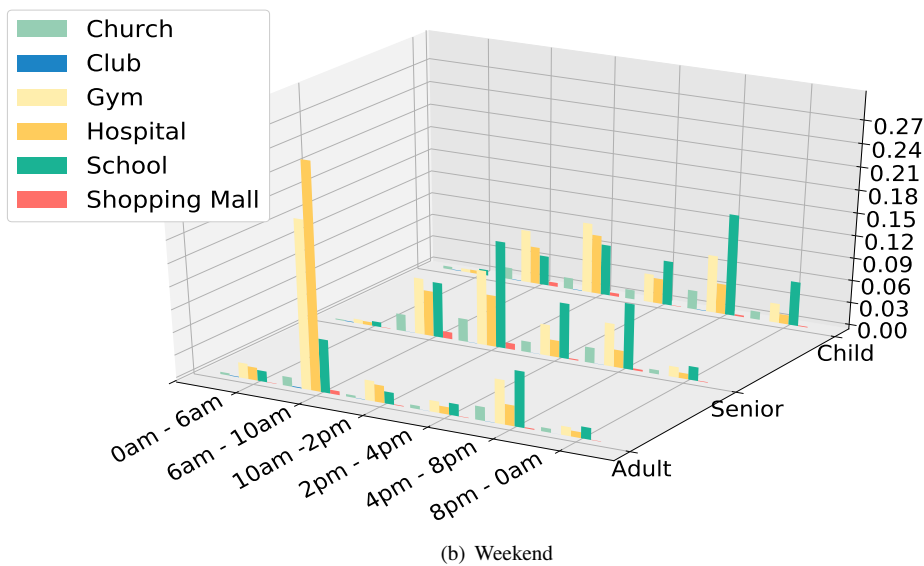
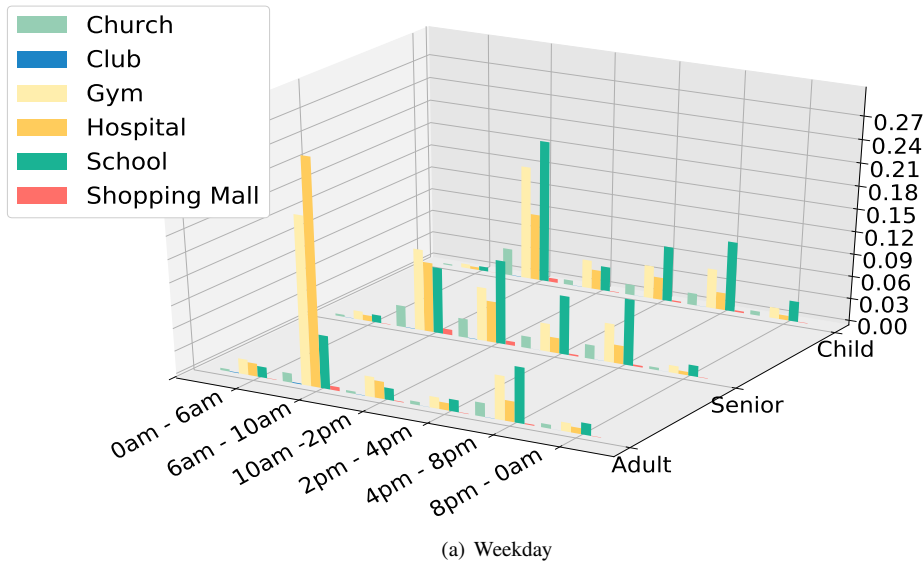
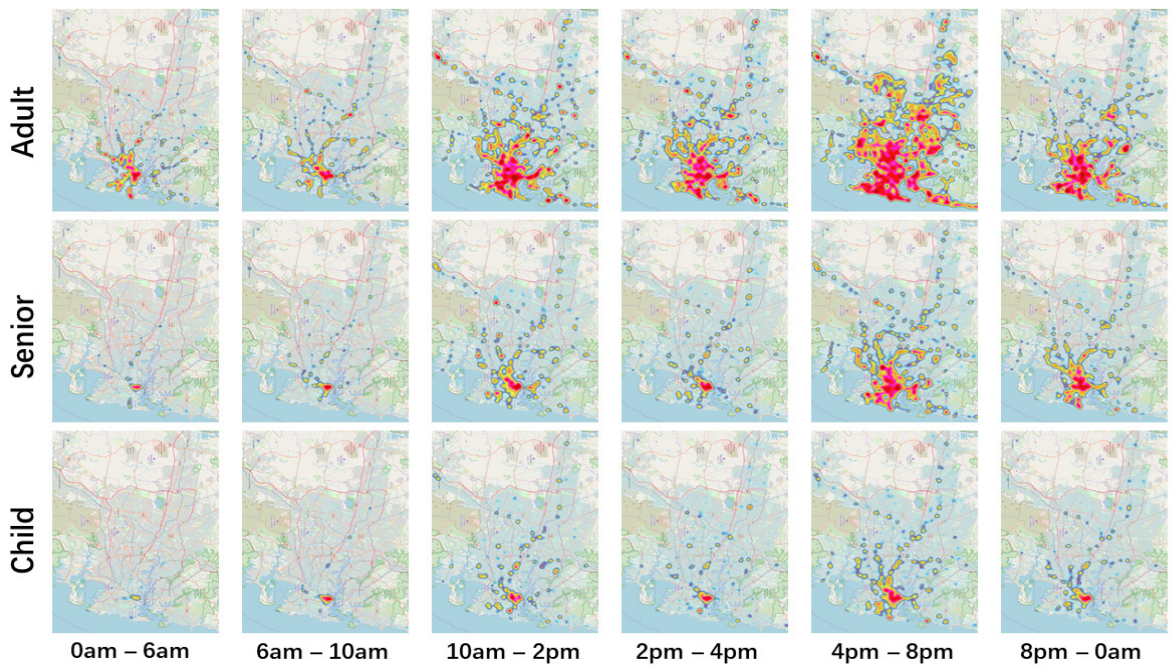


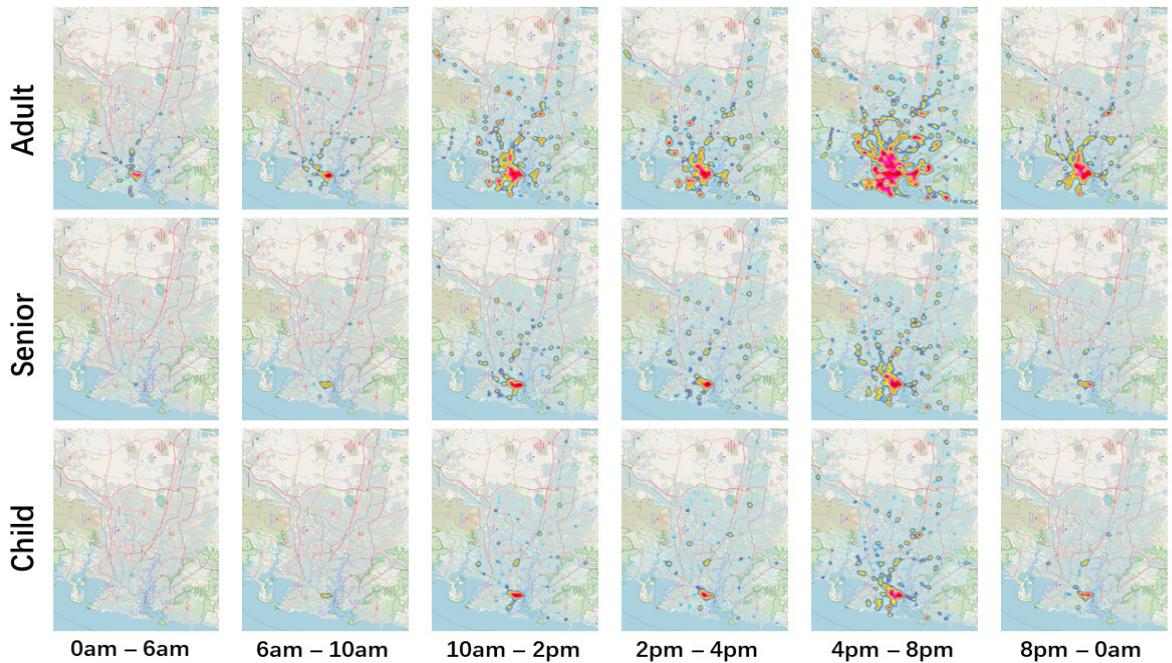
Figure 7: Relationship between PoI and temporal distribution

277 Figure 7 displays the proportions of trips related to PoI (of the destinations) in the two-dimensional domain of age
 278 group and arrival time. As mentioned earlier, there are six categories of PoI, which are also labeled in Figure 7. And
 279 z-axis represents the proportion of passengers going to a certain type of place to the total number of people in that age
 280 group. The PoI that a trip is related is only an estimate of one's destination (i.e., PoI is very close to the destination bus
 281 stop). Therefore, the six categories PoIs we choose do not necessarily mean that people were traveling to these places.
 282 They may indeed travel to other places near to those six types of PoI.

