TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Social demographics imputation based on similarity in multi-dimensional activity-travel pattern

Citation for published version (APA):

Zhang, B., Rasouli, S., & Feng, T. (2024). Social demographics imputation based on similarity in multi-dimensional activity-travel pattern: A two-step approach. *Travel Behaviour and Society*, *37*, Article 100843. https://doi.org/10.1016/j.tbs.2024.100843

Document license:

CC BY-NC-ND

DOI:

10.1016/j.tbs.2024.100843

Document status and date:

Published: 01/10/2024

Document Version:

Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 07. Jul. 2024

Contents lists available at ScienceDirect

Travel Behaviour and Society

journal homepage: www.elsevier.com/locate/tbs

Social demographics imputation based on similarity in multi-dimensional activity-travel pattern: A two-step approach

Bin Zhang [a,*], Soora Rasouli [a], Tao Feng [b]

[a] Urban Planning and Transportation Group, Eindhoven University of Technology, PO Box 513, 5600 MB, Eindhoven, Netherlands
[b] Urban and Data Science Lab, Graduate School of Advanced Science and Engineering, Hiroshima University, Higashi-Hiroshima 739-8511, Japan

ARTICLE INFO

Keywords:
Demographics imputation
Daily activity travel
Feature importance
Dynamic time warping (DTW)
Support vector machine (SVM)

ABSTRACT

In response to the absence of demographics in increasingly emerging big data sets, we propose a novel method for inferring the missing demographic information based on similarity in people's daily multi-dimensional activity-travel patterns as well as the characteristics of the area they move about. Instead of using isolated activity-travel attributes to infer social demographic features, our proposed method first calculates the similarity of people's multidimensional daily activities and travels as well as characteristics of their visiting locations, between those for whom the social demographics are to be imputed (target) and those with known demographics (base) using a polynomial function. The weights of the function are determined using the permutation feature importance method, and then dynamic time warping is used to align the multidimensional activity sequences of the base and target sample and measure their similarities. For each person in the target database, a matched list is created consisting of those with the most similar activity-travel sequences in the base sample. A support vector machine is then trained using the base sample as input to impute the demographics of the target sample. The proposed model is trained using a national travel survey and validated by applying it to a GPS dataset. The results show that the proposed method outperforms existing methods in predicting four selected demographics: gender, age, education level, and work status, with an accuracy range between 91% and 94% for the national dataset and 88% to 91% for the GPS data. This study highlights the importance of considering the multidimensional and sequential nature of peoples' daily activity-travel patterns in the imputation of demographic features.

1. Introduction

With the development of information and communication technology, a substantial amount of data related to human mobility, such as GPS data, mobile phone data, smart card data, and online geo-location data, have become available in the past few years. While these types of data have great value in understanding human mobility at both individual and aggregate levels (Barbosa et al., 2018), they often lack information related to subjects' demographic attributes. Demographic attributes have proven to play an important role in human mobility (Lenormand et al., 2015; Alessandretti et al., 2020) and travel behavior (Mouratidis et al., 2019; Acheampong et al., 2020).

Existing studies suggest that people with the same demographic attributes tend to show considerable similarities in their travel behavior (Lenormand et al., 2015; Goulet-Langlois et al., 2016; Xianyu et al., 2017; Shou and Di, 2018; Xu et al., 2020; Zhou et al., 2021). Such knowledge led to the proposal of a series of imputation methods in recent years, mainly based on discrete choice methods (Auld et al., 2015; Pawlak et al., 2015; Zhao et al., 2022) and machine learning models (Brdar et al., 2012; Wang et al., 2016; Wu et al., 2019). Although these studies included activity-travel characteristics as part of the input data for the model estimation, they did not include the multidimensional or sequential nature of activity-travel patterns. These studies generally found that, regardless of the data source or the method, demographics imputation is complicated and challenging reflected by the typical accuracy of imputation being about in the range of 50%-85%.

In view of the low accuracy of imputed demographics in the existing studies, this study attempts to improve the existing methods by first calculating the similarity in activity-travel patterns, taking into account their multidimensional and sequential nature. In addition, with the premise that built environment characteristics of the areas contribute to the arrangement of activity-travel patterns (Ewing and Cervero, 2001; Wang et al., 2011; Ding et al., 2018; Farinloye et al., 2019; Figueroa Martínez et al., 2019), we included land use mix and density of various

[*] Corresponding author.
E-mail addresses: b.zhang1@tue.nl (B. Zhang), s.rasouli@tue.nl (S. Rasouli), taofeng@hiroshima-u.ac.jp (T. Feng).

https://doi.org/10.1016/j.tbs.2024.100843
Received 8 June 2023; Received in revised form 3 June 2024; Accepted 4 June 2024
2214-367X/© 2024 The Author(s). Published by Elsevier Ltd on behalf of Hong Kong Society for Transportation Studies. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

facility types in the area where the activity takes place in the similarity measurement as well. More precisely, we define a distance metric to measure the distance (or similarity) between two activity-travel patterns characterized by multiple features. These features include departure time, arrival time, travel time, activity duration, trip mode, trip purpose, population density of visited location, bus stop density of visited location, railway station density of visited location, land use mix of visited location, shopping facility density of visited location, leisure facility density of visited location, and surrounding address density of visited location. The weight for each feature is then determined by permutation feature importance which is defined as the decrease in a model score with a single feature value randomly shuffled (Breiman, 2001). In order to account for the multi-dimensionality of activity-travel patterns, dynamic time warping (DTW) is used to calculate the (dis-)similarity in activity-travel patterns of any two persons in the base dataset (contains social demographic) and target dataset (lacks social demographic). The calculated distances are then sorted in ascending order in which the person in the base dataset with the smallest distance (to the person in the target dataset) is ranked first. For each person in the target dataset, a certain number of subjects from the top of the ranked list are selected and are placed in the match list for that person. A support vector machine (SVM) is then constructed to model each social demographic feature of the sample in the target database based on the matched subjects in the base dataset. A large-scale national mobility survey as well as a GPS dataset are used to train and validate the model. The high prediction accuracy suggests the value of the proposed approach.

The rest of the paper is organized as follows. Section 2 reviews current research on activity-travel similarity measurement and demographics imputation. Section 3 describes the details of the proposed method for measuring people's similarity and their demographic imputations. Section 4 describes the data used in the study. Section 5 reports the results. In Section 6 we draw conclusions and discuss the avenue for future research.

2. Literature review

2.1. Activity-travel and demographic association

Existing studies suggest that people with similar demographic attributes tend to exhibit considerable similarities in their mobility behavior. Lenormand et al. (2015) analyzed geolocated credit card transactions from Spain and found that people's mobility patterns vary according to their gender, age, and occupation. Goulet-Langlois et al. (2016) investigated travelers' heterogeneity based on their multi-week activity sequences derived from smart card data of London's public transport network. They found significant associations between travel patterns and demographics, such as age, occupation, household composition, income, and vehicle ownership. Xianyu et al. (2017) analyzed variability in activity-travel diaries imputed from multi-day GPS data. Considering that an activity-travel pattern is multidimensional in nature, involving time, location, mode, purpose, etc., multidimensional sequence alignment was applied to measure the similarity of activity-travel sequences. They then applied panel effect regression models to estimate the effects of demographics and weekdays on the degree of similarity. They found people's gender, age and income have strong influences on the degree of similarity in activity-travel sequences. Shou and Di (2018) proposed a framework to analyze the similarity of activity patterns using frequent sequential pattern mining. Prefix-Span algorithm was used to extract frequent patterns while similarity was defined considering both travel sequence length and frequent patterns. Based on the pairwise similarity between two people, hierarchical clustering was used to divide travelers into communities. A multinomial logistic regression model was employed to model the extent to which the clustered communities can be explained by the similarity of demographics. They found that similarities in demographics, such as employment status and number of children in the household, are closely correlated with the similarity of activity patterns. Xu et al. (2020) proposed a novel semantic-enhanced urban mobility embedding model with a representation learning method for demographics imputation. They designed a user-location network to represent users' physical mobility patterns (i. e., visited locations, visiting frequency and user's time allocation pattern) and underlying semantic information (i.e., urban region's point of interest (POI) distribution). By extracting the most representative temporal patterns and spatial distribution of the most representative locations, and applying Student's *t*-test with Bonferroni correction, they found significant correlations between people's demographics, i.e., gender, age, occupation, and their spatial and temporal visitation patterns. Zhou et al. (2021) employed a Markov-chain-based mixture model to cluster daily activities and detect recurrent patterns. Subsequently, logistic regression models were constructed to examine the postulated associations between activity patterns and socio-demographic characteristics. The study revealed strong correlations between socio-demographic attributes and the determination of daily activity scheduling.

