You Are How You Travel: A Multi-Task Learning Framework for Geodemographic Inference Using Transit Smart Card Data

Yang Zhang[1,*], Nilufer Sari Aslam[1], Juntao Lai[2], Tao Cheng[1]

1. SpaceTimeLab for Big Data Analytics, Department of Civil, Environmental & Geomatic Engineering, University College London, Gower Street, London WC1E 6BT, United Kingdom.
2. Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #10-25 Connexis North Tower, Singapore 138632, Singapore.

Abstract: Geodemographics, providing the information of population's characteristics in the regions on a geographical basis, is of immense importance in urban studies, public policy-making, social research and business, among others. Such data, however, are difficult to collect from the public, which is usually done via census, with a low update frequency. In urban areas, with the increasing prevalence of public transit equipped with automated fare payment systems, researchers can collect massive transit smart card (SC) data from a large population. The SC data record human daily activities at an individual level with high spatial and temporal resolutions. It can reveal frequent activity areas (e.g., residential areas) and travel behaviours of passengers that are intimately intertwined with personal interests and characteristics. This provides new opportunities for geodemographic study. This paper seeks to develop a framework to infer travellers' demographics (such as age, income level and car ownership, et. al.) and their residential areas for geodemographic mapping using SC data with a household survey. We first use a decision tree diagram to detect passengers' residential areas. We then represent each individual's spatio-temporal activity pattern derived from multi-week SC data as a 2D image. Leveraging this representation, a multi-task convolutional neural network (CNN) is employed to predict multiple demographics of individuals from the images. Combing the demographics and locations of their residence, geodemographic information is further obtained. The methodology is applied to a large-scale SC dataset provided by Transport for London. Results provide new insights in understanding the relationship between human activity patterns and demographics. To the best of our knowledge, this is the first attempt to infer geodemographics by using the SC data.

Keywords: Geodemographic inference, smart card data, multi-task CNN, spatio-temporal activity pattern, residential area detection

1 Introduction

Geodemographics is usually defined as the study of people's characteristics (e.g., age and income) by where they live. Obtaining geodemographics of large population is of great significance in social research, business intelligence, public policy, and so on (Martin, Gale, Cockings and Harfoot, 2018). For example,

* Corresponding author: Yang Zhang, E-mail: yang.zhang.16@ucl.ac.uk

geodemographic information can increase the predictability of the market demand in different locations and help commercial sectors to site their stores. It also has considerable use within public service organisations for planning and facilitating resource allocation (Liu and Cheng, 2020). However, geodemographic information is conventionally collected from the census or survey, which is very expensive and time-consuming. Besides, geodemographics is not easy to be obtained on a large scale due to privacy concerns.

To obtain geodemographics, the primary task is to answer the question about 'who you are', which is usually termed as demographic prediction. In the era of big data, with the development of information technology, studies have shown that demographic information can be inferred through many online data sources, such as web browsing logs (Hu, Zeng, Li, Niu and Chen, 2007, Carmel, Lewin-Eytan, Libov, Maarek and Raviv, 2017), mobile phone APP usage data (Zhong, Tan, Mo and Yang, 2013, Dong, Chawla, Tang, Yang and Yang, 2017), and online social network (Perozzi and Skiena, 2015, Volkova, Bachrach and Van Durme, 2016). However, most of the existing literature focuses on predicting demographics based on user's online digital traces, while the discriminative power of human mobility in the physical world has received limited attention. The widespread use of smart devices has been bringing us massive geo-tagged data that record the daily behaviours of individuals, such as GPS trajectories. Many works have reported that human activity patterns are highly intertwined with personal demographic factors (Riederer, Zimmeck, Phanord, Chaintreau and Bellovin, 2015, Zhang, Cheng and Aslam, 2019).

Given the relationship between activity patterns and demographics, researchers have attempted to infer demographics from geo-located data, mainly including GPS trajectories (Ghosh and Ghosh, 2017, Wang et al., 2017) and check-in data (Riederer, Zimmeck, Phanord, Chaintreau and Bellovin, 2015, Zhong, Yuan, Zhong, Zhang and Xie, 2015). For instance, Wu et al. (2019) extracted spatio-temporal and semantic features from GPS trajectories and employed a supervised classification model to infer multiple demographics (e.g. age, gender, and education), separately. Beyond that, another unignorably geotagged data source is public transit smart card (SC) data. Compared to other transport data sources, SC data can provide continuous trip data covering a longer period of time and it is able to link SC data to the individual cardholders (Bagchi and White, 2005). However, existing works about demographic inference based on SC data are rare.

Besides demographic inference, the other key task to obtain geodemographics is to identify 'where you live', namely home location detection. Although a large and growing body of work has studied using various data sources for home location detection (Sari Aslam, Cheng and Cheshire, 2019) or demographic inference (Zhang and Cheng, 2020) in isolation, little attention has been paid to combine them to produce geodemographic mappings. This is because that many data sources (e.g. online digital sources or check-ins) might be suitable to predict demographics but hardly provide sufficient home location information of individuals, and vice versa. In addition, due to privacy concern, increasing people prefer to turn off the smartphone location services, making it challenging to collect geo-tagged data (e.g. GPS) on a large scale for home location inference.

In terms of geodemographic inference, SC data has its unique advantages compared to other data sources. As the modern public transport system plays an increasingly significant role in people's daily life, the equipped Automated Fare Collection (AFC) system can collect massive SC data to reveal activity patterns of individuals. It hence provides an excellent opportunity to explore both the population demographics and home locations for geodemographic mapping. However, to the best of our knowledge, hardly any study has attempted to infer the geodemographics based on transit SC data due to several challenges and limitations. First, the raw SC data are not well represented and noisy. We need to represent SC data properly in order to apply prediction models. Second, the prediction tasks of different demographic characteristics are related to each other. For instance, older people (retired) are unlikely to have a very high income. Existing prediction models always estimate multiple demographics separately, ignoring the correlation between prediction tasks. Third, to date, existing studies about demographic inference seldom contribute to the area of geodemographic research due to the lack of sufficient individual home location information.

To fill these research gaps, this research aims to propose a systematic and feasible framework to obtain geodemographics from human daily activities using SC data. In this framework, the raw SC data are represented as 2D images to capture the spatio-temporal activity patterns of individuals effectively. Based on the representation, we employ a multi-task convolutional neural network (CNN) model to map the activity sequences to demographics. Finally, leveraging the residential areas detected from SC data, geodemographic information can be obtained. The framework is applied to a case study of public transit users in Greater London, UK to validate its effectiveness.

The remainder of this paper is organised as follows. Section 2 reviews the related work. Section 3 describes the datasets used in this study and the pre-processing step. Next, Section 4 provides an overview of the methodology. Section 5 elaborates the case study to evaluate the framework and the paper is concluded in Section 6 with a discussion of future research directions.

2 Related Works

Geodemographic is an analysis of 'who you are' by 'where you live'. The former research question (who you are) is usually referred to as demographics prediction and the latter (where you live) is termed as home location detection. Hence, related works are presented from these two aspects. We also summarise the limitations, the proposed approaches, and the contributions of our work.

2.1 Demographic Inference Using Geo-Tagged Data

Over recent years, a vast amount of data with location information has become available. Such geo-tagged information is timely, detailed and specific to each individual, yet it becomes a valuable source to reconstruct human trajectories and explore the human activities. Extensive studies have shown that the activity patterns correlate with the demographic characteristics of individuals. For example, Siren and Hakamies-Blomqvist (2004) showed that demographics (e.g., gender, the presence/absence of a driver's license and place of

residence) were strongly associated with the mobility of elderly citizens in Finland. Other personal traits, including age, working status, education level, were also identified to be related to individuals' activity patterns (van den Berg, Arentze and Timmermans, 2013, Bantis and Haworth, 2017, Zhang, Zhang and Zhou, 2019). Concerning the relationships between demographics and spatio-temporal activity patterns, researchers began to utilise geo-tagged data to infer demographics. A recent survey has summarised several works about demographic attribute prediction using physical geo-tagged footprints (Gao, Zhang and Zhou, 2019).

From the perspective of data sources, it shows the majority of existing studies mainly focused on inferring demographics from check-ins or GPS trajectories. For example, Zhong, Yuan, Zhong, Zhang and Xie (2015) proposed a 'location to profile' framework, in which spatiality, temporality and location knowledge were extracted from check-in data to profile users. Ghosh and Ghosh (2017) utilised transfer knowledge derived from massive GPS trajectories of a geographically distanced but semantically similar type of region of interest (ROI) to categorise mobile users. Zhu, Gonder and Lin (2017) leveraged individuals' long-term GPS data to retrieve home-based trips. Then, travel behaviour variability features were extracted for estimating the social-demographic information by a supervised learning approach. However, all abovementioned works were based on hand-crafted features extracted from geo-tagged data. The feature engineering process is time-consuming and important information might be lost through the aggregation since activity patterns are not only characterised by the aggregated attributes but also the organisation order of activities (Ilägcrstrand, 1970).

In urban areas, with the increasing prevalence of public transit with AFC systems, researchers can collect massive transit SC data from a large population. The availability of SC data has motivated a considerable number of studies that analyse the relationship between demographics and SC data-based activity patterns (Mohamed, Côme, Oukhellou and Verleysen, 2017, Zhang, Cheng and Aslam, 2019). Although some possible relationship has been well-documented in many works, very limited attention has been paid to estimate demographics from SC data. Lately, Zhang and Cheng (2018) explored inferring demographics by leveraging a variety of spatial and temporal features extracted from the raw transaction records. Ding, Huang, Zhao and Fu (2019) developed a deep learning model to estimate socioeconomic status using temporal-sequential features and general statistical features generated from SC data. However, the success of these works heavily relied on elaborated feature engineering. The extracted features, such as the number of travel days, average travel length, and average departure time of the first trips, were derived from a scalar aggregation of an individual travel diary. These indicators ignored the organisation of multiple journeys over time (Goulet Langlois, Koutsopoulos and Zhao, 2016). In order to take full advantages of the knowledge embedded in SC data, more effective representation of SC data and more advanced mining techniques need to be used. Very recently, Zhang and Cheng (2020) presented an 'end-to-end' thresholding multi-channel CNN model to infer the working status of passengers using 2D temporal profiles reconstructed from SC data. This framework did not require manual feature extraction process, but it ignored the spatial information of the individual activity patterns.