283 As can be observed from Figure 7, adults visit clubs during 0 am - 6 am while kids and seniors do not have this
 284 travel pattern. From 6 am to 10 am on weekdays, children have to attend school while the other two groups hold less
 285 probability to school. From 10 am to 2 pm, the elderly more likely to visit churches and the ratio is higher than the



(a) Weekday



(b) Weekend

Figure 8: The geographical distribution of destinations for passengers

286 other two groups. Furthermore, during 6 am - 10 am, on weekdays and weekends, the proportion of adults to go to
 287 the hospital is relatively high. This is explained below. First, there are a large number of hospitals distributed over
 288 suburbs in Sydney and they often cover large areas. And many bus stops are designed close to hospitals. Second, as
 289 we mentioned in Section 3.2, we only choose six types of PoI (there are some more types from Google API) in this study.
 290 Therefore, people who go to other places different than the six types of PoI may be regarded as going to hospitals.

291 This is to say, many going to work in the morning are counted as going to the hospitals. Thus, the proportion of adults
292 going to the hospital is high.

293 We also visualize the spatial distribution of destinations for the three age groups over the six intervals on weekends
294 and weekdays in Figure 8. The colors indicate the frequency of this location as the destination, i.e., approximately
295 blue → yellow → red → dark red indicating a growing density/frequency of passengers. On weekdays, for seniors and
296 children, the distribution difference between 0 am - 6 am and 6 am - 10 am is larger than adults. And the geographical
297 distribution from 10 am to 2 pm is larger than that from 2 pm to 4 pm while the adults have a similar distribution in
298 these two intervals. On the weekend, the geographic range of children and elderly destinations is more concentrated
299 in the city center while adults have a much broader range. Also, the overall geographical distribution of adults on
300 weekdays is much wider than that on the weekend since they prefer to stay at home or drive to visit other places.

301 It is worth mentioning that depending on the age of the children, they may be codependent with adults in terms
302 of their travel patterns. However, the distinguished features of adults and children in the above analysis indicate that
303 they might be dependent to a very limited extent. The first reason is that the proportion of families with children is
304 only 26.7% in Sydney, which means that a significant number of adults do not have children, and their activity patterns
305 are almost irrelevant to those of children. Second, for those with children, the activity dependency is further governed
306 by the ages of children. In Sydney, children aged 0-14 years made up 4.5% of the population, where those aged 0-4
307 years made up 2.7%, and those aged 5-9 years made up 1.1%, and those aged 10-14 years made up 0.8%.¹ Those
308 aged 0-4 years are not in the dataset (they can travel free in Sydney), those aged 10-14 years already have relatively
309 independent within-city daily trips. This is to say, there is only around 1.1% of the population that may exhibit clear
310 activity dependency on others (adults). Furthermore, people with similar travel patterns may work or live in similar
311 places, which does not necessarily mean that they are a family. They may be just “familiar strangers” who appear on
312 the same bus (Sun et al., 2013). Thus, our method which will be introduced in Section 4 yields very effective inference
313 results by treating adults and children separately since they exhibit different travel patterns.

314 4. Methodology

315 We now present the proposed modeling framework of the hybrid neural network incorporating spatial-temporal
316 dependencies to infer the traveler attributes. The overall architecture of the proposed model is depicted in Figure 9,
317 where we have two parallel sub-networks: (i) a Product-based Spatial and Temporal Module (**PSTM**) with an inner-
318 product operation (**Inner-PNN**) for capturing spatial-temporal dependencies from the extracted features introduced in
319 Section 3.2; (ii) an Auto-Encoder-based Compression Module (**AECM**) for compressing the transit stop sparse matrix
320 shown in Section 3.2. The concatenation of them will be sent for attribute inference (i.e., classifying the age group and
321 residential areas). By doing so, spatial-temporal features of mobility patterns can be well accounted for and the model
322 is able to distinguish passengers of different groups (with different attributes).

323 It is noteworthy that for the two sub-networks introduced in the above, we may replace **Inner-PNN** and **AECM**
324 by other modules, e.g., the Outer-Product-based Module (**Outer-PNN**) and the fully connected layers (**FCLs**). We
325 have tested different combinations for the two sub-networks, which will be presented in Section 5.3. We only present
326 the formulations of **Inner-PNN** and **AECM** in the following since this combination yields the best performance for
327 individual attribute inference.

328 4.1. Inner-Product-based Spatial-Temporal Module

329 We now introduce the Product-based Spatial-Temporal Module with inner-product operation (**Inner-PNN**), a very
330 capable way to capture and investigate the pairwise relevance among the extracted spatial and temporal features for
331 better inference.

332 As can be seen in Section 3.3, different mobility features exhibit dependencies over each other. Following He et al.
333 (2018), we propose an inner-product module in our network to investigate the relations among all fields (each field
334 corresponds to one type of feature) in pairwise since the inner-product module is more powerful than pure concatenation
335 or addition operation, which are not included any correlation among features. Moreover, the product module combined
336 with the fully connected layer is able to inspect high-order feature reactions. Also, the deep neural network is capable
337 to capture non-linear latent patterns among spatial and temporal features. It breaks the limit that traditional models
338 may only identify the shallow relationships among data.

¹https://quickstats.censusdata.abs.gov.au/census_services/getproduct/census/2016/quickstat/SSC13715?opendocument