2.2. Similarity of multi-dimensional activity-travel sequences

Joh et al. (2002) proposed a multidimensional sequence alignment method (MDSAM) based on Levenshtein distance to measure the similarity of people's daily activity-travel patterns. Considering both the multidimensional and sequential nature of activity-travel patterns, MDSAM proved to be a suitable similarity measure for classifying activity-travel patterns. Li et al. (2008) proposed a hierarchical-graph-based similarity measurement framework to mine the similarity between people based on their visited locations. A GPS dataset collected from 65 volunteers over six months was used for the measurement. A relation matrix, generated by these volunteers, was used as the benchmark for similarity measurement. The method proved to be about 10% higher in mean average precision compared with cosine and Pearson similarity measurements. Kwan et al. (2015) presented a methodology for measuring the similarity among individual activity patterns. Multidimensional sequence alignment of daily activities was conceptualized as a multi-objective optimization problem which was further solved with an evolutionary algorithm. They demonstrated the effectiveness of their proposed method by comparing it with ClustalG, a commonly used software package for sequence alignment which is a rewrite of the well-known Clustal series, using 50 car drivers' activity-travel sequences in a day. They concluded that their method outperforms the ClustalG for most of the selected cases. Joh et al. (2016) proposed a position-sensitive sequence-alignment method to measure the similarity between people's daily activities. The method was tested using the activity diary data collected in the Netherlands. Faroqi et al. (2018) introduced a model with two parallel steps for quantifying the activity similarity among public transit passengers. Space-time prism was utilized to measure the spatiotemporal similarity of two activities in a three-dimensional continuous space, and probabilistic decision tree was employed to measure the activity type similarity. The final activity similarity value was defined as the product of the spatiotemporal and activity type similarity values. With the proposed model, they found more than 81 percent of the passengers exhibited partial or complete activity similarity with their fellow passengers. Shou and Di (2018) proposed a framework to analyze the similarity of activity patterns using Prefix-Span algorithm which can discover the frequently occurring ordered subsequences. They defined a quantitative measurement of pattern similarity by considering both travel sequence length and frequent activity patterns. Hierarchical clustering was conducted to divide all travelers into three major clusters based on pairwise similarity. They found that there is a strong association between the similarity of demographics and similarity of activity patterns.

2.3. Demographics imputation

In recent years, with the generation and application of big datasets, the problem of demographic imputation has attracted increasing attention. Early studies imputed people's demographics based on their online behavior (Hu et al., 2007; Mislove et al., 2010; Kosinski et al., 2013; Wang et al., 2019; Cui and He, 2021) and mobile phone usage patterns (Dong et al., 2014, 2017; Sarraute et al., 2014; Jahani et al., 2017) because it is easier to generate big data using these data sources. These studies implemented various techniques to impute missing demographics, such as machine learning algorithms (Hu et al., 2007; Kosinski et al., 2013; Dong et al., 2014, 2017; Jahani et al., 2017; Wang et al., 2019; Cui and He, 2021), and network analysis (Mislove et al., 2010; Sarraute et al., 2014). For example, Hu et al. (2007) utilized users' web browsing behaviors in a Bayesian framework to predict their gender and age. Mislove et al. (2010) used a community detection algorithm to infer missing education information based on other users' profiles in the same online social network. Dong et al. (2014) developed a probabilistic framework based on mobile communication patterns to predict gender and age. Jahani et al. (2017) applied five different machine learning algorithms (logistic regression, SVM with linear kernel, SVM with radial basis function kernel, *k*-nearest neighbors, and random forest) to impute the missing information of gender, age and income using call detail records.

Studies using people's physical activity-travel patterns to impute their demographics are rather scarce. Methods used in the few existing studies can be classified into random utility theory/discrete choice model, and machine learning models. Auld et al. (2015) used GPS traces of 9,736 people and extracted their activity-travel patterns including the number of activities, tour-based transport mode, trip purpose, travel time and activity duration. These characteristics in an isolated form, combined with land use data (i.e., road density, intersection density, block size, employment density, population density, and housing density), were used as determinants in various discrete choice models for the demographic imputation. Work status, education level, age, license possession, and presence of children in households were among the predicted demographics. The accuracy of license possession was up to 90% while for other demographics, the prediction accuracy was generally between 55%-75%. In fact, gender, household size, and number of vehicles were proven to be more difficult to predict. Zhao et al. (2022) proposed an inverse discrete choice modelling (IDCM) approach to infer the demographics based on people's travel behavior (i.e., departure time and travel mode). They assumed that people's choice of departure time and travel mode is partly influenced by demographic variables that can be estimated from observed travel behavior patterns using the IDCM. The approach aimed to maximize the probability that an individual is characterized by a particular demographic attribute considering the observed activity-travel choices. They used each feature of the travel patterns separately as opposed to considering travel patterns as multi-dimensional sequence. They validated their approach over two empirical applications and reported accuracies in the range of 50%-90% for different demographics.

Besides the random utility theory/discrete choice model, machine learning models have also been applied to impute demographic information. One advantage of using machine learning models is that they do not require any pre-defined functional form for the relationship between demographic variables and activity-travel patterns. Brdar et al. (2012) experimented graph-based representation of people where each person was considered as a node in a graph and the relationships (edges in the graph) between people were modeled based on the cosine similarity of their activity-travel features, such as travel time and distance. The demographics were then inferred using three well-known algorithms: *k*-nearest neighbors, radial basis function network and random forest. Demographics of neighbors in the graph were then used to infer the missing demographics. Testing on a large-scale mobility dataset from Nokia mobile data challenge, they reached the accuracy of 70%-80% for gender, 50%-60% for marital status, and 30%-40% for work status. In the conclusion, they stressed the need for future improvement in feature extraction and the measurement of user similarity, particularly highlighting the importance of taking travel sequences into consideration. Zhong et al. (2015) extracted rich semantics of people's check-in on Sina Weibo[1] including time and location. These check-in semantics include the region of check-in, time bins of check-in, and related POI information. They applied SVM and LambdaMART which is a ranking algorithm to infer the demographics based on the check-in semantics. They reported accuracy of gender, age, and education level 80%-85%, while for sexual orientation and marital status, the accuracy dropped to about 50%. Wu et al. (2019) put forward a feature engineering approach that included both spatiotemporal features and semantic features in the analysis. The spatiotemporal features included the number of visited locations, radius of gyration, travel distance per trip, etc. Semantic features refer to the land use which represents the function of a place such as residence, park, hospital, school, and shopping mall. Characteristics of trajectories were not treated as multi-dimensional sequences but as independent features. Based on GPS records of 437 volunteers in 7 days, they compared the performance of three algorithms: SVM, random forest, XGBoost. The imputation accuracy of marital status appeared to be 82% while the accuracies of predicting education level, gender, and age were 74%, 66% and 43% respectively. Zhang et al. (2020) proposed a novel framework with a multi-task convolutional neural network (CNN) for demographic prediction. They converted an individual's spatial–temporal activity pattern from multi-week transit smart card data to a two-dimensional image. CNN was employed to learn features from the images for demographic imputation. The proposed model was validated with smart card data in UK. The accuracy of predicting age, gender, income level, and car ownership was 50%-80%. Xu et al. (2020) proposed a novel semantic-enhanced urban mobility embedding model where SVM was used as the classifier for demographics imputation. They included travelers' activity duration, POI distribution of visited locations, and visiting frequency of different locations as classifying indicators. Their model was validated on two real-world mobility traces and results showed that the proposed model significantly outperforms all baseline methods such as random guess and raw feature. The accuracy of their proposed model for predicting gender, education level, income level, age, and work status were 72%, 79%, 80%, 79% and 59% respectively.

Review of (low) accuracy level in the previous studies, calls for improvement of imputation methods. Previous studies have consistently demonstrated significant associations between demographics and various dimensions of activity-travels, such as travel time (Dharmowijoyo et al., 2017), trip mode (Cao et al., 2022), trip purpose (Su et al., 2020), activity duration (Dharmowijoyo et al., 2015; Garikapati et al., 2016), and so forth. The relationship between the sequential nature of activity-travels in daily life and demographics has also received significant attention and research from scholars (Jiang et al., 2012; Dharmowijoyo et al., 2015, 2017; Goulet-Langlois et al., 2016; Shou and Di, 2018; Hafezi et al., 2019; Su et al., 2020; Zhou et al., 2021). While reviewing existing studies, we realized that neither multidimensional nor sequential nature of activity-travel patterns have been incorporated in imputing the missing demographics. Research has also found that individuals with different demographic attributes exhibit distinct visitation patterns across various built environments (Wang et al., 2011; Cheng et al., 2019; Wu et al., 2021; Duan et al., 2023). Despite this, in many previous studies, the characteristics of the visited location have not been considered when imputing the demographics. Considering such a gap, we propose a novel approach integrating all the above features in calculating similarity between people. We first extract the importance of each activity-travel feature for demographics imputation, and then measure the similarity between people with DTW considering

[1] A Chinese microblogging website similar to Twitter.

multidimensional and sequential nature of their activity-travels. As a result, a matching list is created for each individual in a target dataset after which a SVM model is created for imputing each demographic variable. This paper differs from existing studies in the following ways. First, activity-travel patterns of people (as a measure of similarity) are taken with their multi-dimensional nature in the form of a linear polynomials for which the parameters are measured with permutation feature importance. Second, DTW is used in this study for aligning (the multi-dimensional sequences) for the purpose of imputing demographics for the first time. Third, unlike existing studies that directly input activity-related attributes as features into the model, we first calculate the similarity of people based on their activity-travel sequences, and then use the demographic attributes of similar people as input of the imputation model.

3. Methodology

3.1. Measurement of activity-travel similarity

A trip is typically made to engage in an activity. Each activity and associated trip can be defined by multiple features such as departure time, travel time, transport mode, trip purpose (type of activity for which the trip is made), activity location and its characteristics. If a person has multiple trips and activities per day, daily activity-travel patterns can then be viewed as a multidimensional sequence.