From the perspective of prediction methods, demographic inference has been treated as a supervised learning task. It is found that most previous frameworks have followed a flow chart of extracting measures of temporal and spatial regularity first, and then utilising these features as the input of conventional machine learning methods (e.g. support vector machine, random forest, and neural network) to predict multiple demographics (Zhong, Tan, Mo and Yang, 2013, Zhu, Gonder and Lin, 2017, Zhang and Cheng, 2018, Wu et al., 2019). In these frameworks, besides the abovementioned feature engineering issue, they also overlooked the correlation between different demographic prediction tasks. It has been well-documented that different demographic inference tasks might be correlated.

To tackle this problem, multi-task learning (MTL) can be used to improve the performance by building models collectively to take advantage of the knowledge from all tasks. A typical way of MTL is to learn tasks in parallel with a shared representation. Recently, several studies have employed MTL for demographic prediction. For instance, utilised a multi-task SVM model to infer demographics using mobile phone records. Wang, Guo, Lan, Xu and Cheng (2016) proposed a multi-task representation learning model to predict personal traits of twitter users. Although MTL improves the prediction accuracy, these works also required the manual feature extraction and selection step because the base model used in the MTL framework was the traditional machine learning classifier. Alternatively, deep learning (DL) techniques have been employed as the base model in MTL to avoid task-specific feature engineering. This is because a typical DL model can automatically discover the required features from input data level by level, which is also called 'end-to-end' learning. Many DL or multi-task DL models have been proposed for traffic prediction (Zhang, Zheng, Sun and Qi, 2019, Zhang, Cheng and Ren, 2019, Zhang, Cheng, Ren and Xie, 2020). However, using DL approaches for SC data analysis is quite rare, since the raw SC data cannot be directly fed to DL models. To the best of our knowledge, our work is the very first to use multi-task DL for demographic prediction.

2.2 Home Location Detection From SC Data

Compared to other transport data sources, SC data can provide continuous trip data covering a longer period to reveal the frequently visited locations of individual cardholders (Bagchi and White, 2005). It hence provides an excellent opportunity to infer home locations from SC data. Li, Yu, Ng, Wu and Goh (2015) developed a probabilistic approach to infer the home locations using SC data in Singapore, by exploring the underlying repeated travelling patterns between home and workplace. However, it overlooked that passengers might use different bus/tube stations to access home. Long and Thill (2015) presented a decision tree method to detect home locations from one-week Beijing SC data for commuting analysis, but this approach only considered the first boarding station in each day as the potential home locations. Sari Aslam, Cheng and Cheshire (2019) proposed a heuristic model to detection home locations from London's Oyster card data, but it was only applicable for tube users because the alighting bus stops were missing in SC data. Additionally, in previous works, the home locations detected from SC data are mainly used for demand estimation or commuting/travel behaviour analysis (Zhang, Zhang and Zhou, 2019). Hence, the results of home location identification are

always aggregated in the bus/tube stations. In terms of geodemographic mapping, it is necessary to explore an applicable way to aggregate the results in appropriate geographical unit, e.g. administrative districts.

2.3 Summary

After reviewing the related works, there are three main limitations in these diagrams:

(1) First, the feature extraction approaches are limited in capturing the time-ordered activities during the study period. Besides, handcrafted features are usually non-robust and rest too much on expert knowledge and experience.
(2) Second, existing frameworks overlook the correlation between multiple demographic prediction tasks, which may decrease the prediction accuracy and efficiency (Ruder, 2017).
(3) Third, current studies have limited to demographic inference or home location detection in isolation. To the best of our knowledge, there is no work attempting to link travel behaviour, geography and demography to produce predicted geodemographic mappings, which is significant to expand the real-world applications of the demographic estimates in urban-related fields, such as transport planning, business intelligence and policy-making (Singleton and Spielman, 2014, Liu and Cheng, 2018).

This paper aims to explore whether and how SC data can be used to estimate geodemographics. To cope with abovementioned issues, we present a systematic framework of geodemographic inference based on SC data. The proposed framework has four main components. First, a decision tree diagram is used to detect the passengers' residential areas from SC data. Second, we represent the raw, multi-week SC data as a two-dimension image to depict each user's spatio-temporal activity pattern. Third, benefit from the image-like representation, a multi-task CNN model is provided to infer multiple demographic characteristics based on the 2D images, which overcome the second issue listed above. CNN (LeCun, Bengio and Hinton, 2015) is a state-of-the-art deep learning model for image processing. In addition, the MTL architecture can improve the performance of multiple related tasks by taking advantage of the similarities between different tasks. Finally, the detected residential areas are aggregated in administrative districts and combining demographic estimates, geodemographic mapping can be obtained.

The contributions of this study are twofold. The first is the proposed systematic framework of geodemographic inference from SC data. This framework can improve the prediction accuracy. Compared to census, it provides a feasible and fast way to map out timely geodemographic information from SC data. To the best of our knowledge, this is the very first attempt to infer geodemographics by using SC data. Second, from an empirical perspective, we provide a case study using large-scale SC data collected in London, UK. It is important to understand the association between the passenger's mobility and demographic attributes. In addition, this research also suggests that confidential location information should be hidden to avoid potential privacy leak from the abusive use of the proposed framework.

3 Dataset

The datasets used in this paper come from Transport for London (TfL), which is the public transport authority of Greater London, UK. It consists of a sample of Oyster card data and the London Travel Demand Survey (LDTS) data.

3.1 Oyster Card Data

The Oyster card is a smart card used to pay for journeys on all London public transit systems. According to TfL's report (TfL, 2013), daily journeys made by public transit systems account for about 37% of London's journeys every day. London's public transit has served as the most important travel mode in London. The SC data used in this study is a sample of over 0.86 million transactions made by around 0.3 million passengers in bus and tube networks in London in March 2013. In this paper, a *journey* is defined as one-way travel from one station to another. This dataset consists of 34% tube journeys and 66% bus journeys. Tube records contain the boarding and alighting stations and time. In contrast, bus records only include the origin stations and the start time, but not alighting information. All transactions also include unique user IDs. For privacy concerns, all user IDs were encrypted.

3.2 London Travel Demand Survey

Demographic information was collected via the London Travel Demand Survey (LTDS) in 2013. LTDS is an annual survey carried out in Greater London. Each year, 8000 households are randomly selected by TfL to be interviewed about their travel habits. This survey is completed by every household member aged five or over. The LTDS also collects the demographic information such as gender, household income, car ownership and the outcode (the first part of a UK postcode) of home location. The respondents voluntarily provided their Oyster card IDs. This enables TfL to match the survey to the Oyster card records.

3.3 Data Pre-processing

First, the usage frequency of public transit varies significantly among different users. Some passengers use public transportation for most of their daily travels, while others use it occasionally for specific purposes. The low-resolution SC data of occasional passengers cannot present their daily activities. In order to ensure the Oyster card data can reveal a relatively complete activity pattern of an individual, we first need to identify the frequent passengers. To do so, three variables, namely the number of journeys, the number of travel days, and the spread days between the first and the last travel days during the study period, are used to describe the individual's usage frequency. Subsequently, k-means is utilised to classify the passengers into three groups: occasional, moderate and frequent passengers. The means of the three variables of the three groups are given in Table 1. There are 10,495 passengers (accounting for 34.5% of all passengers in the available sample) are identified as frequent users, who travelled on most of the days during the study period and averagely made about two journeys per day.

Table 1 The mean of the three variables of the different passenger groups.

Categories	Percentage	Mean of the number of journeys	Mean travel days	Mean spread days
Occasional	24.1%	5	2	5
Moderate	41.4%	19	8	24
Frequent	34.5%	73	21	29

Second, we link the LTDS data to the Oyster card data by matching the user IDs. Finally, among the 10,495 frequent users, there are 2493 passengers are identified as LTDS respondents (for whom both demographic data and Oyster card data are available). They carried out about 179,017 journeys. The demographic data of these 2493 LTDS respondents are used as the ground truth to train and validate the proposed demographic prediction model. In this paper, we focus on the inference of four demographic attributes, namely age group, gender, income level and car ownership. The demographic characteristics in LTDS of 2493 passengers are summarised in Table 2.

Table 2 Description of the demographic attributes of the 2493 LTDS respondents.

Demographic attributes	Number of labels	Categories and fraction
Age	3	Young (<30): 28.40% Adults (30-65): 53.87% Elder (>65): 17.73%
Gender	2	Male: 43.68% Female: 56.32%
Income level	3	Low income (<£20k): 32.45% Middle income (£20k-£50k): 39.83% High income (>£50k): 27.72%
Car ownership	2	Have no private cars: 58.56% Have private cars: 41.44%

Next, the Oyster card did not record the alighting stations and time information of bus journeys because the bus ticket prices do not rely on the travel length or traffic zone. To reconstruct the activities from SC data, the first step is to infer the missing alighting information. We employ the method developed by Gordon, Koutsopoulos, Wilson and Attanucci (2013). Approximately 84% of bus journeys can be successfully inferred. For an alighting station that cannot be deducted, we simply assume it to be the bus station nearest to the next boarding station. This assumption has no significant impact on the results.

4 Methodology

4.1 Methodological Framework

To infer the geodemographics from SC data, we propose the concept that 'you are how you travel'. This means that if we can represent SC data describing the activity patterns of individuals (including stays and travels), people's demographics can be inferred. By integrating residential area information, we can then produce the geodemographic mapping of an area of interest. For such a purpose, the proposed framework should be capable of:

- Detecting residential areas from SC data;
- Representing the raw SC data in a proper form, describing the activity patterns of individuals, as well as being used as the input of demographic inference model;
- Building a supervised learning framework for multiple demographic attributes inference;
- Producing the geodemographic mapping of an area of interest leveraging the inferred demographics and residential areas.

The proposed framework consisting of the four main steps is demonstrated by the flow chart in Figure 1. The abovementioned four objectives are realised by four main steps as follows:

1) **Residential Area (RA) detection**: employing a decision tree diagram to identify the RA of each individual from SC data;
2) **SC data representation**: extracting stay areas of individuals from SC data, reconstructing the activity sequences and representing the SC data into 2D images, as their spatio-temporal activity profiles to depict their activity patterns;
3) **Multi-task CNN for demographic inference**: Leveraging a multi-task CNN model to predict multiple demographics simultaneously;
4) **Geodemographic mapping**: Linking the RA to administrative divisions. Combining the demographic estimates and RA information to produce a geodemographic mapping.

The rest of this section is organised according to the flow chart, illustrating the method developed in each step.