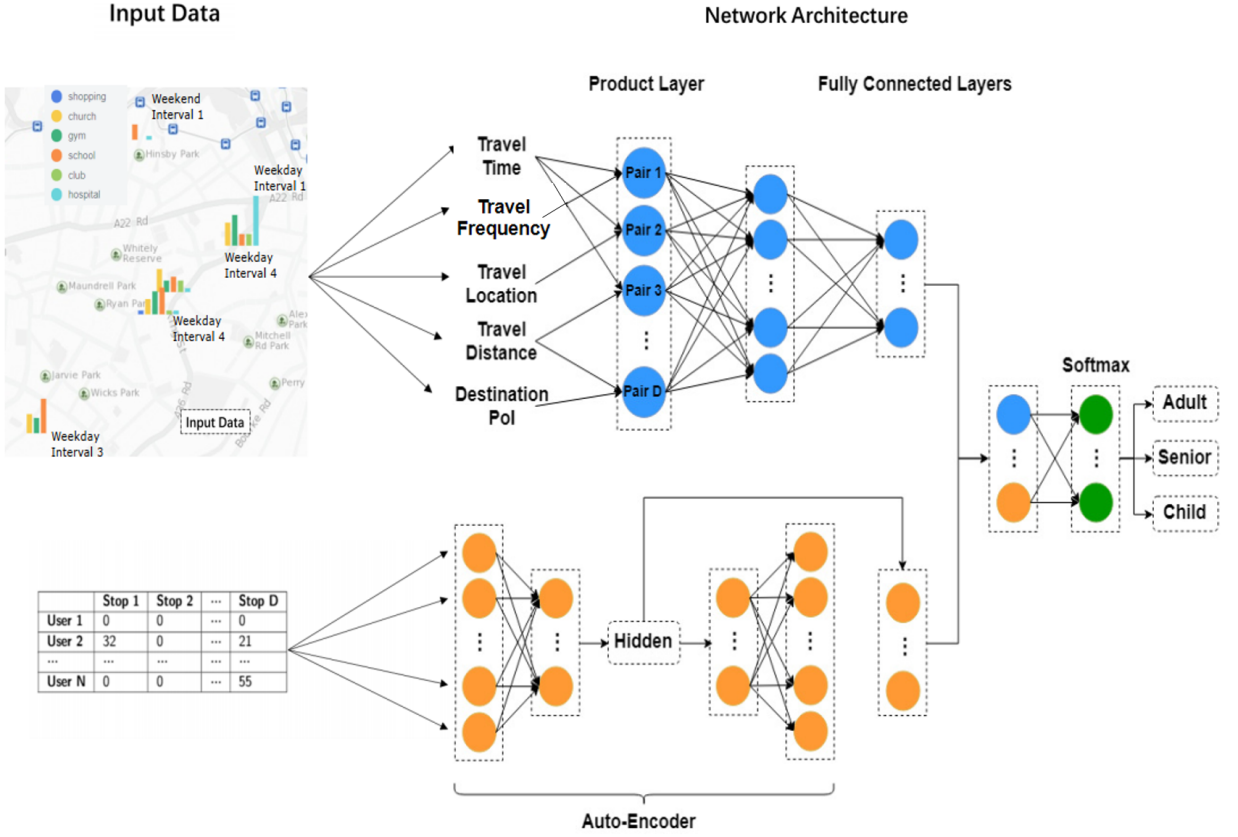


Figure 9: Hybrid Spatial-Temporal Correlation Neural Network Architecture

339 The feature matrix $P_1 \in \mathbb{R}^{N \times D_1}$ is embedded into five fields, where N and D_1 denote the number of test samples
 340 and dimension of all features, respectively. i^{th} field is defined as $v_i \in \mathbb{R}^{N \times \delta_i}$, where δ_i is the length of i^{th} field.
 341 Correspondingly, $\mathbf{V} = (v_1, v_2, \dots, v_i, \dots, v_I)$ is defined as the output of embedding layer that will be sent into inner-
 342 product layer to find the pairwise connection, where I is the number of fields.

Then, the definition of inner-product between two vectors is $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T \mathbf{b}$, where T means the transpose operation. In
 the geometric sense, we can see the proximity of two vectors in the direction from the inner-product values. Therefore,
 we expand the inner-product to two matrices (e.g., matrix A and matrix B) to quantify the relation between them. The
 inner-product in the layer is defined as:

$$A \cdot B = \sum_{m,n} A_{m,n} B_{m,n} \quad (1)$$

343 which is the sum of the product of the elements at the corresponding positions from two matrices. Thus, the inner-
 344 product of two fields is $\langle v_i, v_j \rangle = W_p^i v_i \cdot W_p^j v_j$, where $W_p^i \in \mathbb{R}^{\Gamma \times \delta_i}$ and $W_p^j \in \mathbb{R}^{\Gamma \times \delta_j}$. Γ is the output dimension of
 345 each field. W_p^i is defined as the weight matrix of i^{th} field in the product layer. Therefore, the output of product layer
 346 can be represented as $L_p = (l_1, l_2, \dots, l_\Psi, \dots, l_\Psi)$, where Ψ is the number of pairs (two different fields form a pair),
 347 and l_Ψ is calculated as $l_\Psi = \langle v_i, v_j \rangle$.

L_p is then fed into a fully connected layer and produce the output $L_1 \in \mathbb{R}^{D_1}$:

$$L_1 = \text{relu}(W_1 L_1 + b_1) \quad (2)$$

where W_1 denotes the weight matrices while b_1 represents the bias vector. And relu is the Rectified linear unit activa-

tion function, i.e.,

$$relu(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases} \quad (3)$$

4.2. Auto-Encoder-based Compression Module

As discussed earlier, P_2 is sparse with redundant information, which need to be compressed for further inference. Thus, this section discusses how to take advantage of Auto-Encoder-based Compression Module (**AECM**) to fully utilize the travel destinations information from the transit stop matrix P_2 and reduce the loss of information during the compression process. The learned embedding vectors by **AECM** will be further used to infer the traveler attributes. Note that using the fully connected layers (**FCLs**) may also achieve the same compression objective, however, more information may be lost during the compression process. As discussed in Hinton and Salakhutdinov (2006), Auto-Encoder (**AE**) for dimensional reduction is able to retain most of the original data information which can be more powerful to extract information from P_2 .

As shown in Figure 9, **AECM** is composed of an Auto-Encoder and a fully connected layer. In detail, P_2 is fed into an Auto-Encoder at first to fuse features from different domains together while keeping most of the useful information. The encoding and decoding processes are employed with two-layer fully-connected networks and the transformation of the encoder and decoder can be described as follows:

$$\begin{aligned} H_i(r_i) &= encoder(P_2) \\ \hat{P}_2 &= decoder(H_i(r_i)) \end{aligned} \quad (4)$$

where $encoder(\cdot)$ and $decoder(\cdot)$ represent the transformations of the encoder part and decoder part, respectively. $H_i(r_i)$ is the hidden representation of P_2 , and \hat{P}_2 is the output. The encoder part is used to identify the compressed representation of the input data while the decoder part is used for the representation of the reconstruction that is similar to the original input. The cost function of Auto-Encoder is MSE (mean squared error) of $P_2 - \hat{P}_2$:

$$MSE = \frac{1}{n} \sum_{i=1}^n (P_{2i} - \hat{P}_{2i})^2 \quad (5)$$

and thus we drive \hat{P}_2 to be close to P_2 as much as possible. The hidden representation $H_i(r_i)$ is the most valuable part of Auto-Encoder, which is then fed into one fully connected layer for concatenation. And the result of this module is L_2 . The activation functions of all fully connected layers are again $relu$.

4.3. Combination and Classification

To fuse the spatio-temporal relevance information extracted in Section 4.1 and transit stops information derived in Section 4.2 for traveler attributes inference, we concatenate L_1 with L_2 together to form L_3 . At last, L_3 is fed into one fully connected layer activated by $relu$ to produce the final classification/inference result \hat{y} . The objective function of the proposed network consists of two parts: the constraint loss $Loss_1$ of Auto-Encoder in **AECM** and the loss $Loss_2$ of final classification/inference, which are given as follows:

$$\begin{aligned} Loss_1 &= MSE(P_2, \hat{P}_2) \\ Loss_2 &= Softmax_cross_entropy(y, \hat{y}) \end{aligned} \quad (6)$$

where y is the true label of the input samples, MSE is the mean square error, and $Softmax_cross_entropy$ is the cross entropy loss for softmax function. The softmax function is to map a K -dimensional arbitrary real vector into another K -dimensional real vector, where each element in the vector has a value between zero and one. For example, the probability of z belongs to class j is:

$$softmax(z_j) = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \quad (7)$$

The cross entropy function is used to measure the loss between the predicted standard probability distribution from the softmax function and the correct category label, which is represented as:

$$Loss_2 = -[y \log \hat{y} + (1 - y) \log(1 - \hat{y})] \quad (8)$$

361 Then, the overall loss can be calculated as $Loss(\theta) = (1 - \lambda) \times Loss_1 + \lambda \times Loss_2$, where θ represents all learnable
 362 parameters in the network. The parameter λ is used to balance the two types of loss functions and the suitable values
 363 will be tested in Section 5.4. It is obtained/solved via back-propagation and Gradient Descent optimizer. The detailed
 364 training steps of the proposed model is illustrated in Algorithm 1.