If $\boldsymbol{a}_i$ and $\boldsymbol{b}_j$ represent the *i*-th and the *j*-th activity (and the associated trip) of person a and b respectively, then $a_{i,f}$ and $b_{j,f}$ are feature (attribute) f of the activities. We consider 13 features to characterize each activity and the associated trip, i.e., departure time $a_{i,1}$, arrival time $a_{i,2}$, travel time $a_{i,3}$, activity duration $a_{i,4}$, population density of visited location $a_{i,5}$, bus stop density of visited location $a_{i,6}$, railway station density of visited location $a_{i,7}$, land use mix of visited location $a_{i,8}$, shopping facility density of visited location $a_{i,9}$, leisure facility density of visited location $a_{i,10}$, surrounding address density of visited locations $a_{i,11}$, trip mode $a_{i,12}$, and trip purpose $a_{i,13}$. All density data are calculated at the PC4[2] level in which the visited location is located. Seven trip modes are distinguished, i.e., car, bike, walk, bus/tram/metro, train, motorbike, and others; and eight trip purposes: home, work, shopping, social/leisure/sports, study, pick up/drop off people, services, and others. Each activity and associated trip is therefore a 13-dimensional vector, $\boldsymbol{a}_i = [a_{i,1}, a_{i,2}, \cdots, a_{i,13}]$. All features of the activity-travel patterns, except for trip mode and trip purpose, are normalized with the min–max normalization. The distance $dis(\boldsymbol{a}_i, \boldsymbol{b}_j)$ between two activity-travel patterns for social demographic feature c is then defined as:

$$\underset{c}{dis}(\boldsymbol{a}_i, \boldsymbol{b}_j) = \left(\sum_{f=1}^{11} \left(\beta_{c,f} {}^* (a_{i,f} - b_{j,f})\right)^2 + \left(\beta_{c,mode} {}^* d_{mode}\right)^2 + \left(\beta_{c,purpose} {}^* d_{purpose}\right)^2 \right)^{1/2} \tag{1}$$

where $\beta_{c,f}$ $(f = 1, \cdots, 11)$ is the importance of feature f in defining the distance (dissimilarity) between two people's activity-travel patterns when it comes to the imputation of the demographic characteristic c. $\beta_{c,mode}$ is the feature importance of trip mode and $\beta_{c,purpose}$ is the feature importance of trip purpose. d_{mode} measures whether the trip modes of two compared trips are the same. d_{mode} equals 0 when the two modes are the same and 1 otherwise. $d_{purpose}$ measures whether the trip purposes of the two compared trips are the same. $d_{purpose}$ equals 0 in case of the same trip purpose and 1 otherwise.

We impute four demographics, i.e., gender, age, education level and work status. $\beta_{c,f}$ are calculated using the permutation feature importance method which is defined as the decrease in the model score when a single feature value is randomly shuffled (Breiman, 2001). Random forest with 100 decision trees using Gini impurity for measuring the quality of split is used as the model and accuracy as the score. Accuracy is defined as the ratio $(tp + tn)/(tp + tn + fp + fn)$ where tp is the number of true positives, tn is the number of true negatives, fp is the number of false positives, and fn is the number of false negatives. In developing the decision trees, maximum tree depth is set to 5, minimum number of cases in a parent node is set to be 100 while minimum number of cases in a child node is 50. Input in the feature importance analysis is the characteristics of the activity-travel sequence $\boldsymbol{a}_i = [a_{i,1}, a_{i,2}, \cdots, a_{i,13}]$, and corresponding output is the demographic feature of interest. Each feature is randomly shuffled 10 times resulting in the coefficient $\beta_{c,f}$ to be defined as:

$$\beta_{c,f} = s_c - \frac{1}{10} \sum_{k=1}^{10} s_{c,k,f} \tag{2}$$

where s_c is the accuracy of the trained random forest model for demographics c, and $s_{c,k,f}$ is the accuracy of the trained random forest model for demographics c on randomly shuffled data of feature f. Trip mode and trip purpose are categorical variables, and all other features are continuous variables. One-hot encoding is used to encode categorical variables. Taking the trip mode (seven categories: car, bike, walk, bus/tram/metro, train, motorbike, other) as an example, the permutation feature importance will return seven importance for the mode, the weighted average is then used as the weight of trip mode $\beta_{c,mode}$. The same strategy applies to the weight of the trip purpose $\beta_{c,purpose}$.

Each activity is a vector of features, and all consecutive activities in the day can then be seen as a multidimensional activity sequence. Since the number of daily activities of various people might differ, the compared vectors of activity sequences may have different lengths. Thus, dynamic time warping (DTW) was used to align the two multidimensional activity sequences with different lengths. DTW achieves sequence alignment based on the minimum cumulative distance by warping. The cumulative distance generated during the alignment is regarded as the measure of people's activity-travel (dis-)similarity, that is, the smaller the cumulative distance, the greater the similarity.

Given two sequences $\boldsymbol{A} = [\boldsymbol{a}_i] (i = 1, \cdots, m)$ and $\boldsymbol{B} = [\boldsymbol{b}_j] (j = 1, \cdots, n)$ with lengths m and n, the distance between the two vectors $\boldsymbol{a}_i$ and $\boldsymbol{b}_j$ is $dis(\boldsymbol{a}_i, \boldsymbol{b}_j)$. We create a cumulative distance matrix $\boldsymbol{D} \in \mathbb{R}^{m \times n}$ and initialize it with $D_{1,1} = dis(\boldsymbol{a}_1, \boldsymbol{b}_1)$, $D_{i,1} = dis(\boldsymbol{a}_i, \boldsymbol{b}_1) + D_{i-1,1}$ $(i = 2, \cdots, m)$ and $D_{1,j} = dis(\boldsymbol{a}_1, \boldsymbol{b}_j) + D_{1,j-1}$ $(j = 2, \cdots, n)$. $D_{i,j}$ is the element in the cumulative distance matrix $\boldsymbol{D}$ and stands for the minimum cumulative distance when aligning the two sequences from $(\boldsymbol{a}_1, \boldsymbol{b}_1)$ to $(\boldsymbol{a}_i, \boldsymbol{b}_j)$. For other elements $D_{i,j}$ $(i = 2, \cdots, m;\ j = 2, \cdots, n)$ in the cumulative distance matrix $\boldsymbol{D}$:

$$D_{i,j} = dis(\boldsymbol{a}_i, \boldsymbol{b}_j) + \min \begin{cases} D_{i-1,j} \\ D_{i,j-1} \\ D_{i-1,j-1} \times 2 \end{cases} \tag{3}$$

The coefficient of 2 for the $D_{i-1,j-1}$ is to avoid the warping path towards the diagonal of the cumulative distance matrix $\boldsymbol{D}$ (Giorgino, 2009).

[2] PC4 is 4-digit postcode and 4076 of them existed in the Netherlands in 2013.

For each person whose demographics are unknown (target), based on the daily activity sequence in a day and the method presented above, the (dis-)similarity in activity-travel patterns in terms of the associated features with any other people whose demographics are known (base) is calculated. All calculated distances are sorted in ascending order, and the person in the base sample corresponding to the smallest distance is considered as the most similar one to the person in the target sample whose demographics are to be imputed. In the next step, a matching list is created for each person in the target sample. For certain demographics c, the matching list is considered as $[p_0^c, p_1^c, \ldots, p_s^c]$, where p_0^c is the person in the target sample whose demographic feature c is unknown, and p_1^c to p_s^c are the top s most similar people to p_0^c in the base sample in ascending order of distance. The weight of an activity-travel feature ($\beta_{c,f}$ described in Equations (1) and (2)) may be different for various demographic types, so the matching list for a p_0 may be different when considering various demographics. Considering the day-to-day variability in activity-travel patterns (Buliung et al., 2008; Kang and Scott, 2010; Raux et al., 2016; Deschaintres et al., 2022), as well as the nature of the data available for this study (detailed in Section 4.2: activities were recorded for respondents for only one day), we have chosen to use day of the week as an additional matching item, so the comparison of activity-travel patterns is done for people on the same observed day.

3.2. Imputation of demographics

Based on the matched list, the demographics of each person in the target dataset are imputed by applying a support vector machine (SVM). SVM is a supervised learning method that can be used for classification, regression, and outlier detection. SVM is a robust and efficient prediction method and is specifically effective in cases where the number of dimensions is greater than the number of samples (Pedregosa et al., 2011).

For the binary classification, SVM maps the samples into a high dimensional space, and finds a hyperplane to maximize the gap between the two categories. A hyperplane is a $(d-1)$ dimensional subspace in an d dimensional ambient space that divides the space into two disconnected parts. For other multi-class classifications, we can construct a total of $n_{classes}*(n_{classes}-1)/2$ binary classifiers among which each one trains data from two classes and the final decision is made by aggregating the decisions of these classifiers using a voting scheme.