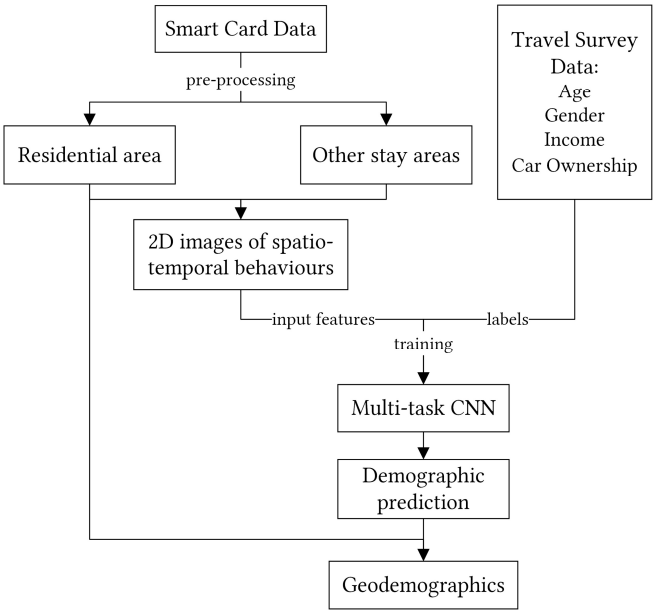


Figure 1. Flow chart of the framework.

4.2 Residential Area Detection

In this paper, a residential area (RA) is defined as a circular area with a radius of 500 m, which is the average radius of a walkable stay area in London (Goulet Langlois, Koutsopoulos and Zhao, 2016, Sari Aslam, Cheng and Cheshire, 2019). To detect residential areas from SC data, we assume that the boarding station of the first journey or the alighting station of the last journey in a day to be the potential RA centre of a cardholder[†]. For convenience, the boarding station of the first journey or the alighting station of the last journey in a day is referred to as 'candidate station' in this paper.

To detect RAs, a decision tree diagram is proposed, as shown in Figure 2. The flowchart and detailed conditional control statements are described as follows:

1) Firstly, we count the *visit-frequency* (vf) of each candidate station during the study period.
2) Next, if there is only one candidate station, this station is identified as the centre of an RA.
3) If not, these candidate stations are clustered into groups based on their geographic coordinates. This is because a passenger may use different stations to access his/her home. The clustering starts by treating each potential station as a separate cluster. Then, it repeatedly executes the following two steps: (1) identify the two clusters that are closest together, and (2) merge the two nearest clusters if the distance between the two clusters is smaller than a predefined distance threshold d_{max}. This continues until the distance between any pair of clusters is greater or equal to d_{max}. The distance between two clusters r and s is defined as:

[†] There could be a minor bias in identifying residential area based upon this assumption since a passenger in London may cycle or drive a private car to the station prior to their first journey or to their residential area after their last journey.

$$D(r,s) = \max\left\{d(i,j):\ \text{where station } i \text{ is in cluster } r \text{ and station } j \text{ is in cluster } s\right\} \qquad (1)$$

The $d(i,j)$ is the Euclid distance between the candidate stations i and j. The threshold distance d_{max} is set to be 1 km, which is twice the radius of an RA.

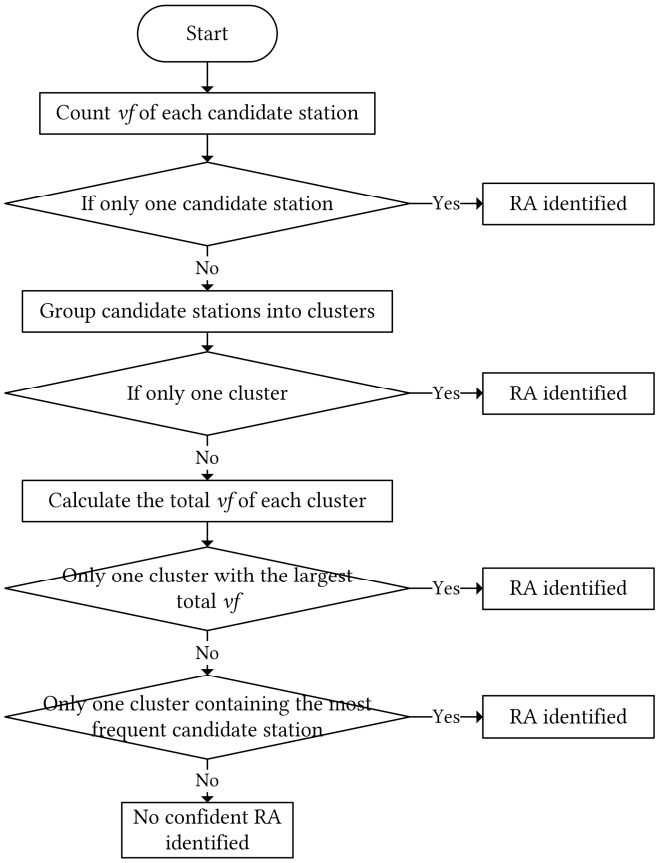


Figure 2 A decision tree diagram for residential area detection.

4) If there is only one cluster, the circular area defined by this cluster is the identified RA.
5) If not, we sum the *vf* of all candidate stations within the same cluster.
6) After that, if a single cluster has the maximum total *vf*, the '*vf*-weighted' centre of this cluster (arithmetic mean coordinate of all stations weighted by the *vf* within this cluster) is taken as the RA centre. For example, in Figure 3 (a), there are three clusters with total *vf* to be 2, 2, and 23, respectively. Thus, the area characterised by the third cluster is taken as the RA.
7) Otherwise, when multiple clusters have the same largest *vf*, the cluster containing the most frequent candidate station is treated as the RA. An example is given in Figure 3 (b). It is observed that the clusters 1 and 3 have the same maximum total *vf*, but cluster 1 contains the most frequent candidate station with *vf*=10. Therefore, the identified RA is determined by cluster 1.
8) Finally, if there are more than one most frequent candidate stations belonging to different clusters, there is no confidential RA can be identified.

Note that in step (6), we choose the cluster with the largest total *vf* as the most potential RA area, no matter whether the candidate station with the highest *vf* is within this cluster. It means the total *vf* of a cluster is prior to the *vf* of a single candidate station in this decision process. This is because users may use different stations to access home.

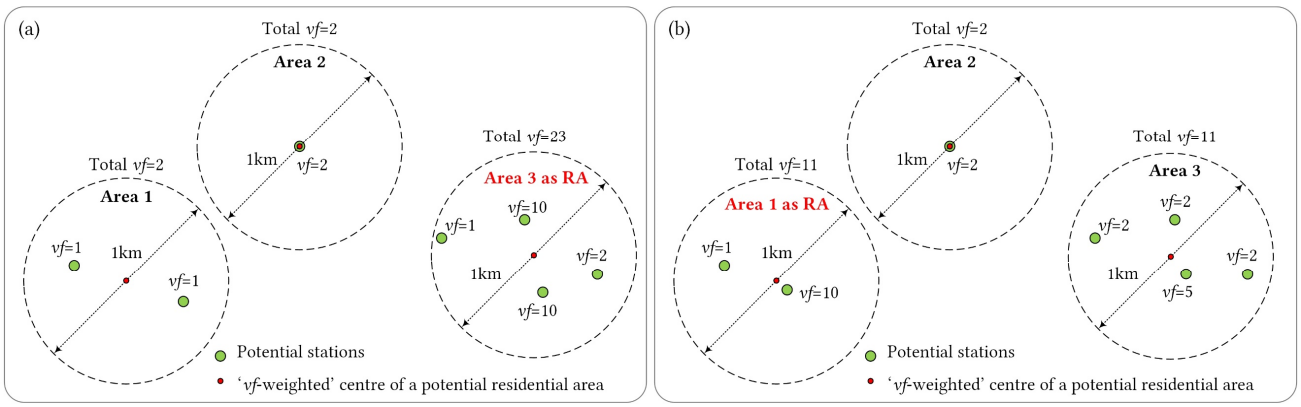


Figure 3 Two examples of the identified RA. (a) RA identified by the cluster with the largest total *vf*; (b) RA identified by the cluster containing the most frequent candidate station.

4.3 Smart Card Data Representation

As suggested by previous works, demographics are related to spatio-temporal activity patterns of individuals, which can be conceptualised as a sequence of stays and travels between some areas. SC data can reveal the series of these activities over time. Thus, effective representation of SC data to capture the activity pattern is fundamental for demographic inference. This is realised by three steps: 1) trip chain generation; 2) stay area identification; 3) SC data representation as two- dimensional images. Details are given below.

4.3.1 Trip Chain Generation

In this paper, a ***journey*** is defined as one-way travel from one station to another. A ***trip chain*** is defined as a series of journeys made by a passenger on a daily basis and is considered as a useful way to demonstrate passengers' activities (Ma, Wu, Wang, Chen and Liu, 2013). In many cases, a passenger may perform an activity while transferring between public transit services. The short-stay between the transfer cannot reflect the passenger's real travel purpose. Thus, identifying transfers and reconstructing the trip chains are therefore crucial for SC data representation. To do so, the journeys of each passenger are ordered by time. Afterwards, public transit transfers are identified based on the temporal and spatial constraints of their SC data to generate trip chains. In terms of Oyster card data, the maximum transfer time is 45 min, and transfer distance is 750 m, referring to (Gordon, Koutsopoulos, Wilson and Attanucci, 2013). This means that if the transaction time between two consecutive journeys in SC data is less than 45 min and the distance between the last alighting and the next boarding stations is less than 750 m, the two consecutive journey stages should be linked. Failing any one condition will label the former journey as not linked to the next.

4.3.2 *Stay Area Identification*

After obtaining the trip chains, all start and end stations of trip chains of each passenger are spatially clustered based upon their geographical coordinates. The clustering method is the same as the approach to group the candidate stations for residential area detection, described in Section 4.2. Each cluster is identified as a stay area of an individual. After that, the total *vf* of each stay area is calculated. Finally, all stay areas of an individual are ranked by their summed *vf* values and the average stay time in descending order. The average stay time is used to rank stay areas with the same summed *vf* values. Both the visit frequency and the stay time can imply the travel purpose to some extend (Wang, de Almeida Correia, de Romph and Timmermans, 2017, Sari Aslam, Cheng, Cheshire and Zhang, 2019), which has been used to profile people (Shen and Cheng, 2016).