Algorithm 1 Spatio-Temporal Neural Network Model Training

Input:

Spatio-Temporal Feature Matrix P_1
 Transit Stop Sparse Matrix P_2
 Learning Rate
 λ in the loss function

Output:

Spatio-Temporal Correlation Model with learned parameters

1: Initialize parameters: weights and biases
 2: **repeat**
 3: for all $(x_k, y_k) \in N$ do
 4: P_1 is embedded and sent into product layer and two fully connected layers to get the output L_1
 5: P_2 is sent into an Auto-Encoder to get the output L_2
 6: Send the Concatenation of L_1 and L_2 into the fully connected layer
 7: Calculate the output \hat{y}_k based on current samples
 8: Estimate the parameters by the loss function $Loss(\theta) = (1 - \lambda) \times Loss_1 + \lambda \times Loss_2$
 9: **end for**
 10: **Until** convergence criterion met
 11: **end procedure**

365 **5. Experiments**

366 In this section, we first introduce the experiment settings and evaluation metrics. Then, we present the experiment
 367 results from three perspectives: overall comparison, network architecture analysis, and ablation study. In particular,
 368 the proposed model is compared with ten existing strategies, i.e., **LDA, QDA, SVM, Ada, DT, XGBoost, MLP, PPC,**
 369 **C2AE and DeepSD**. Also, we test different models for the two sub-networks defined in Section 4 (network architecture
 370 analysis). Moreover, we conduct a number of ablation studies in relation to the model, where in the ablation study we
 371 remove or change some features/components of the model and examine how doing so can affect the model performance
 372 (in terms of inference accuracy). It consists of two parts: feature selection by removing one feature in each experiment;
 373 and parameter tuning on learning rate, λ in the loss function, and the ratio of training data to testing data.

374 **5.1. Experiments Setting**

To test the performance of the proposed traveler attributes inference model, we choose 70% of users in each group
 for training and validation, and use the remaining users for testing. The extracted features are normalized by Euclidean
 norm and shuffled before feeding into the model. The formula of Euclidean norm is: $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$. The output
 dimension Γ of each field is set to 10. We also set the learning rate to 0.1 and batch size to 128. The λ in the loss
 function is set to 0.5. In the confusion matrix, True Positive (TP) means that positive class is predicted as positive class,
 True Negative (TN) means that negative class is predicted as negative class, False Positive (FP) means that negative
 class is predicted as positive class, and False Negative (FN) means that positive class is predicted as negative class.
 Three evaluation metrics used in our work, i.e., accuracy, precision, and recall, are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}; \quad Precision = \frac{TP}{TP + FP}; \quad Recall = \frac{TP}{TP + FN}. \quad (9)$$

375 **5.2. Overall Comparison**

376 We evaluate the effectiveness of the proposed model against several widely used classification algorithms and three
 377 state-of-the-art deep-learning-based models, which are briefly summarized in the following:

- 378 • **Linear Discriminant Analysis (LDA)**: One or more explanatory variables can be used in **LDA** to represent
379 a binomial result. It measures the relationship between a class-dependent variable and one or more indepen-
380 dent variables by using a logic function to estimate the probability, the latter obeying the cumulative logical
381 distribution.
- 382 • **Quadratic Discriminant Analysis (QDA)**: Different from the **LDA** in the above, **QDA** assumes that the clas-
383 sification boundary is quadratic, which provide more accurate prediction than linear models.
- 384 • **Support Vector Machine (SVM)**: Traditional **SVM** algorithm is a binary classification algorithm, which gives
385 two types of points in N-dimensional coordinates. In particular, **SVM** generates an (N-1)-dimensional hyper-
386 plane to divide the points into two groups. Improved **SVM** can be used for multi-label classification, Jahangiri
387 and Rakha (2015) found that **SVM** produced the best performance to classify travel mode.
- 388 • **Adaptive Boosting (Ada)**: The core idea of **Ada** is to train different classifiers (weak classifiers) for the same
389 training set, and then combine these weak classifiers to form a stronger final classifier (strong classifier). For
390 age group inference, the number of estimators is 50 while the learning rate is 0.3. For residential area inference,
391 the number of estimators is 900 while the learning rate is 1.5.
- 392 • **Decision Tree (DT)**: The decision process using **DT** is to start from the root node, test the corresponding feature
393 attributes in the item to be classified, and select the output branch according to its value until the leaf node is
394 reached. The category stored by the leaf node is used as the decision result. The maximum depth of the tree is
395 12 for two inference problems.
- 396 • **eXtreme Gradient Boosting (XGBoost)**: **XGBoost** was proposed by Chen and Guestrin (2016) for classifica-
397 tion. It originated from the Boosting integrated learning method and incorporates the advantages of Bagging
398 integrated learning methods in the evolution process. The custom loss function through the Gradient Boosting
399 framework improves the ability of the algorithm to solve common problems and introduces more controllable
400 parameters for scene optimization. For two reference problems, the number of estimator is 50, the learning rate
401 is 0.01, and the maximum depth is set to 5.
- 402 • **Multilayer Perceptron (MLP)**: **MLP** is an improvement of perceptrons, which overcomes the weakness that
403 perceptrons cannot identify non-linear data. It has fully connected layers and non-linear activation functions. A
404 supervised learning method called back-propagation algorithm is often used to train **MLP**. Note that the way to
405 solve the optimization problem in our experiment is the gradient descent method. Two hidden layers are set for
406 residential area inference and the hidden sizes are 128 and 64. Similarly, the hidden sizes of age group inference
407 are set to 50 and 20. The learning rate is 0.001.
- 408 • **Propagation Path Classification Model (PPC)**: **PPC** was built by Liu and Wu (2018) for fake news detection
409 and classification which adopted recurrent and convolutional neural networks to capture the variations of user
410 features along the propagation path. For two inference problems, the hidden size of RNN encoder is 25 while
411 the output channel of CNN is 10. The kernel size of CNN is set to (1, 56). The learning rates are 0.0008 and
412 0.001 for age group and residential area, respectively.
- 413 • **Canonical Correlated Auto-Encoder (C2AE)**: **C2AE** was proposed by Yeh et al. (2017) for multi-label classifi-
414 cation. It integrated deep Canonical Correlation Analysis and Auto-Encoder structures to learn a latent subspace
415 for better classification. Four fully connected layers are set for embedding the features and labels. For age group
416 inference, the sizes of these layers are 32, 64, 32, 16. The number of units of the output layer is 16. For residen-
417 tial area inference, the sizes of these layers are 64, 128, 64, 32. The number of units of the output layer is 32.
418 The learning rate is set to 0.0001.
- 419 • **Deep Supply-Demand Neural Network (DeepSD)**: Wang et al. (2017) proposed **DeepSD** to automatically learn
420 the patterns across different spatio-temporal attributes for the real-time car-hailing supply-demand forecasting.
421 We use the basic version of this model and change the output layer to infer the age groups and residential areas in
422 the experiments. For the two inference problems, the embedding sizes of temporal features (i.e., departure/arrival
423 time intervals in one day, the day of a week, and the day of a month) are set to 16, 8, and 16. The hidden sizes of
424 other features (i.e., total number of trips, the number of occurrences of the selected routes, PoI of destinations,

Table 2

Overall comparison between the proposed method and several existing methods

Model	Age Group							Residential Area		
	Accuracy			Recall		Precision		Accuracy	Average Recall	Average Precision
	Adult	Senior	Child	Adult	Senior	Child				
LDA	0.6072	0.6859	0.5765	0.5590	0.6787	0.5799	0.5617	0.4273	0.3845	0.3938
QDA	0.4265	0.3741	0.8486	0.0570	0.6345	0.3802	0.3190	0.5672	0.5582	0.5601
SVM	0.5151	0.7720	0.0040	0.7692	0.5955	0.6430	0.4532	0.4194	0.3516	0.4124
Ada	0.6370	0.7518	0.5474	0.5114	0.8729	0.5592	0.5269	0.5769	0.6058	0.5616
DT	0.7613	0.9262	0.7545	0.6027	0.8585	0.7006	0.7143	0.6160	0.6004	0.6325
XGBoost	0.6879	0.8312	0.5982	0.6341	0.8299	0.6112	0.6218	0.6854	0.6604	0.7473
MLP	0.7849	0.8135	0.7512	0.7867	0.8747	0.7226	0.7576	0.6917	0.6835	0.6789
PPC	0.6824	0.6571	0.6490	0.7349	0.6809	0.6649	0.6978	0.6886	0.6766	0.6828
C2AE	0.6354	0.6997	0.6799	0.5247	0.7116	0.5241	0.7074	0.6717	0.6329	0.6889
DeepSD	0.8741	0.8738	0.8210	0.9274	0.8250	0.8506	0.9498	0.6412	0.6234	0.6345
Our Model	0.9237	0.8664	0.9068	0.9989	0.9027	0.8854	0.9831	0.8516	0.8520	0.8618

425 travel distance, and travel location) are set to 64, 32, 64, 32, 64. The hidden size of all features is set to 32. The
426 learning rate is set to 0.001.