Given a set of training instance-label pairs $(\boldsymbol{x}_u, y_u)$ where $u = 1,\cdots,l$, and l is the number of training instance-label pairs. $\boldsymbol{x}_u \in \mathbb{R}^d$ is a d dimensional feature vector (instance) and $y_u \in \{1,-1\}$ is a one-dimensional label where 1 and -1 represent the two categories. The goal of SVM is to find the weight vector $\boldsymbol{w} \in \mathbb{R}^d$ and bias term $b \in \mathbb{R}$ such that the prediction given by $\boldsymbol{w}^T\phi(\boldsymbol{x}_u)+b$ for most samples is correct, that is $y_u(\boldsymbol{w}^T\phi(\boldsymbol{x}_u)+b) \geq 1$ for most samples. $\phi(\boldsymbol{x}_u)$ is the kernel function which can map $\boldsymbol{x}_u$ to the high dimensional space. Basic kernels include linear kernel, polynomial kernel, radial basis function kernel, and sigmoid kernel. SVM then maximizes the margin $\frac{2}{\|\boldsymbol{w}\|}$ between correctly classified samples (equivalent to minimize $\frac{1}{2}\|\boldsymbol{w}\|^2 = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$) and minimize the sum of penalty terms for misclassified samples as shown in Equation (4) (Hsu et al., 2003; Chang and Lin, 2011):

$$\min_{\boldsymbol{w},b,\xi}\left(\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{u=1}^{l}\xi_u\right) \tag{4}$$

subject to:

$$y_u\left(\boldsymbol{w}^T\phi(\boldsymbol{x}_u) + b\right) \geqslant 1 - \xi_u \tag{5}$$

$$\xi_u \geqslant 0, u = 1, .., l \tag{6}$$

where ξ_u is the allowed distance from correct margin boundary of some samples when they are misclassified or within the margin boundaries, and C is the penalty for such imperfect classification.

Due to the potential high dimensionality of the weight vector $\boldsymbol{w}$ in SVM, it is computationally demanding to solve the above primal problem directly. In order to address this issue, a dual problem is often considered as an alternative approach, as shown in Equation (7):

$$\min_{\alpha}\left(\frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{Q}\boldsymbol{\alpha} - \boldsymbol{e}^T\boldsymbol{\alpha}\right) \tag{7}$$

subject to:

$$\boldsymbol{y}^T\boldsymbol{\alpha} = 0 \tag{8}$$

$$0 \leqslant \alpha_u \leqslant C, u = 1, .., l \tag{9}$$

where $\boldsymbol{e} = [1, \cdots, 1]^T$ is the vector of all ones, and $\boldsymbol{Q}$ is an $l \times l$ positive semidefinite matrix. Such a matrix $\boldsymbol{Q}_{u,v} \equiv y_u y_v K(\boldsymbol{x}_u, \boldsymbol{x}_v)$, where $K(\boldsymbol{x}_u, \boldsymbol{x}_v) = \phi(\boldsymbol{x}_u)^T\phi(\boldsymbol{x}_v)$ is the kernel. The dual coefficient α_u is upper bounded by penalty C. By solving this easier dual optimization problem, $\boldsymbol{w}$ and b will be found and the hyperplane is determined.

As in our study, the ultimate goal is to impute the social demographic c of person p_0 given the matched list created for this person, the matched list is passed to SVM where the d dimensional vectors $\boldsymbol{x}_{uc} = \left[p_{1,c}, \cdots, p_{d,c}\right]$ is the social demographic c of these d $(d \leq s)$ matched people and used as input in SVM. The corresponding output, inferred social demographic c for p_0, is $y_{uc} = p_{0,c}$. As the number of matched people d $(d \leq s)$ may affect the imputation accuracy of the SVM model, the optimum number of d needs to be determined when training the model.

The training and testing of the imputation model is based on two datasets that contain activity-travel and demographic information: the Onderzoek Verplaatsingen in Nederland (OViN, (CBS, 2017)) dataset and the GPS dataset collected in Rotterdam, the Netherlands. Detailed information on these two datasets is provided in Section 4. In order to train the models, we use the full data from OViN dataset to obtain feature importance $\beta_{c,f}$. Then 20% of the people in OViN dataset (6,675) is randomly selected and used as the target sample, while the remaining 80% of OViN sample is used as the base. The dissimilarity between the activity-travel patterns of each individual in the target sample with all people in the base sample is calculated using $\beta_{c,f}$. A matched list of different lengths d is generated for each demographic information c (for each person in the target dataset). The matched list is used to train the SVM model. Different lengths of the matched lists are tested and the optimum length for training SVM for each demographic information c is determined. In the model application stage, people in the GPS dataset are used as the target group while all people in OViN dataset are considered as the base group. The trained SVM model is then used to impute the demographics of people in GPS dataset. A framework describing the imputation process is shown in Fig. 1.

Multidimensional activity-travel sequences during one day are considered for the calculation of similarity. If multiple days trajectories are available for one person, the data for each day will be matched separately and the SVM model is applied for each day. Eventually, the mode of the inferred demographics in multiple days will be regarded as the imputed demographics for the person.

When training the SVM model, the radial basis function (RBF) kernel $\exp(-\gamma\|x-x'\|^2)$ where $\gamma > 0$ is used to map the input x to the high dimensional space. Such an RBF kernel requires no prior knowledge about the data. Then two parameters of penalty C (described in Equations (4) and (9)) and γ in the RBF kernel need to be determined. A proper choice of C and γ is critical to the SVM's performance. The selection of parameter C is a trade-off between the misclassification of training examples against the simplicity of the classification hyperplane. A higher C aims at classifying more training examples correctly, while a lower C makes the decision surface smoother. The parameter γ defines how much influence a single training instance has, with smaller values

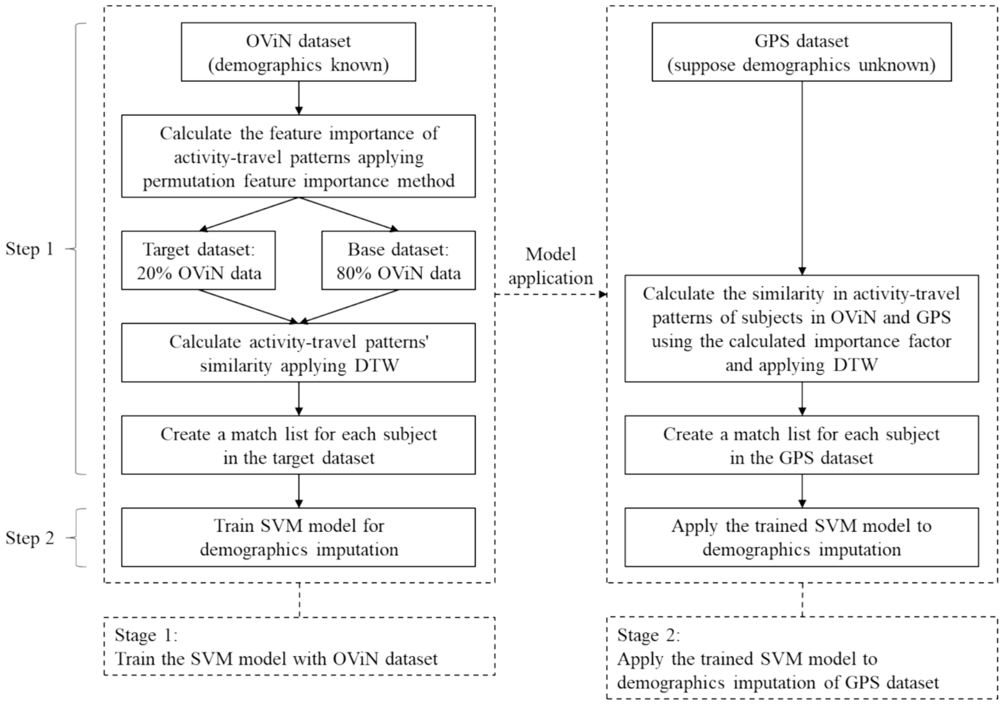
Step 1
Step 2
OViN dataset
(demographics known)
Calculate the feature importance of activity-travel patterns applying permutation feature importance method
Target dataset:
20% OViN data
Base dataset:
80% OViN data
Calculate activity-travel patterns' similarity applying DTW
Create a match list for each subject in the target dataset
Train SVM model for demographics imputation
Stage 1:
Train the SVM model with OViN dataset
Model application
GPS dataset
(suppose demographics unknown)
Calculate the similarity in activity-travel patterns of subjects in OViN and GPS using the calculated importance factor and applying DTW
Create a match list for each subject in the GPS dataset
Apply the trained SVM model to demographics imputation
Stage 2:
Apply the trained SVM model to demographics imputation of GPS dataset

Fig. 1. Methodological framework of demographics imputation.

indicating a wider influence and larger values indicating a narrower influence (Pedregosa et al., 2011). The optimum values of these two parameters can be obtained using a cross-validation grid search process.

4. Data

4.1. *GPS data*

The GPS dataset was collected in 2013 in the Rotterdam region, Netherlands. Respondents were randomly recruited and invited to take part in the study of their activity-travel patterns for three consecutive months. GPS loggers recorded time, longitudinal and latitudinal coordinates every three seconds. The respondents were required to upload the GPS traces to the server, after which an imputation algorithm was used to extract travel purposes and transportation modes (Feng and Timmermans, 2014). Respondents were then asked to check and confirm or correct the imputed aspects. 86,733 activities (27,698 location related) from 175 respondents are used for the current research. The demographic information of these respondents are summarized in Table 1.