4.3.3 *Activity Profiles as 2D Images*

SC data cannot be directly used for demographic prediction. To represent the spatio-temporal activity pattern, we propose to reconstruct the SC data of a passenger as a two-dimensional image. In this approach, the start and end stations of trip chains are linked to these identified residential or stay areas. If a journey's destination does not match the origin of the next journey, the interval between the two journey stages are equally divided and linked to the two different stay areas, respectively. If there is no journey observed in a day, the intraday status is denoted as unknown. In this way, all journeys made by a passenger can hence be linked to a sequence of activities, characterised by stay/residential areas or travel behaviours with specific durations. Furthermore, the activity pattern is represented by an $N \times M$ array, where N is the number of predefined time slots and M is the number of days during the study period. Each cell of an image is associated with a stay/residential area, or travel behaviour, or unknown status. Note that the duration of travelling is always shorter than staying in an area. In order to contain the travel behaviours in an activity pattern as possible, a time slot will be labelled as travelling by tube/bus once there is a tube/bus journey being detected during this time slot. If both tube and bus journeys exist in the same slot, we select the travel behaviour with a longer duration as the activity status of this time bin.

The notation of different activities is summarised in Table 3. For each individual, his/her stay areas are ranked by the summed *vf* values and the average stay time in descending order for notation. It is worth noting that the number of stay areas is varied from user to user. Some users may have many occasional visits to different areas, hence their maximum number of identified stay areas might be much larger than that of the majority. To deal with these exceptional situations, we propose to only differentiate the K_{3sigma} most frequently visited areas (including home location). The value of K_{3sigma} is determined based on all passengers' number of activities using the three-sigma rule. For example, if K_{3sigma} is identified to be ten, it means 99.7% passengers have less than ten stay areas. Consequently, for passengers who have more than ten activity areas, their first ten activities are denoted using number 1 to 10 (as shown in Table 3). All the other occasional stay areas are identically denoted using number 11. The reason for this is that we usually need to rescale the 2D image into

a predefined range (i.e. usually [0, 1] or [-1, 1]) before inputting them into a predictive model. Limiting the maximum value using three-sigma rule is equivalent to remove outliers (noise) from the images. In addition, we assume that the stay areas with higher visit frequency and longer stay duration play a more important role than very occasional stay areas in terms of passenger profiling.

Using one-hour time bin, an example of the 2D image generated from a passenger's Oyster card data is given in Figure 4. For visualisation, the activities (stays, travels or unknown status) are depicted using different colours in Figure 4. The duration of each activity is captured by the length alongside the y-axis. Each column then denotes a daily spatio-temporal activity pattern.

Table 3 The notation of activities.

Notation	Activity
-2	Travel by bus
-1	Travel by tube
0	Unknown status
1	Stay in the esidential area
2	Stay in stay area 1
3	Stay in stay area 2
⋮	⋮
k+1	Stay in stay area k
⋮	⋮
K_{3sigma}+1	Stay in stay areas K>K_{3sigma}

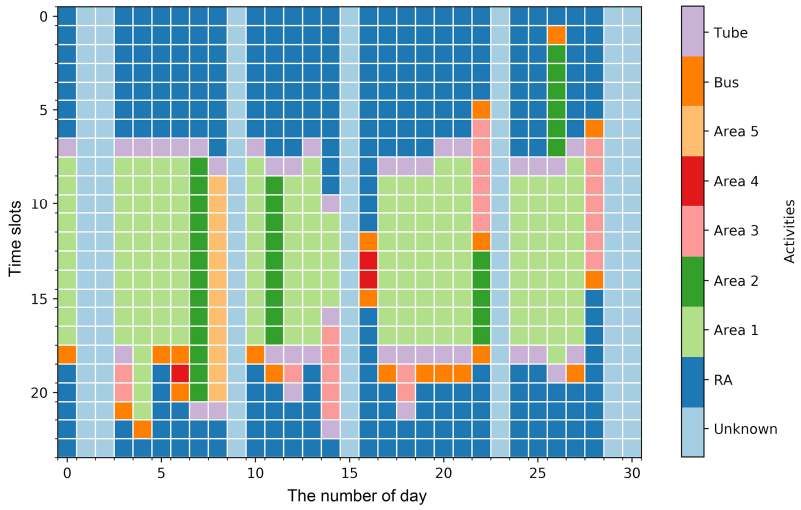


Figure 4 A two-dimensional image generated from a passenger's SC data. It describes the activities distributed over one-hour time slots of each day of the month.

4.4 Multi-Task CNN for Demographic Prediction

After extracting activity profiles of individuals, the demographics can be inferred using classifiers. A wide range of traditional machine learning models has been applied for performing the classification task, including support vector machine, decision tree, and Naïve Bayes (Zhong, Tan, Mo and Yang, 2013, Zhao et al., 2017). These classifiers require hand-crafted features as input for training. In this study, as SC data are represented as 2D images, we propose to employ the state-of-the-art image processing model, namely CNN, to learn features from images for demographic inference automatically.

Additionally, the conventional approach for multiple demographic attributes prediction is to train different classifiers for different prediction tasks. However, it is observed that the four tasks are relevant. For example, the average income level of elderly passengers is lower than that of the middle age. The car ownership of female passengers is lower than the male. In this case, multi-task learning (MTL) may improve prediction performance, as suggested by (Zhong, Tan, Mo and Yang, 2013, Vijayaraghavan, Vosoughi and Roy, 2017). Motivated by this, we propose to combine MTL with CNN framework by sharing some layers between different tasks.

As CNN has been widely used in imaging processing, this paper only provides a simple introduction to CNN. More details of the CNN model can be referred to (LeCun, Bengio and Hinton, 2015). Briefly, a CNN generally consists of multiple hidden layers between an input and an output layer. The hidden layers of a CNN typically consist of a series of convolutional layers and pooling layers, as well as fully-connected layers. A *convolutional layer* comprises a set of learnable convolutional kernels (defined by a width and height) to convolve the input. A non-linear activation function (e.g., the Rectified Linear Unit (ReLU) function) is usually performed after the convolution to add non-linearity to the network. Subsequently, a convolutional layer produces a feature map and passes it to the next layer, typically a pooling layer. The *pooling layer* reduces the dimensions of the image by subsampling the feature map extracted by the convolutional layers. It is commonly inserted between successive convolutional layers. In this paper, a popular pooling method is used, i.e. average pooling, which keeps the average value within a subarea (typically 2×2 size) of the feature map. After a sequence of convolutional layers and pooling layers, the outputs are flattened and concatenated as a single vector. Afterwards, the vector is fed into several *fully-connected layers*, which connect every neuron in one layer to every neuron in the next layer. For a classification task, a *softmax layer* is added before the final output layer to determine the probability of multiple classes at once. Finally, an object (e.g., an image) is identified to belong to the class with the maximum probability.

Based on the basic components of CNN, a multi-task CNN is proposed, visualised in Figure 5. In this configuration, the images are first fed into several shared layers. The weights of these shared layers are common to all the tasks. Subsequently, the output of the shared layers goes into task-specific layers, where computation related to each task is carried out separately without sharing the learnable parameters across

the layers. In these task-specific layers, the network layers learn features to a specific task. Finally, these layers produce the outputs of each task separately.

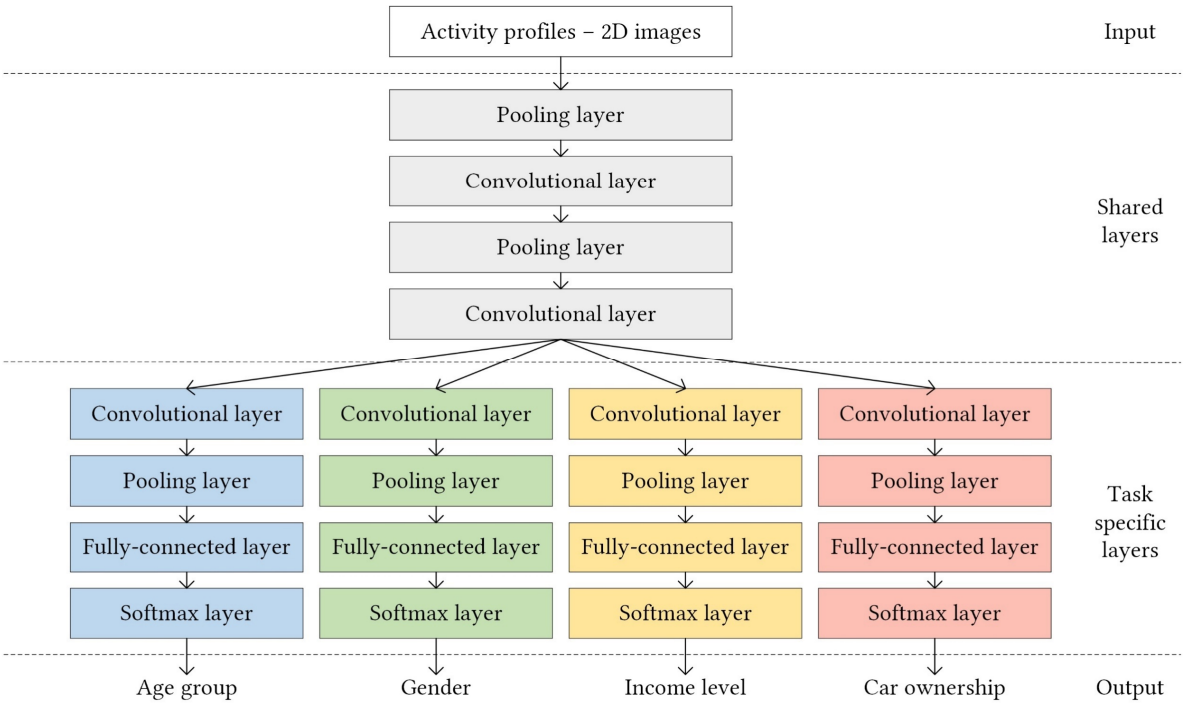


Figure 5 The configuration of the proposed multi-task CNN

The target of model training is to jointly minimise the loss of all tasks. The softmax cross entropy is used as the loss of each classification task. In the proposed model, the loss function is defined as a weighted sum of the loss of every task, written as

$$loss = \sum_{k=0}^{T} w_k \cdot \left(-\sum_i y'_{i,k} \log\left(y_{i,k}\right) \right) \tag{2}$$

where $y'_{i,k}$ and $y_{i,k}$ are the output of the softmax layer and the real label of the i-th sample in the k-th task, respectively, w_k is the weight assigned to the k-th task's softmax cross entropy loss and T is the total number of tasks. The optimal weight of each task is manually tuned using the grid search method. During model training, the model calculates the loss after forward-propagation, and then optimises all the learnable parameters by back-propagation with the optimiser Adam (Kingma and Ba, 2014). By minimising the loss function, all learnable parameters in the multi-task CNN are well trained.