427 Table 2 summarizes the inference results of the age group and residential areas for the proposed method and
428 the aforementioned existing tools. We test different settings of hyperparameters for all the compared methods based
429 on the validation set. The setting of hyperparameters that yields the best performance is then utilized to conduct
430 testing and comparison based on the testing set. Several observations from Table 2 are discussed below. First, the
431 proposed method significantly outperforms the traditional classification strategies including **LDA**, **QDA**, **SVM**, **Ada**,
432 **DT**, and **XGBoost**. This is because the proposed model based on deep neural networks can better accommodate
433 non-linear relations among different features. Specifically, the Auto-Encoder-based Compression Module carries out
434 sparse matrix analysis and compression operation that **SVM** cannot solve. Second, the Inner-Product-based Module
435 explores the interactive patterns among spatial and temporal features, which yields higher accuracy than **MLP**. Thus,
436 the proposed model performs better than other three listed deep-learning-based models, i.e., **PPC**, **C2AE**, and **DeepSD**.
437 As for **PPC** which is based on the recurrent neural network, it is hard to capture the correlations among features since
438 RNN-based models are often used for sequential series analyzing not for capturing parallel relations. Similarly, the
439 component of **C2AE**, deep Canonical Correlation Analysis, is often used to joint feature and label embedding but
440 not to find correlations among features. Moreover, **DeepSD** is developed for car-hailing supply-demand prediction
441 which can extract spatial-temporal correlations but performs worse in attributes inference/classification. Overall, the
442 proposed approach achieves better performance than the listed approaches in inferring the age group and residential
443 areas, indicating the effectiveness of the proposed modeling framework that consists of **Inner-PNN** and **AECM** for
444 traveler attributes inference. Note that although our definition of residential areas does not necessarily represent true
445 information of residential area for passengers, under the same assumptions/conditions, our model can still yield higher
446 accuracy than other listed strategies.

447 It is observed that the inference results of children have relatively high precision and recall. This is further explained
448 below. Generally speaking, the mobility patterns of kids have a high level regularity and their travels are often more
449 limited. For instance, as can be seen from Section 3.3, most kids travel in the morning rush hour (6 am - 10 am) and
450 the afternoon rush hour (2 pm - 4 pm) on weekdays. And the five selected high-frequency travel routes for children
451 account for a high percentage of total trips. This observed regularity in relation to children indeed helps the inference.

452 5.3. Network Architecture Analysis

453 As discussed in Section 4, we propose a basic framework consisting of two sub-networks for traveler attributes
454 inference. In this subsection, utilizing age group inference as an example, we compare the performance of utilizing
455 different components in the framework including the Fully Connected Layer (**FCL**), Auto-Encoder-based Compression
456 Module (**AECM**), Inner-Product-based Module (**Inner-PNN**), and Outer-Product-based Module (**Outer-PNN**). We
457 also test each of the above four modules for inference.

458 For the combination of different modules in the two sub-networks framework, **FCL** or **AECM** is utilized to com-

Table 3
Performance under different combinations of model components

Model	Accuracy			Recall			Precision	
		Adult	Senior	Child	Adult	Senior	Child	
FCL	0.8051	0.7936	0.7354	0.8864	0.8105	0.7952	0.8086	
AECM	0.8194	0.6226	0.8518	0.9849	0.8404	0.7275	0.9046	
Inner-PNN	0.8624	0.8874	0.8304	0.8694	0.8575	0.8238	0.9087	
Outer-PNN	0.8009	0.8682	0.7254	0.8090	0.8130	0.7868	0.8009	
Outer-PNN + FCL	0.8792	0.7711	0.8774	0.9897	0.8754	0.8193	0.9441	
Inner-PNN + FCL	0.9053	0.8352	0.8896	0.9719	0.8807	0.8615	0.9719	
Outer-PNN + AECM	0.9023	0.8295	0.8890	0.9889	0.8792	0.8552	0.9723	
Inner-PNN + AECM	0.9237	0.8664	0.9068	0.9989	0.9027	0.8854	0.9831	

459 press the transit stop matrix and extract useful information for further inference. **Inner-PNN** or **Outer-PNN** inves-
 460 tigate the relevance among the extracted spatial and temporal features to infer different groups. Therefore, we have
 461 four combinations, where the results are summarized in Table 3. We discuss the four modules and their combinations
 462 in the following.

463 **FCL** or **AECM** can be used to compress the public transit stop matrix and extract useful knowledge from it (e.g.,
 464 destination information). However, the transit stop matrix does not contain much temporal information. Also, the
 465 transit stop matrix consisting of all destinations of passengers may enclose some noises (occasional travel destinations
 466 cannot be counted as a part of passenger’s regular travel patterns). Therefore, **FCL**-only or **AECM**-only yields worse
 467 performance than the combination of two sub-networks. Both **AECM** and **FCLs** are able to extract important infor-
 468 mation from the transit stop matrix. However, compared to **AECM**, more information may be lost when compressing
 469 the transit stop matrix with **FCLs**. As introduced in Section 4.2, the constraint loss function in **AE** helps drive the
 470 output from the decoder to be close to the input data as much as possible. Thus, the derived features from the encoder
 471 are more powerful and complete than that from **FCLs**, and thus may help yield a higher level of inference accuracy in
 472 many occasions.

473 **Inner-PNN** and **Outer-PNN** both help capture and utilize the relevance among extracted spatial and temporal fea-
 474 tures. However, only using **PNN**-based modules for inference may ignore some destination information of passengers
 475 since the extracted features only contain the features of the most frequent travel routes. The overall destination infor-
 476 mation is better captured by the transit stop matrix. Thus, the performance of **PNN**-based modules is worse than the
 477 complete framework with two sub-networks. Furthermore, the results in Table 3 show that **Inner-PNN** achieves better
 478 performance than **Outer-PNN**. **Outer-PNN** has been proposed by He et al. (2018) to explicitly model the pairwise
 479 correlations among different types of one-hot encoded features for the recommendation system. Although the data in
 480 our experiments consists of several types of features, it is not a one-hot encoded sparse matrix. Thus, directly adapting
 481 **Outer-PNN** for attributes inference may not obtain distinguished performance. Also, in terms of the geometric mean-
 482 ing, inner-product operation indeed judges the angle between two vectors but the outer-product does not have such a
 483 meaning.

484 The combination of **Inner-PNN** and **AECM** achieves the highest accuracy, since it takes advantage of two sub-
 485 networks, and provides a capable and balanced way to capture information in the dataset. It is worth mentioning
 486 that **Inner-PNN** performing better than this combination for the adult group indicates that the extracted features are
 487 sufficient to identify adults and the transit stop matrix may contain noises (occasional travel destinations), which may
 488 have a negative impact on the age group inference.

489 5.4. Ablation Study on Feature Selection and Parameter Tuning

490 In this section, taking age group inference as an example, we evaluate the importance of features as shown in
 491 Section 3.2 by an ablation study of removing one feature in each experiment. Then, we evaluate the performance of
 492 the proposed method under different learning rates, λ in the loss function, and the ratio of training and testing data.

493 Figure 10(a) shows the importance score of each type of feature listed in Section 3.2 in our model. As can be seen,
 494 all the features positively contribute to the classification/inference while their significance differs. To further verify
 495 that the extracted features do make a contribution in the proposed model, we remove one feature each time and conduct
 496 an independent experiment to infer the age group and the testing results are shown in Table 4. Keeping all features

Table 4
Performance under different combinations of features

Features	Accuracy			Recall			Precision		
		Adult	Senior	Child	Adult	Senior	Child		
Without Pol	0.8938	0.8076	0.8808	0.9922	0.8714	0.8453	0.9626		
Without Time	0.8930	0.8033	0.8875	0.9900	0.8728	0.8383	0.9700		
Without Distance	0.8974	0.8686	0.8440	0.9787	0.8349	0.8784	0.9785		
Without Location	0.9133	0.9135	0.8305	0.9956	0.8458	0.9185	0.9811		
Without Frequency	0.9166	0.8935	0.8630	0.9926	0.8642	0.9062	0.9791		
All Features	0.9237	0.8664	0.9068	0.9989	0.9027	0.8854	0.9831		

497 yields a higher level of accuracy than other settings. It is also noteworthy that removing location or travel frequency
 498 yields relatively close performance to the proposed model with all five features. This is partially due to that the sparse
 499 matrix training by **AECM** also includes location and travel frequency information of passengers.