4.2. *OViN data*

Onderzoek Verplaatsingen in Nederland (OViN) is an annual survey conducted from 2010 to 2017 (CBS, 2017) by the Centraal Bureau voor de Statistiek (CBS) aiming at providing adequate information about the daily mobility of the Dutch population. The respondents reported their activities and trips on a predetermined day of the year, as well as their social demographics. These trips are described by reporting the origin, destination, departure time, arrival time, trip mode and trip purpose. In the survey, a mixed approach was used. Respondents were asked by letter to complete the OViN questionnaire via the Internet. If people were unwilling or unable to respond via the Internet, they were contacted by telephone if the telephone number was available. In case the telephone number was not available, the questionnaire was administered at the respondent's home.

Table 1
Demographics of GPS dataset.

Variable	Description
Gender	Male: 86
	Female: 89
Age	<18: 0
	18–24: 10
	25–44: 53
	45–64: 88
	>=65: 24
Education level	Low: 3
	Medium: 133
	High: 35
	Unknown: 4
Work status	No job: 43
	With job: 100
	Study: 6
	Retired: 25
	Unknown:1
Number of days recorded	Min.: 7; Max.: 92

Table 2
Demographics of OViN dataset.

Variable	Description
Gender	Male: 16,298
	Female: 17,088
Age	<18: 8,223
	18–24: 2,446
	25–44: 7,800
	45–64: 9,832
	>=65: 5,085
Education level	Low: 2,131
	Medium: 15,602
	High: 8,188
	Unknown: 7,465
Work status	No job: 2,994
	With job: 1,4585
	Study: 7,254
	Retired: 5,469
	Unknown: 3,084
Number of days recorded	1

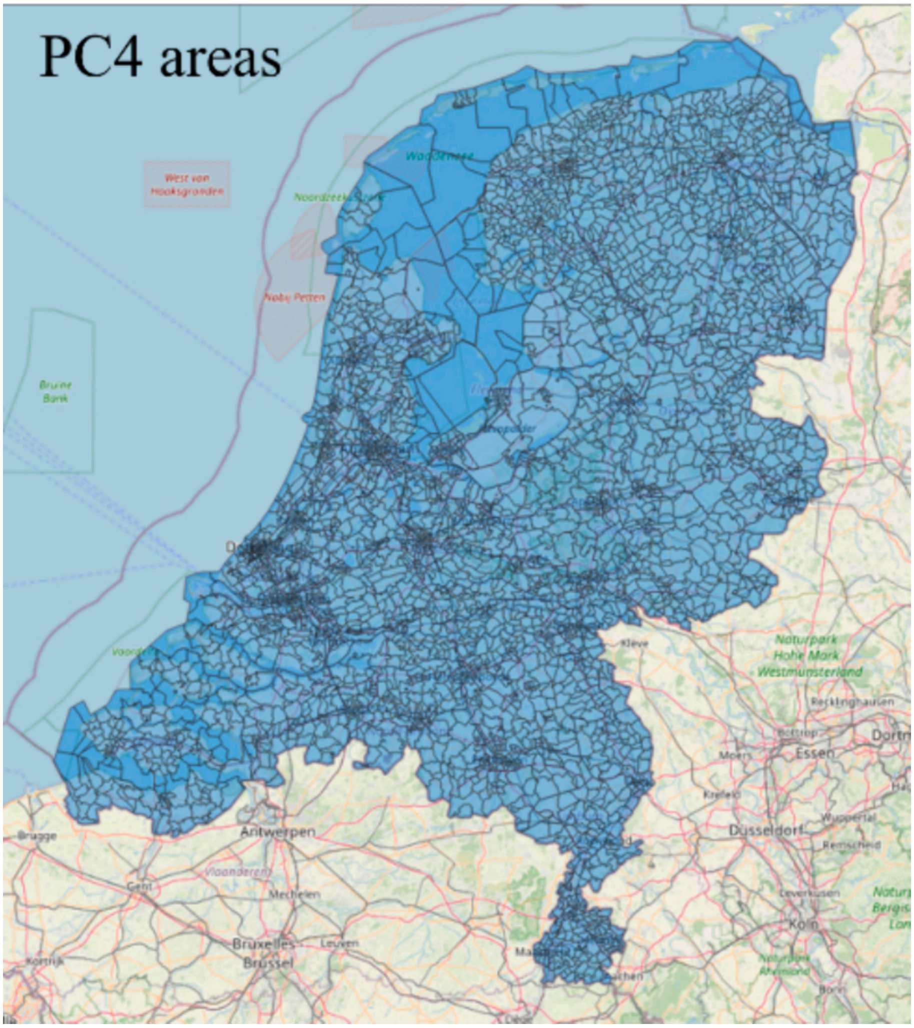

Fig. 2. Boundaries of PC4 areas in the Netherlands.

In 2013, there were 135,762 trips reported from 42,350 respondents. Some incomplete or duplicated reported trips have been deleted during the data cleaning. Eventually, 109,936 trips of 33,386 respondents are used in this research. The demographic information of these respondents are presented in Table 2.

4.3. Spatial data

Although GPS data has accurate location information, the OViN data are at the level of postcodes. In 2013 there existed 4,076 PC4 areas in the Netherlands (CBS, 2013) with the boundaries shown in Fig. 2.

Table 3
Facility and associated POI.

Facility	POI
Shopping facility	Supermarket, bakery, kiosk, mall, department_store, general, convenience, clothes, florist, chemist, bookshop, butcher, shoe_shop, beverages, optician, jeweller, gift_shop, sports_shop, stationery, outdoor_shop, mobile_phone_shop, toy_shop, newsagent, greengrocer, beauty_shop, vedio_shop, car_dealership, bicycle_shop, doityouself, furniture_shop, computer_shop, garden_centre, hairdresser, car_repair, car_rental, car_wash, car_sharing, bicycle_rental, travel_agent, laundry, vending_machine, vending_cigarette, vending_parking
Leisure facility	Theater, nightclub, cinema, park, playground, dog_park, sports_centre, pitch, swimming_pool, tennis_court, golf_course, stadium, ice_rink

Seven built environment-related variables consisting of population density, bus stop density, railway station density, land use mix, shopping facility density, leisure facility density, and surrounding address density are used as features fs. All density data are calculated at PC4 levels. Population and address data are extracted from the census data (CBS, 2013). All other data are extracted from the OpenStreetMap (OSM). Land use mix is measured as the entropy of different land use types following Equation (10).

$$Entropy = -\frac{\sum_{r=1}^{R} P^r \ln P^r}{\ln R} \tag{10}$$

where P^r is the percentage of area of land use type r in the postcode area, and R is the total number of land use types. Table 3 lists the categorization of facilities and associated POIs extracted from OSM.

5. Results

The results of activity-travel feature importance $\beta_{c,f}$ are shown in Table 4. A larger feature importance for the demographics means there is a stronger correlation between the feature and demographics. The results suggest that departure time, arrival time, travel time and activity duration are the top four important features for all four demographics. The railway station density of visited PC4 area is the least important feature regardless of the type of imputed demographics. This may be attributed to the fact that the number of railway stations is extremely limited, with a very low density across all PC4 areas, rendering it unable

Table 4
Weights and ranks of each activity-travel feature for different demographics.

Activity-travel features	Gender	Age	Education level	Work status	Rank (Gender)	Rank (Age)	Rank (Education level)	Rank (Work status)
Departure time	0.069	0.165	0.148	0.214	4	1	3	1
Arrival time	0.080	0.155	0.161	0.201	3	4	2	2
Travel time	0.146	0.164	0.202	0.150	1	2	1	4
Activity duration	0.094	0.163	0.140	0.185	2	3	4	3
Population density[a]	0.029	0.036	0.052	0.056	8	10	10	9
Bus stop density[a]	0.031	0.051	0.086	0.069	6	7	6	6
Railway station density[a]	0.005	0.006	0.009	0.013	13	13	13	13
Land use mix[a]	0.040	0.039	0.069	0.059	5	9	8	8
Shopping facility density[a]	0.028	0.041	0.059	0.052	9	8	9	10
Leisure facility density[a]	0.017	0.030	0.050	0.045	10	12	11	12
Surrounding address density[a]	0.030	0.056	0.098	0.064	7	6	5	7
Trip mode	0.008	0.077	0.076	0.098	11	5	7	5
Trip purpose	0.007	0.031	0.032	0.050	12	11	12	11

[a] In the visited PC4 area.

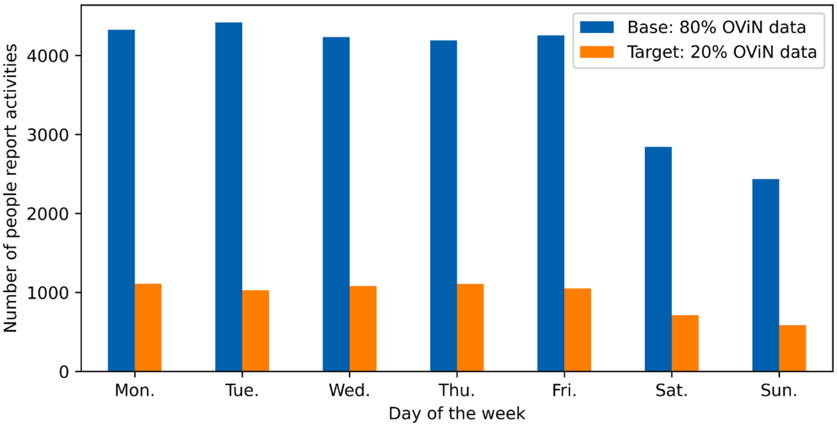

Fig. 3. Number of people each day in target and base samples.

to exhibit the correlation with demographics. Trip mode is an unimportant feature for gender (ranked 11), while relatively important for other demographics (ranked 5, 7, 5). Trip purpose exhibits a relatively weak correlation with all four demographics (ranked 11–12).