4.5 Geodemographic Mapping

Geodemographics study the demographics of people based on where they live. Geodemographic mapping can widen the application of demographic inference results, such as helping life insurance companies and pension funds to assess longevity for pricing and reserving. This study combines the inferred home location information (i.e. RA) and demographics to produce geodemographic mappings.

To do so, we first snap the RAs to the geographic units, such as the administrative districts of the study area. Note that some tube/bus stations within an RA might be located at the boundary of two or more adjacent units. If there is only a single station within the RA, we think that the passenger lives in the geographic unit which contains the station. If there are multiple stations within the RA, we propose to measure the overlapping area between the RA and each geographic unit. The unit with the maximum overlapping area is hence taken as the one where the passenger lives. For example, in Figure 6, the recognised RA overlaps with three districts. District 1 is treated as the most likely home location since it has the largest overlapping area. Afterwards, the statistics of demographics can be calculated in each of the geographic units to produce a geodemographic mapping, for instance, the distribution of elderly people across administrative districts. Geodemographic data can be used for geodemographic classification or segmentation for intelligence business, etc., but this is beyond the scope of this study. We do not discuss here.

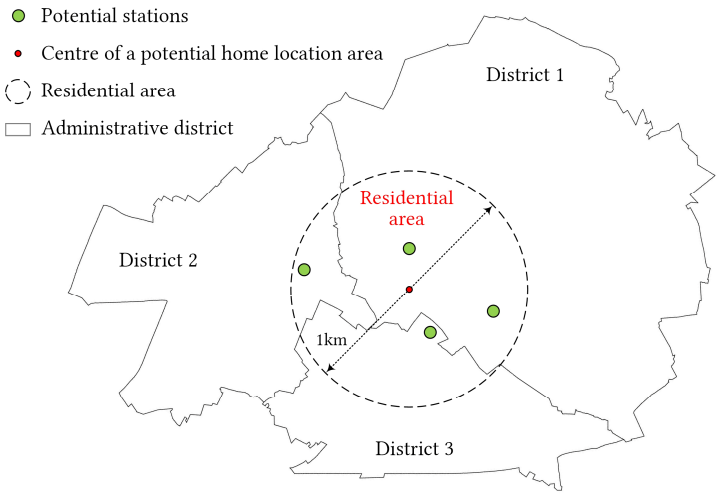


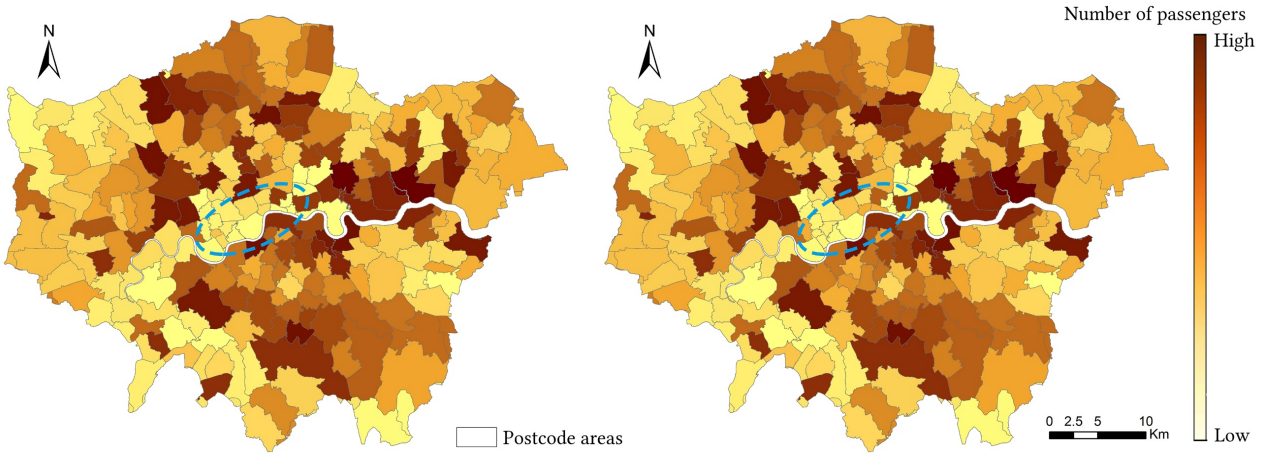
Figure 6 An example of assigning a residential area to an administrative district.

5 Case Study

5.1 Residential Area Detection Results

Leveraging the approach proposed in Section 4.2, there are 99.1% of 10,495 frequent users' residential areas can be identified based upon the first boarding and last alighting stations in Oyster card data. To validate the results, we use the 2493 LTDS respondents' home location information in the survey, in which they provided the outcodes of their home addresses to measure the accuracy. In the UK, the postcoding system was designed by the Royal Mail for efficient mail delivery. A postcode consists of two parts. The first part, commonly known as the 'outcode', suggests a postcode area while the second part (i.e., 'incode') indicates a postcode district. For results validation, the identified RA of each LTDS respondent is assigned to a postcode area using the method described in Section 4.5. The accuracy of RA detection of these LTDS respondents is 97.4%, which shows the effectiveness of the proposed method. The distributions of the inferred and the real RAs of the

LTDS respondents in the sample are shown in Figure 7. Overall, the inferred results are highly consistent with the real RA distribution. Slight differences between the two maps can be observed in central London (highlighted using the blue dotted circle). This is probably because the inner area has a denser distribution of bus/tube stations. There are hence more passengers whose RAs overlapped multiple postal areas, leading to a relatively high error ratio when snapping the RAs to the postal areas. Overall, the proposed method can produce the correct results in most of these cases.



(a) The inferred residential area distribution (b) The real residential area distribution

Figure 7 The identified and the real residential area distribution of LTDS respondents.

5.2 Oyster Card Data Representation

In this paper, the time slot for SC data representation is set to be one hour. Hence, the one-month Oyster card data is represented as an image of size 24×31. To construct the 2D images of spatio-temporal activity patterns, we first detect the stay areas from Oyster card data for each passenger. Figure 8 displays the histogram of the number of identified stay areas of all passengers. Briefly, the maximum number of stay areas is 39 among all passengers and the average stay area count is 8. According to three-sigma rule, the value of K_{3sigma} is 23, thus the maximum value in the 2D image is 24. Afterwards, each time slot is associated with a stay area or bus/tube travel activity, as shown in Figure 4. Before demographic prediction, all pixel values of these images are rescaled to [0, 1] using the min-max normalisation.

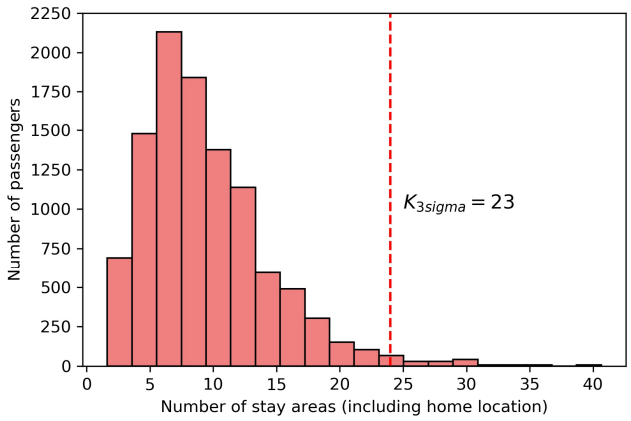


Figure 8 The histogram of the number of stay areas (including home location) of passengers

5.3 Demographic Prediction Results

5.3.1 Experiment Settings

After representing the SC data into images, the proposed multi-task CNN is used for multiple demographic attributes inference. The prediction is formulated as several supervised classification tasks. Therefore, we use the Oyster card and LTDS data of the 2493 respondents to train, validate and test the proposed multi-task CNN model. The dataset is split into three parts: 70% for training, 10% for validation and 20% for testing. The validation set is not used for model training but for monitoring the predictive accuracy change of multi-task CNN during the training process, in order to avoid the over-fitting problem.

The prediction accuracy might be affected by the hyper-parameter settings of the multi-task CNN. We need to determine the optimal configuration of the proposed model. The hyper-parameters primarily include the number of a combination of the convolutional and average pooling layers in both shared and task-specific structures (denoted as L_s and L_{ts}, respectively), the number and size of the convolutional kernels in each convolutional layer, and the weight of each task's loss (see Eq. (2)). This paper employs the commonly-used grid search approach to determine the optimal hyper-parameters of the multi-task CNN, because it can be easily parallelised for time-saving. Grid search means changing one of the hyperparameters while keeping the others unchanged during the tuning process. As suggested in (Zhang and Cheng, 2020), the network structure does not need to be very deep because the 2D images of activity patterns are not as complex as common images (e.g., street view images). Thus, L_s and L_{ts} is chosen from 1 to 3, each layer's number of filters is selected from [4, 8, 16] and kernel size is chosen from [(2×2), (3×3), (5×5)]. Finally, the weight of each task's loss is selected from [0.3, 0.5, 1]. To simplify the grid search process, we set L_{ts} in each specific task to be the same and the kernel size in each pooling layer is fixed to be (2×2). Additionally, all tasks employ only one fully-connected layer.

The multi-task CNN model is implemented using the GPU-version Tensorflow 1.2 (Abadi et al., 2016). In the training process, the batch size is 32 and the number of epochs is 50. The learning rate is set to be 0.01. The optimal configuration of the multi-task CNN could then be determined. The best results presented in the next section is obtained using a multi-task CNN with two shared convolutional+pooling layers, and one task-specific convolutional+pooling layers (as illustrated in Figure 5). In both shared convolutional layers, there are 8 kernels of size (5×5) and (3×3), respectively. In each task-specific convolutional layer, the number of kernels is 16. However, it is found that the optimal kernel size in each task-specific convolutional layer varies slightly. In age, income level and car ownership prediction tasks, the optimal kernel size is (3×3). Meanwhile, the kernel size for gender prediction is set to be (2×2), since empirical studies show a large kernel size makes it easily overfitting. As that smaller kernel size will capture more details of the image, it might imply that the gender prediction requires more spatio-temporal information extracted from the activity pattern than the other prediction tasks. Finally, the loss's weight of the age, gender, income and car ownership inference is set to be 1, 0.3, 0.3 and 1, respectively.