500 In Figure 10(b), we display the values of loss functions $Loss_1$ (from the constraint of Auto-Encoder) and $Loss_2$
 501 (the loss of final classification in the concatenation part) when we vary λ from 0.1 to 0.9 ($\lambda = 0.5$ in the benchmark
 502 case). As can be seen, when $Loss_1$ decreases, $Loss_2$ increases. These points are expected to lie along the Pareto
 503 frontier, where we minimize the two objectives $Loss_1$ and $Loss_2$ simultaneously.

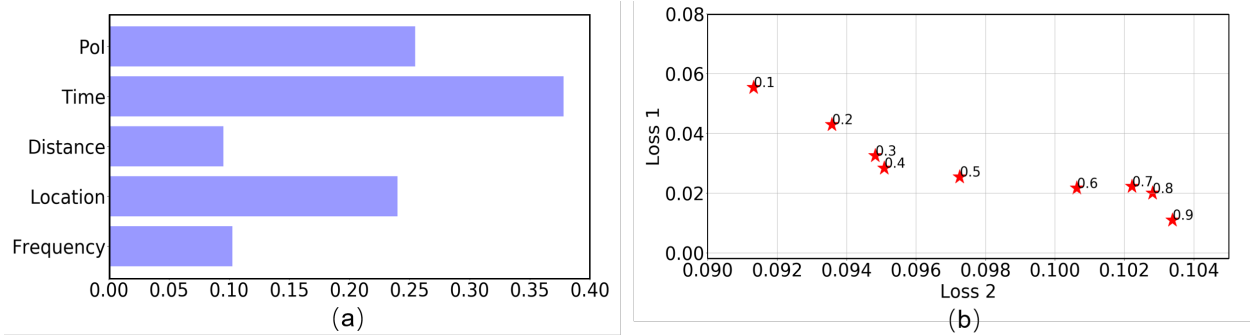


Figure 10: (a) Features importance ranking; (b) Loss value

504 In Figure 11, we show how the level of accuracy of the age group inference based on the proposed model varies
 505 under different parameter settings. We compare diverse learning rate values and change the value of λ in the loss
 506 function. As the results show, $learning_rate = 0.1$ and $\lambda = 0.5$ yield the best performance. The previous experiments
 507 are tested with a fixed percentage of training and validation data (70%). When the value of λ varies from 0.1 to 1,
 508 $Loss_1$ weights less in the overall loss, and $Loss_2$ weights more. When λ equals one, $Loss_1$ is not counted and the
 509 Auto-Encoder-based Compression Module becomes **FCLs**, which still has the ability for compression and extracting
 510 information from the transit stop matrix. The overall model becomes a combination of **Inner-PNN** and **FCLs**. The
 511 result is consistent with the result in Table 3. The two loss functions do not necessarily always conflict with each other
 512 since they all tend to minimize the difference between observed values and certain calculated/predicted values/metrics.
 513 Therefore, varying λ , i.e., changing the weights for the two loss functions does not lead to substantial changes in the
 514 prediction accuracy. At last, the ratio of training data (including training data and validation data) and testing data is
 515 changed from 7 : 3 to 1 : 9. Even only 10% of data is used for training, the proposed method still produces an overall
 516 level of accuracy at 77.29% for inferring the age group.

517 6. Discussion and Conclusion

518 6.1. Discussion

519 This study demonstrates the potential of inferring or recovering traveler attributes or labels based on the observed
 520 trajectories of travelers in public transit systems. Moreover, this study provides new perspectives on utilizing public

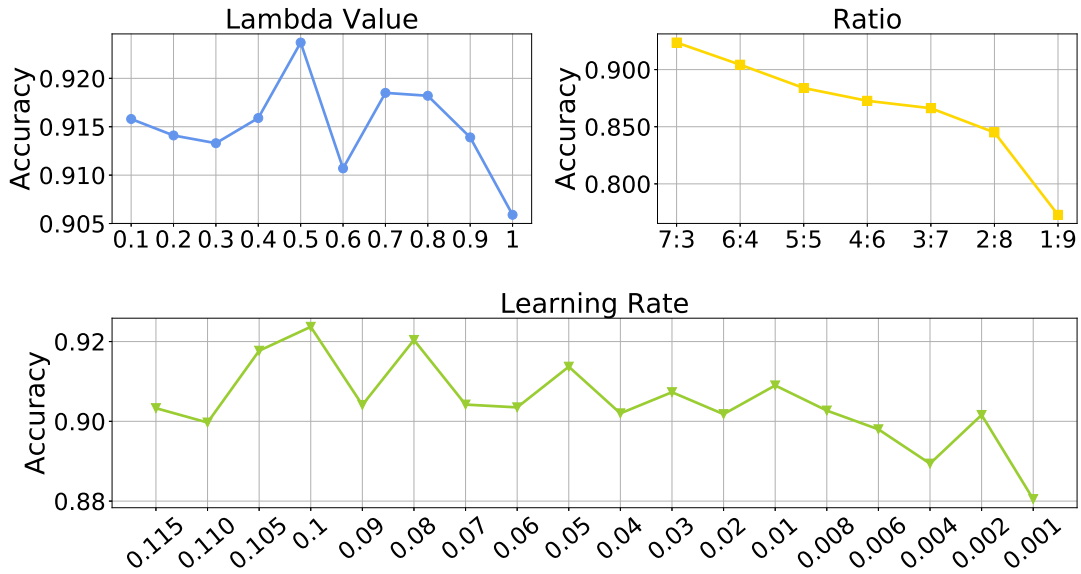


Figure 11: Performance under parameter tuning

transport data to understanding travel behavior patterns, which may be further utilized to improve public mobility services and added services. Several potential applications or implications from this study are briefly discussed below.

523 • Inferring other attributes:

524 In the experiments, this study focuses on inferring the age groups and residential areas of travelers since age
 525 group information can be validated and residential areas can be approximated. Other traveler attributes such as
 526 gender, occupation, income level, may also be inferred through the proposed model, given that these attributes
 527 are highly related to travel behaviours. The proposed method has to be further tested for inferring other attributes
 528 that have a low correlation with travel behaviours.

529 This study indeed offers a general framework to infer individual attributes based on user behaviours. For exam-
 530 ple, based on the extracted features of purchase records (e.g., the types, prices, and the number of the purchased
 531 commodities of the consumers), the attributes (related to purchase behaviors) of these consumers can be inferred.
 532 However, the effectiveness is expected to rely on the level of matching between behaviors and attributes.

533 • Recovering attributes under limited information:

534 If only a part of the data contains a certain attribute or label, the proposed model can be used to infer or recover
 535 attributes or labels for those without true attributes or labels (e.g., due to data loss). During the training process,
 536 only the part of the data with the attributes or labels in concern is used to train the model and fit the parameters of
 537 the model. Since the training and test data are drawn from the same feature space and have the same distribution,
 538 the rest data without true attributes or labels can be inferred by the model.

539 Furthermore, if a dataset does not contain the useful labels or attributes, one may conduct small-scale surveys
 540 and train the proposed model with survey data. Then, the proposed model can be used to infer the attributes or
 541 labels for the whole dataset, which avoids infeasible or very costly large-scale surveys for the whole dataset.

542 The proposed method can also be coupled with transfer learning techniques in order to produce or recover
 543 attributes/labels of passengers in datasets without labels or with insufficient labels for model training. Suppose
 544 we have a dataset with sufficient labels as the source domain and a dataset without labels (or with insufficient
 545 labels) as the target domain, the method of domain adaption, a strategy of transfer learning (Pan and Yang, 2009),
 546 can be coupled with the proposed model to infer the attributes in the target domain. In particular, one can first
 547 filter out the data with high similarities between the source domain and the target domain based on the extracted

548 features, and then train the proposed model in this paper based on these selected data with similarities and infer
549 the attributes of passengers in the target domain.