20% of the individuals (6,675) from the OViN dataset are set as p_0 (target) and the similarity between each person in the target and the remaining 80% of the dataset (base) is calculated using the estimated weight of the features applying DTW. The number of people on each weekday in the 80% and 20% sub datasets is shown in Fig. 3. In the base database (80% of OViN data), the lowest number of people's reported activity-travel patterns belongs to Sunday including 2,435 individuals. Therefore, for each person in the target dataset, the maximum number of persons matched in the base dataset is set to 2,400 to make matched lists for all days of the week having the same maximum length. By sorting the distance in ascending order, a matched list $[p_0, p_1, \cdots, p_s]$ is obtained. This process generates a matched list for each individual and each demographic feature in the target dataset.

To find the optimum value of the length for the base dataset to be matched as well as the two hyper-parameters C and γ in the RBF kernel, a cross-validated grid search is applied. C is chosen from $\{10^{-1}, 10^{0}, 10^{1}\}$, γ is chosen from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$, and d (number of matched subjects) is chosen from $\{50, 100, 150, \cdots, 2400\}$. In total, there are $3 \times 5 \times 48 = 720$ parameter combinations for each demographic attribute. For each parameter combination, 5-fold cross-validation (Pedregosa et al., 2011; Jahani et al., 2017) is implemented as the final metric. The accuracy of each parameter combination for four demographic attributes is shown in Fig. 4.

When C is 10^0 or 10^1, for all four demographics, increasing the number of people in the matched list increases the accuracy, which is especially evident when γ is 10^{-4}, 10^{-3} and 10^{-2}. When C is 10^{-1}, the accuracy is lower than 0.6. When C is set as 10^0, the highest accuracy can be obtained, exceeding 0.9. In certain subplots, we are unable to observe the accuracy data corresponding to each γ. This is because they are overshadowed by the accuracy results associated with other larger γ. For instance, in subplots (a) and (c), the results for $\gamma = 10^{-1}$ (green dots) are almost entirely overshadowed by the results for $\gamma = 10^0$ (blue dots). Table 5 lists the best parameter combinations for achieving the highest accuracy of each demographic variable. If different parameter combinations lead to the same accuracy, a smaller d (length of the matched list) is considered for further analysis.

After obtaining the optimum parameters, the SVM model is used to impute the demographics of the respondents in GPS dataset (Stage 2). Now, the individuals from the GPS dataset are set as p_0 (target) and OViN dataset is the base dataset. Four metrics, i.e., accuracy, precision, recall and F1 score, are used to measure the performance of the SVM imputation model. Demographics imputation results are shown in Table 6.

Precision is the ratio $tp/(tp + fp)$. Recall is the ratio $tp/(tp + fn)$. F1 score is a harmonic mean of the precision and recall. The relative contribution of precision and recall to the F1 score are equal.

$$F1\ \ score = 2*\frac{precision*recall}{precision + recall} \tag{11}$$

For multilabel targets, the precisions, recalls and F1 scores are calculated for each class and their average value weighted by the number of true instances for each label is regarded as the final score.

It can be found the accuracy exceeds 0.88 for all four imputed demographics, precision exceeds 0.86, recall exceeds 0.88, and F1 score

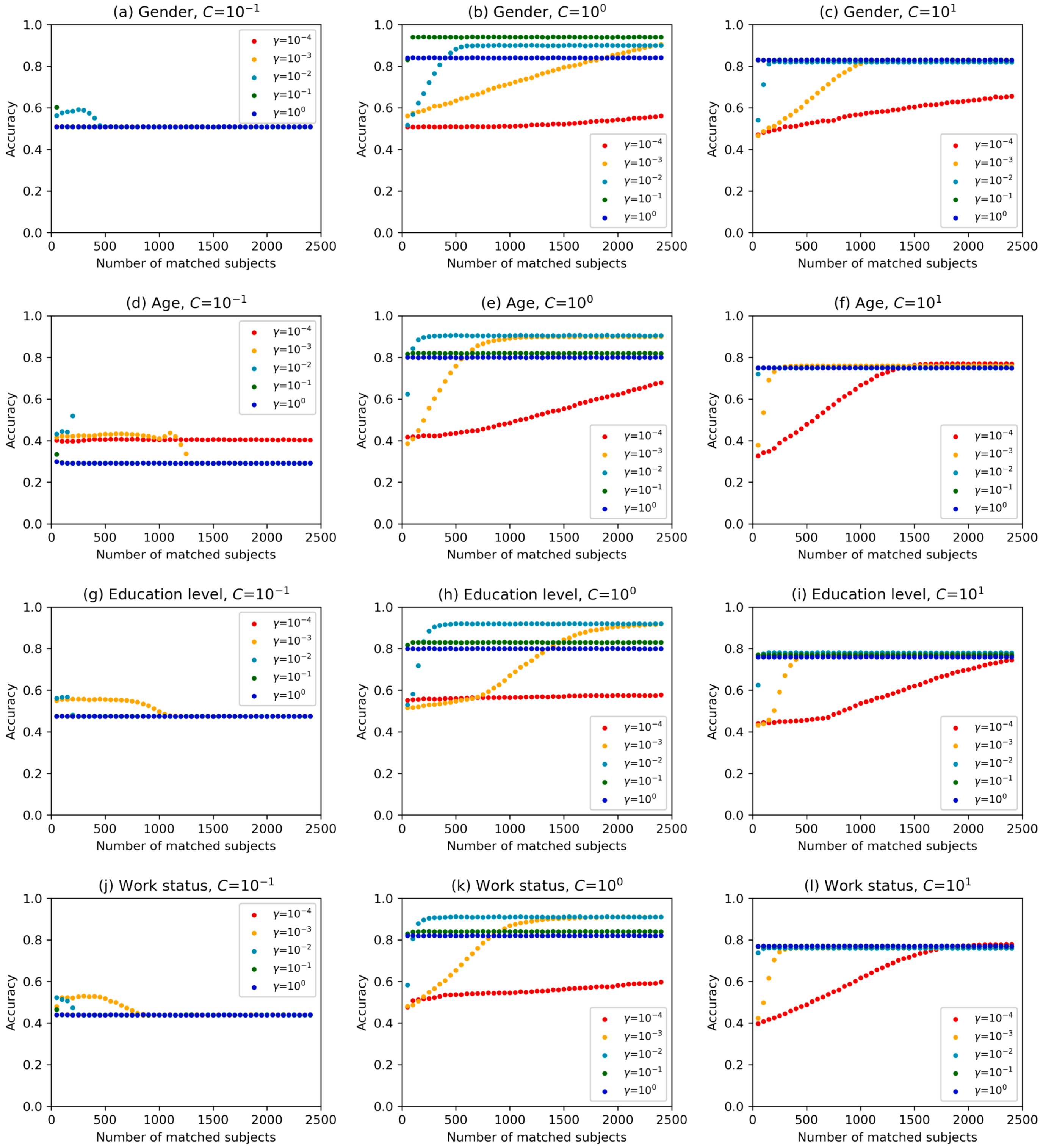

Fig. 4. Accuracy of demographics prediction vs. model parameters.

Table 5
Optimum parameter values.

Demographics	d	C	γ	Accuracy
Gender	200	10^{0}	10^{-1}	0.941
Age	500	10^{0}	10^{-2}	0.906
Education level	700	10^{0}	10^{-2}	0.921
Work status	500	10^{0}	10^{-2}	0.911

Table 6
Demographics imputation results of GPS dataset.

Demographics	Accuracy	Precision	Recall	F1 score
Gender	0.914	0.915	0.914	0.914
Age	0.886	0.906	0.886	0.878
Education level	0.897	0.864	0.897	0.873
Work status	0.880	0.866	0.880	0.860

which balances precision and recall exceeds 0.86. Comparing these values with the ones obtained in the previous research (discussed in Section 2.3), we can conclude that the proposed method outperforms the existing ones in terms of the prediction accuracy.

Further examination of the demographics for which imputation failed reveals an even distribution of misclassified respondents by gender, with 7 males and 8 females. In the meanwhile, the gender distribution in the overall sample is approximately equal. Among respondents with incorrect age imputation, errors occur within the second, third, and fifth age groups, where respondents are erroneously imputed into the fourth age group. This fourth age group, 45–64, comprises the largest subgroup in terms of population. Regarding education level related imputation errors, inaccuracies are observed across low, high, and unknown levels, with respondents have been mistakenly inferred to belong to the medium level, which is the most prevalent category among all education levels. In the case of work status imputation errors, inaccuracies are observed when other categories of work status are imputed as the second status, “with job”, which is also the most common status among all work categories. It is evident that a disproportionate representation of a certain category in demographics can bias the model towards this category during training.

6. Conclusions

While understanding citizens’ travel behavior is imperative for urban and transport policy makers, having knowledge about the social demographics of people is equally important allowing them to customize policies in order to reduce social exclusion and transport poverty. With the development of information and communication technology, the acquisition of travel behavior is becoming increasingly more convenient, while these datasets usually lack demographic information. To address this lack of demographics, we propose a two-step approach to infer the missing information. (Dis)similarity of activity-travel patterns is calculated with weighted features and applying DTW after which a matched list is created for each person in the target dataset. The matched lists are then used to train the SVM imputation model. The results show that our proposed method performs well in imputing the four selected demographics.