5.3.2 Prediction Accuracy and Model Comparison

In this section, the demographic prediction accuracy obtained by the proposed model is compared with many baselines. The benchmarks include Random Forests (RF) (Breiman, 2001), Support Vector Machine (SVM) (Cortes and Vapnik, 1995), XGBoost (Chen, He, Benesty, Khotilovich and Tang, 2015), because these models have been widely adopted for demographic prediction in existing literature (Zhong, Tan, Mo and Yang, 2013, Wu et al., 2019). However, these models cannot directly accept two- dimensional images. Thus, each two-dimensional image is flattened into a one-dimensional vector, and principal component analysis is utilised to extract one-dimensional feature vectors as the input of these models. Furthermore, to verify the effectiveness of the MTL structure, the proposed model is also compared with the standard CNN. Additionally, we also compare the proposed model with residual CNN (ResNet) (He, Zhang, Ren and Sun, 2016), which is a state-of-the-art algorithm for image processing.

By feeding features to all algorithms, results are obtained and depicted in Figure 9. It shows that leveraging the proposed model, the highest accuracy achieves 80% in the prediction of car ownership, followed closely by 76% of age group inference. The performance of income level ranks third, with an accuracy of 69%. The accuracy of gender prediction is about 64%, which is the worst among all prediction tasks. This implies that spatio-temporal activity patterns of frequent public transit users have a stronger association with age and car ownership than gender and household income level. Additionally, this empirical study suggests that gender prediction is the toughest prediction task among others, therefore it might require more spatio-temporal details, which probably further explains why the kernel size of gender prediction is smaller than that of the others.

Compared to different algorithms, results show that CNN-based methods perform better than traditional classifiers. This is because CNN is more powerful to extract features from the image-structured spatio-

temporal activity profiles. Among the CNN-based models, the performance of the standard CNN is comparable to the ResNet. This may be because the spatio-temporal activity pattern images are not as complex as ordinary images. Increasing the complexity of the CNN architecture cannot significantly increase the prediction accuracy. In addition, comparing the performance between CNN/ResNet and the multi-task CNN, overall, the performance can be improved using the proposed model, especially for gender and income level inference (the two hardest prediction tasks). This might be because the MTL framework could help them to take advantage of the knowledge learned from other tasks. However, the situation is opposite in car ownership inference. The accuracy of this prediction task slightly decreases after adopting MTL architecture. This is probably because the noises in other demographic inference tasks have a negative impact on car ownership inference under the MTL framework. As the accuracy decrease is not significant, summarily, jointly predicting multiple demographics can improve the performance.

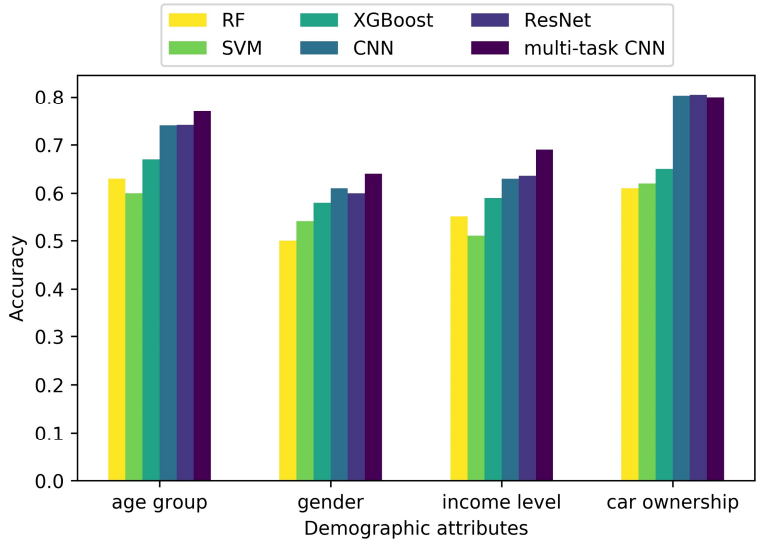


Figure 9 Prediction accuracy of different classification models.

5.4 Geodemographic Mapping Results

To obtain the geodemographics, we employ the well-trained multi-task CNN to infer the demographics of the passengers' whose Oyster card data is available but LTDS data not. The RAs detected in Section 5.1 are used to link each passenger to a specific administrative area. Note that although the postcode area is used as the geographic unit in Section 5.1 to validate the results of RA detection, we will not use this unit in this section. This is because we need to utilise the open geodemographic data to verify the geodemographic mappings. However, postcode area is not the administrative area in the UK. It means the demographic attributes at the postcode area level are not available. To verify the geodemographic mapping gained from SC data, this section uses the 'borough' as the geographic basis. The London boroughs are the 32 local authority districts that makeup Greater London. Generally, the area of a borough is larger than a postcode area. There are fewer cases that an identified RA overlaps multiple boroughs, which ensures that the accuracy of snapping RAs to boroughs will not be lower than linking to postcode areas.

5.4.1 *Geodemographic Data Post-Processing*

In this section, we use the open data provided by the Office for National Statistics (ONS)[‡] of UK in 2013 to validate the geodemographic mappings. In order to compare the real and inferred geodemographics, the predicted demographic attributes and the ONS data are post-processed using the following methods:

- **Age**: The ONS provides the borough-level population estimates by single year of age[§] in 2013. But these figures are less reliable and ONS advises that they should be aggregated to at least five-year age groupings for use in further calculations. Differently, passengers in this study are classified into three groups by age. To make them comparable, we first use the ONS age data to calculate the number of people in each group by the same grouping rule in each borough, as the ground truth. We can then compute the ratio of the young, middle-aged and elderly in each borough using the predictions and ground truth, respectively. Finally, the three ratios are used to produce the young, middle-aged and elderly distribution maps. An example of the ONS age data post-processing procedure is given in Figure 10.
- **Gender**: We use both the ONS gender data[§] in 2013 and the inferred results to compute the female/male ratio of each borough, respectively. The female to male ratio of each borough is used to compare the inferred and the real geodemographics.
- **Income** and **car ownership**: We retrieved the household income estimates[**] and the licensed car[††] data in 2013 from ONS as the ground truth. For comparison, we calculate the real and the estimated average income and average car ownership in each borough for validation.

After the post-processing, we use the rank of the boroughs to produce the geographical distribution maps of these demographic attributes.

Boroughs	Age 1	Age 2	...	Age 90+
Borough 1	10	15	...	18
Borough 2	2	8	...	20
Borough 3	20	22	...	15
Borough 4	4	6	...	14

grouping →

Boroughs	Young	Middle-age	Elderly
Borough 1	1169	1842	597
Borough 2	1107	1369	365
Borough 3	1620	1795	660
Borough 4	951	1150	468

compute ratio →

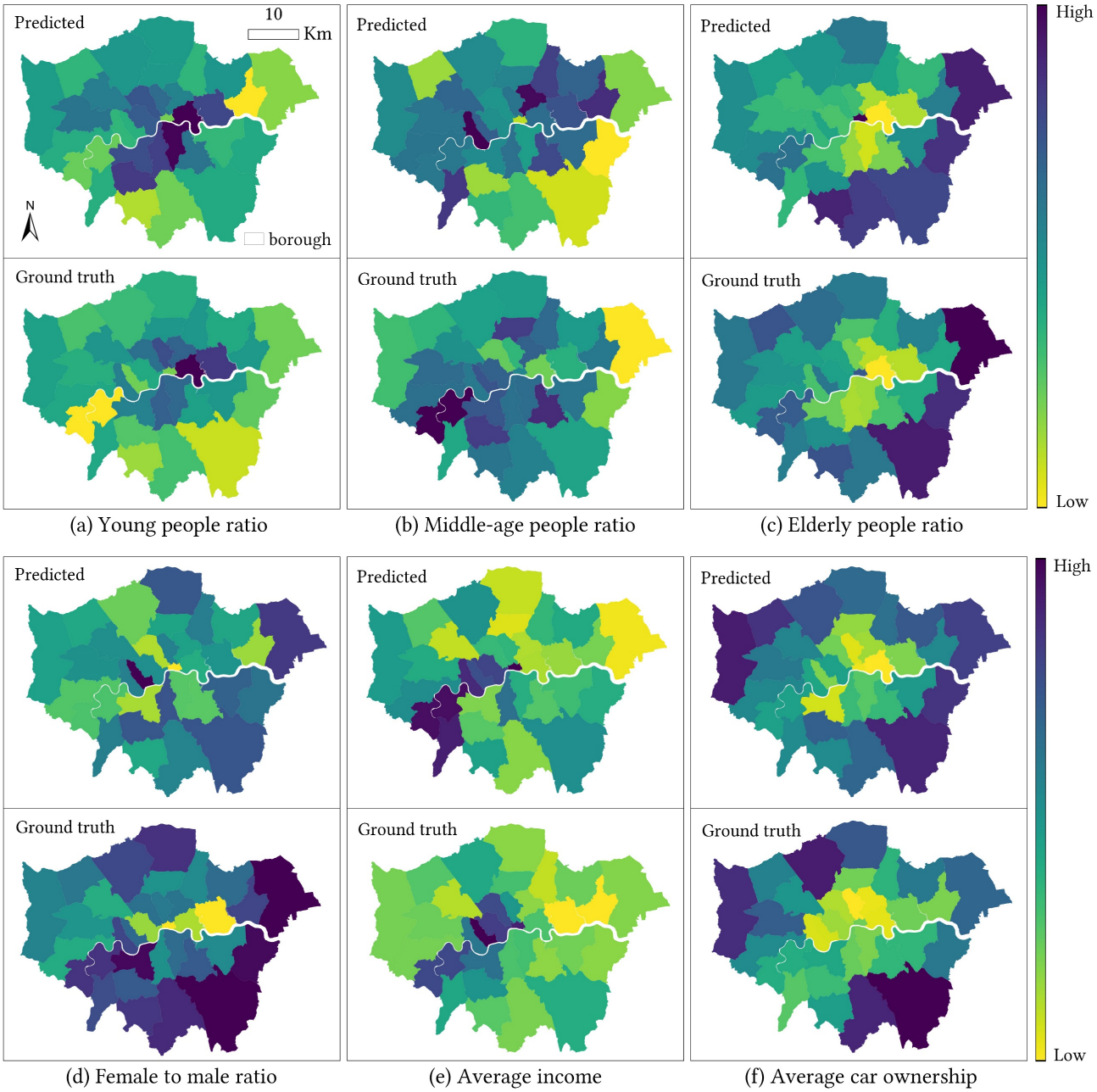
Boroughs	Young ratio	Middle-age ratio	Elderly ratio
Borough 1	0.32	0.51	0.17
Borough 2	0.39	0.48	0.13
Borough 3	0.40	0.44	0.16
Borough 4	0.37	0.45	0.18

Figure 10 An example of the ONS age data post-processing procedure.