550 • Planning of Urban Public Services:

551 Since travel patterns can be linked to personal attributes and individual preferences, the attributes of passengers
552 inferred by the proposed model can be helpful for decision making in public services development (e.g., person-
553 alized recommendation or advertising systems, commercial/entertainment site selection, and urban planning),
554 where those inferred attributes of travelers could further help identify the travel preferences of travelers.

555 For instance, knowing the preferred transit lines or departure time of travelers with different attributes, we can
556 design corresponding user-centered public transit systems (including vehicle sizes, transit routes, transit stop
557 spacing, and scheduling). As discussed in Section 1, the transit lines preferred by the elderly may have to be
558 quieter. This may improve the attractiveness and enhance the usability of the public transport systems to different
559 passengers. Also, advertisements aimed at adults can be broadcast more often on the public transit lines that
560 they take more often.

561 Similarly, knowing the origins/destinations of citizens with diverse travel behaviours and attributes can be useful
562 in planning and operation for urban public facilities (Belanche et al., 2016). For instance, by analyzing the
563 destination distribution of passengers, the preference of people in each age group can be further inferred. If the
564 elderly prefers to visit place A, relevant planning authorities may provide more dedicated facilities (e.g., clinics)
565 for the elderly. If the adults are likely to go to place B, more clubs or other entertainment facilities may have to be
566 planned nearby. Overall, the proposed model can produce output that is the input for advertising, planning, and
567 management problems of city-wide public facilities in order to provide a more user-centered service planning.

568 • Implications for Mobility Patterns and Demand Estimation:

569 This study provides additional evidence to the literature that mobility patterns and traveler attributes are highly
570 correlated (e.g., see distinguished mobility patterns for different age groups in Section 3.3). More importantly,
571 the matching between mobility patterns and traveler attributes can be well modelled and captured through data
572 mining techniques (see competitive performance of the proposed method in this paper in Section 5). While
573 this study does not focus on estimation of mobility patterns and demand, it indicates that further data mining
574 techniques can be developed (in the reverse direction compared to this study) to quantify and utilize the match-
575 ing between traveler attributes and mobility patterns, and thus help provide better estimations or forecasts for
576 mobility patterns and travel demand.

577 • Model Improvement for Attributes Inference:

578 Based on the extracted features presented in Section 3.2, the proposed inference method can be further improved
579 by adopting an attention module (Vaswani et al., 2017) to assign different weights to different types of features,
580 which may further improve individual attributes inference quality. Also, the compression module can be en-
581 hanced by more complicated compression frameworks such as Deep Compression model (Han et al., 2016),
582 which is a three-stage pipeline including pruning, trained quantization, and Huffman coding. Such compression
583 operation with a high compression rate can retain important information, while reduce the storage need, and
584 may also further improve the inference accuracy.

585 Furthermore, one may infer the attributes of passengers based on the raw data (e.g., the transit stop matrix of
586 origins/destinations for passengers, the matrix to record the number of trips for passengers in each time period)
587 rather than the pre-defined features proposed in Section 3.2. For instance, utilizing the transit stop matrix of
588 origins/destinations, the graph-based convolutional neural network (Bai et al., 2019b) can be adapted to analyze
589 the spatial correlations among the origins/destinations of passengers. Utilizing the travel frequency matrix, the
590 recurrent neural network (Ke et al., 2017) or its variants can be adapted to analyze the temporal relations and
591 extract useful implicit information. The extracted knowledge can then be fused for further attributes inference.
592 It should be noted that there is a large number of individual trajectories in the dataset, which may contain many
593 irregular trips (or noises). Thus, the proposed model may have to be coupled with techniques to manage the
594 negative effects of noise observations in the dataset on the inference.

6.2. Conclusion

This paper explores the relevance among temporal and spatial information of transit data to identify and visualize the travel patterns of travelers. These patterns are related to individual attributes (e.g. age groups and residential areas), where the patterns of correlations are identified by analyzing the extracted spatial and temporal mobility features. We then develop a new hybrid spatio-temporal neural network to infer the attributes from mobility patterns extracted from the observable trajectories. The proposed model is compared with several benchmark algorithms in the literature including LDA, QDA, SVM, Ada, DT, XGBoost, MLP, PPC, C2AE, and DeepSD. We evaluate the proposed approach by inferring age groups and residential areas of travelers based on real-world large-scale public transit data, which outperforms all the algorithms tested in this paper. The effectiveness of the proposed method mainly benefits from its capability of capturing the spatio-temporal characteristics of travel patterns. This is achieved by taking full advantage of the overall framework and further enhanced by the Inner-Product-based strategy and Auto-Encoder-based method. The pieces of evidence from the comparison of different components of the architecture and the features also suggest that characterizing spatial and temporal correlations in models can greatly improve the classification/inference accuracy. This study also provides insights regarding mobility patterns associated with traveler attributes and shows that people with different attributes may behave differently in a systematic manner.

Acknowledgments

We would like to thank the handling editor, Prof. Zhen (Sean) Qian, and the anonymous referees for their constructive comments, which have helped improve both the technical quality and exposition of this paper substantially. Dr. Wei Liu thanks the funding support from the Australian Research Council through the Discovery Early Career Researcher Award (DE200101793).

References

- Axhausen, K. W., Zimmermann, A., Schönfelder, S., Rindsfuser, G., and Haupt, T. (2002). Observing the rhythms of daily life: A six-week travel diary. *Transportation*, 29(2):95–124.
- Bai, L., Yao, L., Kanhere, S. S., Wang, X., Liu, W., and Yang, Z. (2019a). Spatio-temporal graph convolutional and recurrent networks for citywide passenger demand prediction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2293–2296.
- Bai, L., Yao, L., Kanhere, S. S., Wang, X., and Sheng, Q. Z. (2019b). Stg2seq: spatial-temporal graph to sequence model for multi-step passenger demand forecasting. In *28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1981–1987.
- Bai, L., Yao, L., Kanhere, S. S., Yang, Z., Chu, J., and Wang, X. (2019c). Passenger demand forecasting with multi-task convolutional recurrent neural networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 29–42. Springer.
- Belanche, D., Casaló, L. V., and Orús, C. (2016). City attachment and use of urban services: Benefits for smart cities. *Cities*, 50:75–81.
- Calabrese, F., Di Lorenzo, G., Liu, L., and Ratti, C. (2011). Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):36–44.
- Chang, B., Park, Y., Park, D., Kim, S., and Kang, J. (2018). Content-aware hierarchical point-of-interest embedding model for successive poi recommendation. In *29th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3301–3307.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.
- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.
- Chu, K. F., Lam, A. Y., and Li, V. O. (2018). Travel demand prediction using deep multi-scale convolutional lstm network. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1402–1407. IEEE.
- Cottrill, C. D. (2009). Approaches to privacy preservation in intelligent transportation systems and vehicle–infrastructure integration initiative. *Transportation Research Record*, 2129(1):9–15.
- Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779.
- Guo, K., Hu, Y., Qian, Z., Liu, H., Zhang, K., Sun, Y., Gao, J., and Yin, B. (2020). Optimized graph convolution recurrent neural network for traffic prediction. In *IEEE Transactions on Intelligent Transportation Systems*, pages 1–12.
- Han, S., Mao, H., and Dally, W. J. (2016). Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In *4th International Conference on Learning Representations, ICLR, May 2-4, 2016, Conference Track Proceedings*.
- He, X., Du, X., Wang, X., Tian, F., Tang, J., and Chua, T.-S. (2018). Outer product-based neural collaborative filtering. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2227–2233.
- Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.
- Jahangiri, A. and Rakha, H. A. (2015). Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. *IEEE Transactions on Intelligent Transportation Systems*, 16(5):2406–2417.
- Karlaftis, M. G. and Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3):387–399.