This two-step demographics imputation approach holds promise for diverse domains where inferring missing demographic information from activity-travel patterns or leveraging inferred demographics for further research is necessary. The domains can include, but not limited to, urban and transportation planning, social science and market research. In urban and transportation planning, this method can provide invaluable insights into the mobility patterns of different demographic groups within a city or region. By identifying social demographic characteristics, planners can develop more effective and equitable transportation policies and infrastructure designs. In social science research, the application of this approach can contribute to a deeper understanding of human behavior and social dynamics. By analyzing activity-travel patterns and inferred demographic attributes, researchers can explore topics such as social inequality, urban segregation, and community development, thereby informing policy interventions and societal initiatives. Furthermore, in market research, this approach offers the opportunity to do market segment analysis and find out needs and preferences of customers at different life stages.

It is important to realize that GPS tracking solely can provide limited information regarding the activity-travel patterns as it basically can only register time stamp and longitude-latitude coordination. It is important therefore to develop or have access to reliable imputation algorithms by which travel mode and trip purpose can be imputed from the GPS tracking. It is only then, when our proposed method can be applied to impute missing social demographic attributes. Availability of detailed built environment characteristics is equally important in any new area which the proposed methodology deems to be tested. In terms of temporal coverage, the collected data should span various times and days to capture a comprehensive range of activity-travel patterns, considering both weekday and weekend behaviors as well as variations across different times of the day.

Concurrently, data privacy remains a crucial concern. It is of utmost importance that the users whose data is collected are informed about any future attempt to synthesize their social demographic. It can be communicated with them either via privacy policy information sheet or any other types of consent forms. In our research, it was not an issue because our GPS data already contained detailed social demographics of the users.

Although the developed method contributes to the existing body of knowledge, the work still leaves space for further research. Firstly, the determination of feature importance as an integrated part of imputation is worth exploration although it could be a challenge. Secondly, devising a method to find the weights for each person in the matched list can also be added as part of the imputation process. Since the matched list has the order in terms of the extent of similarity, higher weights can be assigned to those in a higher order of the list. Thirdly, there might exist interaction between activity-travels across multiple days (Fu and Lam, 2014, 2018; Astroza et al., 2018; Nayak and Pandit, 2023). Our dataset did not allow to include such interaction because OViN data which was used for training only records the activity-travel pattern for one single day. In addition, our GPS data, although contains observations of multiple days for each person, the days are not necessarily consecutive. Expansion of our proposed framework to include such interaction across days is straightforward once data becomes available. Lastly, applying the same methodology to the data from other areas with substantially different contexts and demographic profiles will allow the examination of the generalibility of the proposed method beyond the current application.

CRediT authorship contribution statement

Bin Zhang: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Soora Rasouli:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Tao Feng:** Conceptualization, Data curation, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The first author of this research was supported by the China Scholarship Council (CSC). We appreciate the insightful comments provided by the anonymous reviewers and the support from the editor.

References

Acheampong, R.A., Siiba, A., Okyere, D.K., Tuffour, J.P., 2020. Mobility-on-demand: An empirical study of internet-based ride-hailing adoption factors, travel characteristics and mode substitution effects. Transp Res Part C Emerg Technol 115, 102638. https://doi.org/10.1016/J.TRC.2020.102638.

Alessandretti, L., Aslak, U., Lehmann, S., 2020. The scales of human mobility. Nature 2020 587:7834 587, 402–407. https://doi.org/10.1038/s41586-020-2909-1.

Astroza, S., Bhat, P.C., Bhat, C.R., Pendyala, R.M., Garikapati, V.M., 2018. Understanding activity engagement across weekdays and weekend days: A multivariate multiple discrete-continuous modeling approach. Journal of Choice Modelling 28, 56–70. https://doi.org/10.1016/J.JOCM.2018.05.004.

Auld, J., Mohammadian, A.K., Oliveira, M.S., Wolf, J., Bachman, W., 2015. Demographic characterization of anonymous trace travel data. Transportation Research Record: Journal of the Transportation Research Board 2526, 19–28. https://doi.org/10.3141/2526-03.

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C.R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J.J., Simini, F., Tomasini, M., 2018. Human mobility: Models and applications. Phys Rep 734, 1–74. https://doi.org/10.1016/J.PHYSREP.2018.01.001.

Brdar, S., Culibrk, D., Crnojevic, V., 2012. Demographic attributes prediction on the real-world mobile data, in: Proc. Mobile Data Challenge by Nokia Workshop, in Conjunction with Int. Conf. on Pervasive Computing, Newcastle, UK. Vol. 1.

Breiman, L., 2001. Random forests. Mach Learn 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Buliung, R.N., Roorda, M.J., Remmel, T.K., 2008. Exploring spatial variety in patterns of activity-travel behaviour: Initial results from the Toronto Travel-Activity Panel Survey (TTAPS). Transportation (Amst) 35, 697–722. https://doi.org/10.1007/S11116-008-9178-4/TABLES/2.

Cao, W.R., Huang, Q.R., Zhang, N., Liang, H.J., Xian, B.S., Gan, X.F., Xu, D.R., Lai, Y.S., 2022. Mapping the travel modes and acceptable travel time to primary healthcare institutions: A case study in Inner Mongolia autonomous region, China. J Transp Geogr 102, 103381. https://doi.org/10.1016/J.JTRANGEO.2022.103381.

CBS, 2013. Kerncijfers per postcode [WWW Document]. URL https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode (accessed 1.16.23).

CBS, 2017. Onderzoek Verplaatsingen in Nederland (OViN) [WWW Document]. URL https://www.cbs.nl/nl-nl/onze-diensten/methoden/onderzoeksomschrijvingen/korte-onderzoeksbeschrijvingen/onderzoek-verplaatsingen-in-nederland–ovin–(accessed 1.16.23).

Chang, C.C., Lin, C.J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2, 1–27. https://doi.org/10.1145/1961189.1961199.

Cheng, L., Chen, X., Yang, S., Cao, Z., De Vos, J., Witlox, F., 2019. Active travel for active ageing in China: The role of built environment. J Transp Geogr 76, 142–152. https://doi.org/10.1016/J.JTRANGEO.2019.03.010.

Cui, Y., He, Q., 2021. Inferring twitters' socio-demographics to correct sampling bias of social media data for augmenting travel behavior analysis. Journal of Big Data Analytics in Transportation 3, 159–174. https://doi.org/10.1007/S42421-021-00037-0.

Deschaintres, E., Morency, C., Trépanier, M., 2022. Cross-analysis of the variability of travel behaviors using one-day trip diaries and longitudinal data. Transp Res Part A Policy Pract 163, 228–246. https://doi.org/10.1016/J.TRA.2022.07.013.

Dharmowijoyo, D.B.E., Susilo, Y.O., Karlström, A., Adiredja, L.S., 2015. Collecting a multi-dimensional three-weeks household time-use and activity diary in the Bandung Metropolitan Area, Indonesia. Transp Res Part A Policy Pract 80, 231–246. https://doi.org/10.1016/J.TRA.2015.08.001.

Dharmowijoyo, D.B.E., Susilo, Y.O., Karlström, A., 2017. Analysing the complexity of day-to-day individual activity-travel patterns using a multidimensional sequence alignment model: A case study in the Bandung Metropolitan Area, Indonesia. J Transp Geogr 64, 1–12. https://doi.org/10.1016/J.JTRANGEO.2017.08.001.

Ding, C., (Jason) Cao, X., Næss, P., 2018. Applying gradient boosting decision trees to examine non-linear effects of the built environment on driving distance in Oslo. Transp Res Part A Policy Pract 110, 107–117. https://doi.org/10.1016/J.TRA.2018.02.009.

Dong, Y., Chawla, N.V., Tang, J., Yang, Y., Yang, Y., 2017. User modeling on demographic attributes in big mobile social networks. ACM Trans Inf Syst 35, 1–33. https://doi.org/10.1145/3057278.

Dong, Y., Yang, Y., Tang, J., Chawla, N. V., 2014. Inferring user demographics and social strategies in mobile social networks, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, pp. 15–24. https://doi.org/10.1145/2623330.2623703.

Duan, Z., Zhao, H., Li, Z., 2023. Non-linear effects of built environment and socio-demographics on activity space. J Transp Geogr 111, 103671. https://doi.org/10.1016/J.JTRANGEO.2023.103671.

Ewing, R., Cervero, R., 2001. Travel and the Built Environment: A Synthesis. Transportation Research Record: Journal of the Transportation Research Board 1780, 87–113. https://doi.org/10.3141/1780-10.

Farinloye, T., Mogaji, E., Aririguzoh, S., Kieu, T.A., 2019. Qualitatively exploring the effect of change in the residential environment on travel behaviour. Travel Behav Soc 17, 26–35. https://doi.org/10.1016/J.TBS.2019.06.001.

Faroqi, H., Mesbah, M., Kim, J., Tavassoli, A., 2018. A model for measuring activity similarity between public transit passengers using smart card data. Travel Behav Soc 13, 11–25. https://doi.org/10.1016/J.TBS.2018.05.004.