5.4.2 *Geodemographic Visualisation, Validation and Analysis*

Figure 11 displays the comparison of the geodemographic maps produced by using the inferred demographics and the ONS data. In each subfigure in Figure 11, the upper plot is the predicted geodemographic map, and the lower plot is the ground truth. The darkness of colour is proportional to the value of a borough in terms of a specific demographic attribute. The darker colour implies a higher value of a borough.

[‡] In this paper, all sources from ONS are used under the terms of the Open Government License (OGL) and UK Government Licensing Framework.
[§] Retrieved from: https://data.london.gov.uk/dataset/office-national-statistics-ons-population-estimates-borough.
[**] Retrieved from: https://data.london.gov.uk/dataset/household-income-estimates-small-areas.
[††] Retrieved from: https://data.london.gov.uk/dataset/licensed-vehicles-type-0.



(a) Young people ratio (b) Middle-age people ratio (c) Elderly people ratio

(d) Female to male ratio (e) Average income (f) Average car ownership

Figure 11 Geographic distribution of multiple demographic statistics.

Overall, the predicted geodemographic pattern is approximately consistent with the ground truth. For example, both maps in the first subplot suggest that the ratio of young people in inner London is higher than in outer London, which is contrary to the distribution of high-elderly-ratio boroughs. Additionally, boroughs of a higher middle-age ratio exhibit a distribution running on the southeast/northwest diagonal. Moreover, it can be observed in Figure 11 (e) that the boroughs in southwest London have a higher average household income than other areas. Furthermore, Figure 11 (f) shows that the households in outer London averagely

own more private cars than those in inner London. However, the two maps in Figure 11 (d) do not match well as others. This is because the prediction accuracy of gender is the lowest among all tasks.

To quantitatively measure the correlation between the predicted geodemographics and the ground truth, we employ the Pearson correlation coefficient ρ, which is a number between -1 and 1 that indicates the extent to which two variables are linearly related. Evans (1996) has suggested using the following rule of thumb to interpret the coefficient: very week correlation ($|\rho| < 0.2$), week correlation ($0.2 \leq |\rho| < 0.4$), moderate correlation ($0.4 \leq |\rho| < 0.6$), strong correlation ($0.6 \leq |\rho| < 0.8$) and very strong correlation ($|\rho| \geq 0.8$). The p-value indicates whether the testing result is significant. Table 4 presents the results of the Pearson correlation measurement. It shows that the predicted maps of young people ratio, the elderly ratio, average income and average car ownership have a significantly strong linear correlation with the corresponding true maps. Among the three age groups, the correlation of the middle-aged group is lower than the others. This might be because empirically the middle-aged have more travel mode choices than the youth or the elderly. Consequently, the distribution of frequent middle-aged passengers may be biased in the sample. Furthermore, the correlation of the ranks by female to male ratio is the weakest due to the poor prediction accuracy. To further validate the effectiveness of the proposed model, we also report the Pearson coefficients between the estimated geodemographics produced by using ResNet and the ground truth. As ResNet is the best predictive model among all benchmarks, we do not present other baselines' Pearson coefficients for conciseness. Table 4 shows the MTL framework can produce better geodemographic maps due to its higher demographic prediction accuracy. The results in Table 4 are consistent with the visualisation presented in Figure 9.

Table 4 Pearson correlation coefficients between the real and the estimated geodemographic attributes produced by multi-task CNN and ResNet, respectively

Geodemographic	Multi-task CNN		ResNet	
	Pearson coefficient	p-value	Pearson coefficient	p-value
Young people ratio	0.707246	3.161e-06	0.672575	1.817e-05
Middle-age people ratio	0.541675	9.974e-04	0.488839	3.893e-03
Elderly people ratio	0.754120	3.262e-07	0.720923	2.023e-06
Female to male ratio	0.411758	1.048e-02	0.373647	3.21 e-02
Average income	0.600297	2.224e-04	0.56146	3.162e-05
Average car ownership	0.651941	2.945e-05	0.659877	1.845e-05

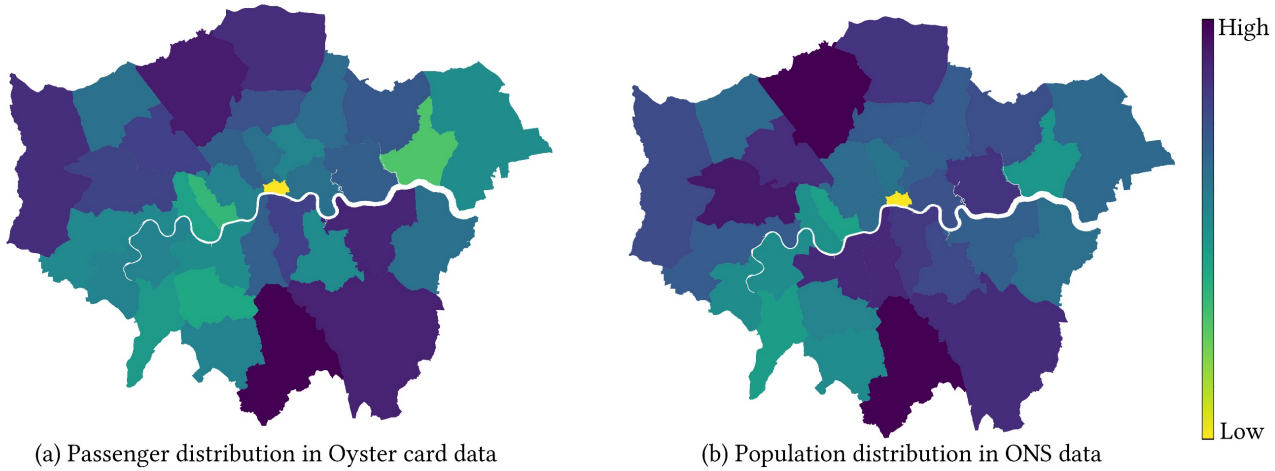
This experiment suggests that the SC data is a potential data source to generate the geographical mappings of some demographic attributes. If a demographic trait can be inferred from SC data with high accuracy, a satisfying geodemographic map can be made leveraging residential area detection. Additionally, although the resolution of activities embedded in the smart card data is much lower than that in GPS trajectories, the

demographic prediction accuracy is satisfying. Hence, for personal privacy concern, the results suggest that the entire trajectories or travel diaries should be partially hidden and protected to avoid privacy leak.

5.5 Discussion about Data Bias

A series of experiments shows that it would be a promising way to use SC data to infer the timely geodemographic information, which can be a supplement of the conventional census survey. However, for geodemographics study, the success of the proposed framework also relies on the representativeness of the SC data. For instance, according to the results in Section 5.4.2, the inferred geographic distribution of the middle-age population is inconsistent with the ground truth to some extent. The potential reason is that bias might exist in SC data as middle-age people have more travel mode choices than other age groups. Considering the fact that not all the population would take public transport for daily travel, this section provides a brief discussion about the influence of the data bias.

To measure the representativeness of the Oyster card data, we compare the geographic distribution of all passengers in the Oyster card data with the population distribution in the ONS data (as shown in Figure 12). The passenger distribution map is produced using their inferred residential areas. The Pearson coefficient of the two distributions is 0.856 (p-value= 2.17e-10), indicating a very strong linear correlation. Overall, Oyster card data are representative in terms of population distribution. However, Figure 12 shows the ratio of public transit users in Outer London is marginally lower than that in Inner London. This is probably because the services by public transport are more limited in Outer than in Inner London.



(a) Passenger distribution in Oyster card data (b) Population distribution in ONS data

Figure 12 Geographic distributions of passengers in Oyster card data and the population in ONS.

To deal with the slight data bias, resampling techniques can be employed. We use a bootstrapping approach (Mooney, Mooney, Mooney, Duval and Duvall, 1993) to resample the passengers from the total Oyster card data samples according to the population distribution in ONS data. We then reproduce the geodemographic mappings and compare the new maps with the ground truths. Results are given in Table 5. It shows the

accuracy of the geographic distribution of different demographics can be improved to varying degrees. However, the accuracy of the geodemographic mapping is still mainly limited by demographic prediction accuracy. Note that the resampling techniques can be used to alleviate the bias issue only when the residential areas of passengers can be inferred with high accuracy and the real population distribution is available. If the real population distribution is unknown, there is no easy way to tackle the data bias issue.

In summary, the representativeness of Oyster card data is satisfying for geodemographic study. Alleviating the data bias issue can help further increase the estimated results. We understand that we may miss out those who never use public transport, but our finding has shown the high potential of using SC for geodemographic mapping, which has the advantages over conventional travel survey or census. Even if the bias exists in the geodemographic maps produced based on SC data, the results are still useful and meaningful in many real-world applications, such as optimising transport planning (Liu and Cheng, 2020) and improving the delivery of regional public transport services (Zhang and Cheng, 2020).

Table 5 Pearson correlation coefficients between the ground truth and the estimated geodemographics produced after bootstrapping

Geodemographic	Pearson coefficient	p-value
Young people ratio	0.726955	1.658313e-06
Middle-age people ratio	0.566292	5.918843e-04
Elderly people ratio	0.770002	1.614154e-07
Female to male ratio	0.462407	6.740838e-03
Average income	0.676857	1.524531e-05
Average car ownership	0.674038	1.705585e-05

6 Conclusion

The contribution of this study can be summarised into two aspects. The first contribution resides in the proposed geodemographic inference framework using SC data. In this framework, we propose a novel residential area detection approach. In addition, by representing the raw SC data into 2D images, the spatio-temporal activity patterns can be revealed. A multi-task CNN is then utilised to infer multiple demographic characteristics simultaneously to improve the performance of demographic inference. Finally, the inferred home location and demographics are leveraged to produce geodemographic maps. The second contribution emerges from the application of the framework using large-scale Oyster card data in Greater London, UK. Results validate the effectiveness of the residential area identification approach. In addition, the case study shows that car ownership and age group can be inferred with high accuracy from SC data. On the contrary, the prediction performance of gender and household income level is relatively low. This phenomenon indicates that the predictive power of spatio-temporal activity patterns may have inclinations to some types

of demographic attributes. Finally, by comparing the estimated geodemographic maps with the ground truth, it is concluded that SC data is potential to generate the geographical distribution of some demographics. If a demographic attribute can be inferred from SC data with high accuracy, a satisfying geodemographic map can be made. Also, it would be a more efficient and timely way to infer the geodemographic information using the SC data, as a supplement of the conventional census survey (every 10 years in UK). Furthermore, this study suggests the importance of geotagged data protection for privacy concerns.