- 650 Ke, J., Zheng, H., Yang, H., and Chen, X. M. (2017). Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal
651 deep learning approach. *Transportation Research Part C: Emerging Technologies*, 85:591–608.
- 652 Kieu, L. M., Bhaskar, A., and Chung, E. (2014). Passenger segmentation using smart card data. *IEEE Transactions on Intelligent Transportation
653 Systems*, 16(3):1537–1548.
- 654 Kusakabe, T. and Asakura, Y. (2014). Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part
655 C: Emerging Technologies*, 46:179–191.
- 656 Lathia, N., Froehlich, J., and Capra, L. (2010). Mining public transport usage for personalised intelligent transport systems. In *2010 IEEE Interna-
657 tional Conference on Data Mining (ICDM)*, pages 887–892. IEEE.
- 658 Li, C., Bai, L., Liu, W., Yao, L., and Waller, S. T. (2019). Passenger demographic attributes prediction for human-centered public transport. In
659 *International Conference on Neural Information Processing*, pages 486–494. Springer.
- 660 Li, C., Bai, L., Liu, W., Yao, L., and Waller, S. T. (2020). Knowledge adaption for demand prediction based on multi-task memory neural network.
661 In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management, CIKM 2020, Beijing, China, October
662 19-23, 2020*. ACM.
- 663 Li, X., Zhao, K., Cong, G., Jensen, C. S., and Wei, W. (2018). Deep representation learning for trajectory similarity computation. In *2018 IEEE
664 34th International Conference on Data Engineering (ICDE)*, pages 617–628. IEEE.
- 665 Li, Y., Wang, X., Sun, S., Ma, X., and Lu, G. (2017). Forecasting short-term subway passenger flow under special events scenarios using multiscale
666 radial basis function networks. *Transportation Research Part C: Emerging Technologies*, 77:306–328.
- 667 Liu, L., Qiu, Z., Li, G., Wang, Q., Ouyang, W., and Lin, L. (2019). Contextualized spatial–temporal network for taxi origin-destination demand
668 prediction. *IEEE Transactions on Intelligent Transportation Systems*, 20(10):3875–3887.
- 669 Liu, W., Wu, W., Thakuriah, P., and Wang, J. (2020). The geography of human activity and land use: a big data approach. *Cities*, 97:102523.
- 670 Liu, Y., Liu, C., Yuan, N. J., Duan, L., Fu, Y., Xiong, H., Xu, S., and Wu, J. (2014). Exploiting heterogeneous human mobility patterns for intelligent
671 bus routing. In *2014 IEEE International Conference on Data Mining (ICDM)*, pages 360–369. IEEE.
- 672 Liu, Y. and Wu, Y.-F. B. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional
673 networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 354–361. AAAI Press.
- 674 Long, Y. and Thill, J.-C. (2015). Combining smart card data and household travel survey to analyze jobs–housing relationships in beijing. *Computers,
675 Environment and Urban Systems*, 53:19–35.
- 676 Luo, F., Cao, G., Mulligan, K., and Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via twitter: A case
677 study of chicago. *Applied Geography*, 70:11–25.
- 678 Ma, W. and Qian, Z. S. (2018). Estimating multi-year 24/7 origin-destination demand using high-granular multi-source traffic data. *Transportation
679 Research Part C: Emerging Technologies*, 96:96–121.
- 680 Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., and Wang, Y. (2017a). Learning traffic as images: a deep convolutional neural network for large-scale
681 transportation network speed prediction. *Sensors*, 17(4):818.
- 682 Ma, X., Liu, C., Wen, H., Wang, Y., and Wu, Y.-J. (2017b). Understanding commuting patterns using transit smart card data. *Journal of Transport
683 Geography*, 58:135–145.
- 684 Ma, X., Tao, Z., Wang, Y., Yu, H., and Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote
685 microwave sensor data. *Transportation Research Part C: Emerging Technologies*, 54:187–197.
- 686 Ma, X., Wu, Y.-J., Wang, Y., Chen, F., and Liu, J. (2013). Mining smart card data for transit riders’ travel patterns. *Transportation Research Part
687 C: Emerging Technologies*, 36:1–12.
- 688 Mohamed, K., Côme, E., Oukhellou, L., and Verleysen, M. (2016). Clustering smart card data for urban mobility analysis. *IEEE Transactions on
689 Intelligent Transportation Systems*, 18(3):712–728.
- 690 Olmos, L. E., Çolak, S., Shafiei, S., Saberi, M., and González, M. C. (2018). Macroscopic dynamics and the collapse of urban traffic. *Proceedings
691 of the National Academy of Sciences*, 115(50):12654–12661.
- 692 Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.
- 693 Pelletier, M.-P., Trépanier, M., and Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C:
694 Emerging Technologies*, 19(4):557–568.
- 695 Ren, H., Song, Y., Wang, J., Hu, Y., and Lei, J. (2018). A deep learning approach to the citywide traffic accident risk prediction. In *2018 21st
696 International Conference on Intelligent Transportation Systems (ITSC)*, pages 3346–3351. IEEE.
- 697 Shifftan, Y., Outwater, M. L., and Zhou, Y. (2008). Transit market research using structural equation modeling and attitudinal market segmentation.
698 *Transport Policy*, 15(3):186–195.
- 699 Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.
- 700 Sun, L. and Axhausen, K. W. (2016). Understanding urban mobility patterns with a probabilistic tensor factorization framework. *Transportation
701 Research Part B: Methodological*, 91:511–524.
- 702 Sun, L., Axhausen, K. W., Lee, D.-H., and Huang, X. (2013). Understanding metropolitan patterns of daily encounters. *Proceedings of the National
703 Academy of Sciences*, 110(34):13774–13779.
- 704 Van Hinsbergen, C. I., Van Lint, J., and Van Zuylen, H. (2009). Bayesian committee of neural networks to predict travel times with confidence
705 intervals. *Transportation Research Part C: Emerging Technologies*, 17(5):498–509.
- 706 Van Lint, J. (2008). Online learning solutions for freeway travel time prediction. *IEEE Transactions on Intelligent Transportation Systems*, 9(1):38–
707 47.
- 708 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In
709 *Advances in Neural Information Processing Systems*, pages 5998–6008.
- 710 Wang, D., Cao, W., Li, J., and Ye, J. (2017). DeepSD: Supply-demand prediction for online car-hailing services using deep neural networks. In
711 *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 243–254. IEEE.
- 712 Wang, S., Cao, J., and Yu, P. S. (2019). Deep learning for spatio-temporal data mining: A survey. *arXiv preprint arXiv:1906.04928*.

- 713 Wang, X., Yao, L., Liu, W., Li, C., Bai, L., and Waller, S. T. (2020). Mobility irregularity detection with smart transit card data. In *Pacific-Asia*
714 *Conference on Knowledge Discovery and Data Mining*, pages 541–552. Springer.
- 715 Wu, L., Yang, L., Huang, Z., Wang, Y., Chai, Y., Peng, X., and Liu, Y. (2019). Inferring demographics from human trajectories and geographical
716 context. *Computers, Environment and Urban Systems*, 77:101368.
- 717 Xu, Y., Belyi, A., Bojic, I., and Ratti, C. (2018). Human mobility and socioeconomic status: Analysis of singapore and boston. *Computers,*
718 *Environment and Urban Systems*, 72:51–67.
- 719 Yang, S., Ma, W., Pi, X., and Qian, S. (2019). A deep learning approach to real-time parking occupancy prediction in transportation networks
720 incorporating multiple spatio-temporal data sources. *Transportation Research Part C: Emerging Technologies*, 107:248–265.
- 721 Yao, D., Zhang, C., Zhu, Z., Huang, J., and Bi, J. (2017). Trajectory clustering via deep representation learning. In *2017 International Joint*
722 *Conference on Neural Networks (IJCNN)*, pages 3880–3887. IEEE.
- 723 Yeh, C.-K., Wu, W.-C., Ko, W.-J., and Wang, Y.-C. F. (2017). Learning deep latent space for multi-label classification. In *Thirty-First AAAI*
724 *Conference on Artificial Intelligence*, pages 2838–2844. AAAI Press.
- 725 Zhao, X., Zhang, Y., Hu, Y., Wang, S., Li, Y., Qian, S., and Yin, B. (2020). Interactive visual exploration of human mobility correlation based on
726 smart card data. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–13.
- 727 Zhong, Y., Yuan, N. J., Zhong, W., Zhang, F., and Xie, X. (2015). You are where you go: Inferring demographic attributes from location check-ins.
728 In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 295–304. ACM.