However, there is still room for improvement. Future work can be conducted based on the work presented herein. For example, the current SC data representation method only captures the spatial and temporal information of activities. The semantic interpretations of these activities are ignored in current work. Future research can combine multiple data sources, such as land use and POI data, to add semantics of activities, which might be able to further improve the prediction accuracy. In addition, the demographic inference model is a supervised learning approach, which requires sufficient demographic information as true labels for training. However, the demographic data obtained via survey is usually limited. Future work can consider leveraging semi-supervised learning, which uses a small amount of labelled data with a large amount of unlabelled data for model training, to further improve the prediction accuracy in large-scale application scenarios.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G. & Isard, M. TensorFlow: A System for Large-Scale Machine Learning. 12th Symposium on Operating Systems Design and Implementation (OSDI), 2016. 265-283.

Bagchi, M. & White, P. R. 2005. The potential of public transport smart card data. *Transport Policy,* 12**,** 464-474.

Bantis, T. & Haworth, J. 2017. Who you are is how you travel: A framework for transportation mode detection using individual and environmental characteristics. *Transportation Research Part C: Emerging Technologies,* 80**,** 286-309.

Breiman, L. 2001. Random forests. *Machine learning,* 45**,** 5-32.

Carmel, D., Lewin-Eytan, L., Libov, A., Maarek, Y. & Raviv, A. The demographics of mail search and their application to query suggestion. Proceedings of the 26th International Conference on World Wide Web, 2017. International World Wide Web Conferences Steering Committee, 1541-1549.

Chen, T., He, T., Benesty, M., Khotilovich, V. & Tang, Y. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2***,** 1-4.

Cortes, C. & Vapnik, V. 1995. Support-vector networks. *Machine learning,* 20**,** 273-297.

Ding, S., Huang, H., Zhao, T. & Fu, X. Estimating Socioeconomic Status via Temporal-Spatial Mobility Analysis-A Case Study of Smart Card Data. 2019 28th International Conference on Computer Communication and Networks (ICCCN), 2019. IEEE, 1-9.

Dong, Y., Chawla, N. V., Tang, J., Yang, Y. & Yang, Y. 2017. User Modeling on Demographic Attributes in Big Mobile Social Networks. *ACM Transactions on Information Systems (TOIS),* 35**,** 35.

Evans, J. D. 1996. *Straightforward statistics for the behavioral sciences*, Thomson Brooks/Cole Publishing Co.

Gao, J., Zhang, Y.-C. & Zhou, T. 2019. Computational socioeconomics. *Physics Reports*.

Ghosh, S. & Ghosh, S. K. Modeling of Human Movement Behavioral Knowledge from GPS Traces for Categorizing Mobile Users. Proceedings of the 26th International Conference on World Wide Web Companion, 2017. International World Wide Web Conferences Steering Committee, 51-58.

Gordon, J. B., Koutsopoulos, H. N., Wilson, N. H. & Attanucci, J. P. 2013. Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. *Transportation research record,* 2343**,** 17-24.

Goulet Langlois, G., Koutsopoulos, H. N. & Zhao, J. 2016. Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies,* 64**,** 1-16.

He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. 770-778.

Hu, J., Zeng, H.-J., Li, H., Niu, C. & Chen, Z. Demographic prediction based on user's browsing behavior. Proceedings of the 16th international conference on World Wide Web, 2007. ACM, 151-160.

Ilägcrstrand, T. What about people in regional science? Papers of the Regional Science Association, 1970.

Kingma, D. P. & Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lecun, Y., Bengio, Y. & Hinton, G. 2015. Deep learning. *nature,* 521**,** 436.

Li, G., Yu, L., Ng, W. S., Wu, W. & Goh, S. T. Predicting home and work locations using public transport smart card data by spectral analysis. 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 2015. IEEE, 2788-2793.

Liu, Y. & Cheng, T. 2018. Understanding public transit patterns with open geodemographics to facilitate public transport planning. *Transportmetrica A: Transport Science***,** 1-28.

Liu, Y. & Cheng, T. 2020. Understanding public transit patterns with open geodemographics to facilitate public transport planning. *Transportmetrica A: Transport Science,* 16**,** 76-103.

Long, Y. & Thill, J.-C. 2015. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems,* 53**,** 19-35.

Ma, X., Wu, Y.-J., Wang, Y., Chen, F. & Liu, J. 2013. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies,* 36**,** 1-12.

Martin, D., Gale, C., Cockings, S. & Harfoot, A. 2018. Origin-destination geodemographics for analysis of travel to work flows. *Computers, Environment and Urban Systems,* 67**,** 68-79.

Mohamed, K., Côme, E., Oukhellou, L. & Verleysen, M. 2017. Clustering smart card data for urban mobility analysis. *IEEE Transactions on Intelligent Transportation Systems,* 18**,** 712-728.

Mooney, C. F., Mooney, C. L., Mooney, C. Z., Duval, R. D. & Duvall, R. 1993. *Bootstrapping: A nonparametric approach to statistical inference*, sage.

Perozzi, B. & Skiena, S. Exact age prediction in social networks. Proceedings of the 24th International Conference on World Wide Web, 2015. ACM, 91-92.

Riederer, C. J., Zimmeck, S., Phanord, C., Chaintreau, A. & Bellovin, S. M. I don't have a photograph, but you can have my footprints.: Revealing the Demographics of Location Data. Proceedings of the 2015 ACM on Conference on Online Social Networks, 2015. ACM, 185-195.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Sari Aslam, N., Cheng, T. & Cheshire, J. 2019. A high-precision heuristic model to detect home and work locations from smart card data. *Geo-spatial Information Science,* 22**,** 1-11.

Sari Aslam, N., Cheng, T., Cheshire, J. & Zhang, Y. Trip purpose identification using pairwise constraints based semi-supervised clustering. The 27th Conference on Geographical Information Science Research UK, 2019 Newcastle, UK. 97-102.

Shen, J. & Cheng, T. 2016. A framework for identifying activity groups from individual space-time profiles. *International journal of geographical information science,* 30**,** 1785-1805.

Singleton, A. D. & Spielman, S. E. 2014. The past, present, and future of geodemographic research in the United States and United Kingdom. *The Professional Geographer,* 66**,** 558-567.

Siren, A. & Hakamies-Blomqvist, L. 2004. Private car as the grand equaliser? Demographic factors and mobility in Finnish men and women aged 65+. *Transportation Research Part F: Traffic Psychology and Behaviour,* 7**,** 107-118.

Tfl 2013. Travel in London Report 7.

Van Den Berg, P., Arentze, T. & Timmermans, H. 2013. A path analysis of social networks, telecommunication and social activity–travel patterns. *Transportation Research Part C: Emerging Technologies,* 26**,** 256-268.

Vijayaraghavan, P., Vosoughi, S. & Roy, D. Twitter Demographic Classification Using Deep Multi-modal Multi-task Learning. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2017. 478-483.

Volkova, S., Bachrach, Y. & Van Durme, B. Mining user interests to predict perceived psycho-demographic traits on Twitter. Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on, 2016. IEEE, 36-43.

Wang, P., Guo, J., Lan, Y., Xu, J. & Cheng, X. 2016. Multi-task Representation Learning for Demographic Prediction. *In:* FERRO, N., CRESTANI, F., MOENS, M.-F., MOTHE, J., SILVESTRI, F., DI NUNZIO, G. M., HAUFF, C. & SILVELLO, G. (eds.) *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings.* Cham: Springer International Publishing.

Wang, P., Sun, F., Wang, D., Tao, J., Guan, X. & Bifet, A. Inferring Demographics and Social Networks of Mobile Device Users on Campus From AP-Trajectories. Proceedings of the 26th International

Conference on World Wide Web Companion, 2017. International World Wide Web Conferences Steering Committee, 139-147.

Wang, Y., De Almeida Correia, G. H., De Romph, E. & Timmermans, H. 2017. Using metro smart card data to model location choice of after-work activities: An application to Shanghai. *Journal of Transport Geography,* 63**,** 40-47.

Wu, L., Yang, L., Huang, Z., Wang, Y., Chai, Y., Peng, X. & Liu, Y. 2019. Inferring demographics from human trajectories and geographical context. *Computers, Environment and Urban Systems,* 77**,** 101368.

Zhang, J., Zheng, Y., Sun, J. & Qi, D. 2019. Flow Prediction in Spatio-Temporal Networks Based on Multitask Deep Learning. *IEEE Transactions on Knowledge and Data Engineering***,** 1-1.

Zhang, Y. & Cheng, T. Inferring Social-Demographics of Travellers based on Smart Card Data. 2nd International Conference on Advanced Research Methods and Analytics, 2018 Valencia, Spain. Editorial Universitat Politècnica de València, 55-62.

Zhang, Y. & Cheng, T. 2020. A Deep Learning Approach to Infer Employment Status of Passengers by Using Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems,* 21**,** 617-629.

Zhang, Y., Cheng, T. & Aslam, N. S. Exploring the Relationship Between Travel Pattern and Social-Demographics Using Smart Card Data and Household Survey. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2019 Netherland. Copernicus Publications, 1375-1382.

Zhang, Y., Cheng, T. & Ren, Y. 2019. A graph deep learning method for short-term traffic forecasting on large road networks. *Computer-Aided Civil and Infrastructure Engineering,* 34**,** 877–896.

Zhang, Y., Cheng, T., Ren, Y. & Xie, K. 2020. A novel residual graph convolution deep learning model for short-term network-based traffic forecasting. *International Journal of Geographical Information Science,* 34**,** 969-995.

Zhang, Y., Zhang, Y. & Zhou, J. 2019. A novel excess commuting framework: Considering commuting efficiency and equity simultaneously. *Environment and Planning B: Urban Analytics and City Science***,** 2399808319851517.

Zhao, S., Pan, G., Zhao, Y., Tao, J., Chen, J., Li, S. & Wu, Z. 2017. Mining user attributes using large-scale app lists of smartphones. *IEEE Systems Journal,* 11**,** 315-323.

Zhong, E., Tan, B., Mo, K. & Yang, Q. 2013. User demographics prediction based on mobile data. *Pervasive and Mobile Computing,* 9**,** 823-837.

Zhong, Y., Yuan, N. J., Zhong, W., Zhang, F. & Xie, X. You are where you go: Inferring demographic attributes from location check-ins. Proceedings of the eighth ACM international conference on web search and data mining, 2015. ACM, 295-304.

Zhu, L., Gonder, J. & Lin, L. 2017. Prediction of individual social-demographic role based on travel behavior variability using long-term GPS data. *Journal of Advanced Transportation,* 2017.