Feng, T., Timmermans, H.J.P., 2014. Enhanced Imputation of GPS Traces Forcing Full or Partial Consistency in Activity Travel Sequences: Comparison of Algorithms. Transportation Research Record: Journal of the Transportation Research Board 2430, 20–27. https://doi.org/10.3141/2430-03.

Figueroa Martínez, C., Hodgson, F., Mullen, C., Timms, P., 2019. Walking through deprived neighbourhoods: Meanings and constructions behind the attributes of the built environment. Travel Behav Soc 16, 171–181. https://doi.org/10.1016/J.TBS.2019.05.006.

Fu, X., Lam, W.H.K., 2014. A network equilibrium approach for modelling activity-travel pattern scheduling problems in multi-modal transit networks with uncertainty. Transportation (Amst) 41, 37–55. https://doi.org/10.1007/S11116-013-9470-9/FIGURES/6.

Fu, X., Lam, W.H.K., 2018. Modelling joint activity-travel pattern scheduling problem in multi-modal transit networks. Transportation (Amst) 45, 23–49. https://doi.org/10.1007/S11116-016-9720-8/FIGURES/10.

Garikapati, V.M., Pendyala, R.M., Morris, E.A., Mokhtarian, P.L., McDonald, N., 2016. Activity patterns, time use, and travel of millennials: a generation in transition? Transp Rev 36, 558–584. https://doi.org/10.1080/01441647.2016.1197337.

Giorgino, T., 2009. Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. J Stat Softw 31, 1–24. https://doi.org/10.18637/JSS.V031.I07.

Goulet-Langlois, G., Koutsopoulos, H.N., Zhao, J., 2016. Inferring patterns in the multi-week activity sequences of public transport users. Transp Res Part C Emerg Technol 64, 1–16. https://doi.org/10.1016/J.TRC.2015.12.012.

Hafezi, M.H., Liu, L., Millward, H., 2019. A time-use activity-pattern recognition model for activity-based travel demand modeling. Transportation (Amst) 46, 1369–1394. https://doi.org/10.1007/S11116-017-9840-9/TABLES/4.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A Practical Guide to Support Vector Classification [WWW Document]. URL https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (accessed 1.13.23).

Hu, J., Zeng, H.-J., Li, H., Niu, C., Chen, Z., 2007. In: Demographic Prediction Based on User's Browsing Behavior, in. ACM Press, New York, New York, USA, pp. 151–160. https://doi.org/10.1145/1242572.

Jahani, E., Sundsøy, P., Bjelland, J., Bengtsson, L., Pentland, A.S., de Montjoye, Y.A., 2017. Improving official statistics in emerging markets using machine learning and mobile phone data. EPJ Data Sci 6, 1–21. https://doi.org/10.1140/EPJDS/S13688-017-0099-3.

Jiang, S., Ferreira, J., González, M.C., Wang, F., Tong, H., Yu, P., Jiang, A.S., Ferreira, J., González, M.C., 2012. Clustering daily patterns of human activities in the city. Data Min Knowl Discov 25, 478–510. https://doi.org/10.1007/S10618-012-0264-Z.

Joh, C.H., Arentze, T., Hofman, F., Timmermans, H., 2002. Activity pattern similarity: a multidimensional sequence alignment method. Transp. Res. B Methodol. 36, 385–403. https://doi.org/10.1016/S0191-2615(01)00009-1.

Joh, C.H., Arentze, T.A., Timmermans, H.J.P., 2016. A position-sensitive sequence-alignment method illustrated for space-time activity-diary data. Environment and Planning a: Economy and Space 33, 313–338. https://doi.org/10.1068/A3323.

Kang, H., Scott, D.M., 2010. Exploring day-to-day variability in time use for household members. Transp Res Part A Policy Pract 44, 609–619. https://doi.org/10.1016/J.TRA.2010.04.002.

Kosinski, M., Stillwell, D., Graepel, T., 2013. Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci U S A 110, 5802–5805. https://doi.org/10.1073/pnas.1218772110.

Kwan, M.P., Xiao, N., Ding, G., 2015. Assessing activity pattern similarity with multidimensional sequence alignment based on a multiobjective optimization evolutionary algorithm. Geogr Anal 46, 297. https://doi.org/10.1111/GEAN.12040.

Lenormand, M., Louail, T., Cantú-Ros, O.G., Picornell, M., Herranz, R., Arias, J.M., Barthelemy, M., Miguel, M.S., Ramasco, J.J., 2015. Influence of sociodemographic characteristics on human mobility. Sci Rep 5, 1–15. https://doi.org/10.1038/srep10075.

Li, Q., Zheng, Y., Xie, X., Chen, Y., Liu, W., Ma, W.Y., 2008. Mining user similarity based on location history. GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems 298–307. https://doi.org/10.1145/1463434.1463477.

Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P., 2010. You are who you know: Inferring user profiles in online social networks, in: WSDM 2010 - Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. pp. 251–260. https://doi.org/10.1145/1718487.1718519.

Mouratidis, K., Ettema, D., Næss, P., 2019. Urban form, travel behavior, and travel satisfaction. Transp Res Part A Policy Pract 129, 306–320. https://doi.org/10.1016/J.TRA.2019.09.002.

Nayak, S., Pandit, D., 2023. A joint and simultaneous prediction framework of weekday and weekend daily-activity travel pattern using conditional dependency networks. Travel Behav Soc 32, 100595. https://doi.org/10.1016/J.TBS.2023.100595.

Pawlak, J., Zolfaghari, A., Polak, J., 2015. Imputing Socioeconomic Attributes for Movement Data by Analysing Patterns of Visited Places and Google Places Database: Bridging between Big Data and Behavioural Analysis, in: International Choice Modelling Conference 2015.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research 12, 2825–2830. https://doi.org/10.5555/1953048.2078195.

Raux, C., Ma, T.Y., Cornelis, E., 2016. Variability in daily activity-travel patterns: the case of a one-week travel diary. Eur. Transp. Res. Rev. 8, 1–14. https://doi.org/10.1007/S12544-016-0213-9/TABLES/9.

Sarraute, C., Blanc, P., Burroni, J., 2014. A study of age and gender seen through mobile phone usage patterns in Mexico, in: ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Institute of Electrical and Electronics Engineers Inc., pp. 836–843. https://doi.org/10.1109/ASONAM.2014.6921683.

Shou, Z., Di, X., 2018. Similarity analysis of frequent sequential activity pattern mining. Transp Res Part C Emerg Technol 96, 122–143. https://doi.org/10.1016/J.TRC.2018.09.018.

Su, R., McBride, E.C., Goulias, K.G., 2020. Pattern recognition of daily activity patterns using human mobility motifs and sequence analysis. Transp Res Part C Emerg Technol 120, 102796. https://doi.org/10.1016/J.TRC.2020.102796.

Wang, D., Chai, Y., Li, F., 2011. Built environment diversities and activity–travel behaviour variations in Beijing, China. J Transp Geogr 19, 1173–1186. https://doi.org/10.1016/J.JTRANGEO.2011.03.008.

Wang, P., Guo, J., Lan, Y., Xu, J., Cheng, X., 2016. Your Cart tells You: Inferring Demographic Attributes from Purchase Data, in: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. ACM, New York, NY, USA, pp. 173–182. https://doi.org/10.1145/2835776.

Wang, Z., Hale, S.A., Adelani, D., Grabowicz, P.A., Hartmann, T., Flöck, F., Jurgens, D., 2019. Demographic inference and representative population estimates from multilingual social media data. The Web Conference 2019 - Proceedings of the World

Wide Web Conference, WWW 2019 12, 2056–2067. https://doi.org/10.1145/3308558.3313684.

Wu, J., Wang, B., Wang, R., Ta, N., Chai, Y., 2021. Active travel and the built environment: A theoretical model and multidimensional evidence. Transp Res D Transp Environ 100, 103029. https://doi.org/10.1016/J.TRD.2021.103029.

Wu, L., Yang, L., Huang, Z., Wang, Y., Chai, Y., Peng, X., Liu, Y., 2019. Inferring demographics from human trajectories and geographical context. Comput Environ Urban Syst 77, 101368. https://doi.org/10.1016/j.compenvurbsys.2019.101368.

Xianyu, J., Rasouli, S., Timmermans, H., 2017. Analysis of variability in multi-day GPS imputed activity-travel diaries using multi-dimensional sequence alignment and panel effects regression models. Transportation (amst) 44, 533–553. https://doi.org/10.1007/s11116-015-9666-2.

Xu, F., Lin, Z., Xia, T., Guo, D., Li, Y., 2020. SUME: Semantic-enhanced Urban Mobility Network Embedding for User Demographic Inference, in: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. ACM PUB27 New York, NY, USA, pp. 1–25. https://doi.org/10.1145/3411807.

Zhang, Y., Sari Aslam, N., Lai, J., Cheng, T., 2020. You are how you travel: A multi-task learning framework for Geodemographic inference using transit smart card data. Comput Environ Urban Syst 83, 101517. https://doi.org/10.1016/J.COMPENVURBSYS.2020.101517.

Zhao, Y., Pawlak, J., Sivakumar, A., 2022. Theory for socio-demographic enrichment performance using the inverse discrete choice modelling approach. Transp. Res. B Methodol. 155, 101–134. https://doi.org/10.1016/J.TRB.2021.11.004.

Zhong, Y., Yuan, N.J., Zhong, W., Zhang, F., Xie, X., Research, M., 2015. You Are Where You Go: Inferring Demographic Attributes from Location Check-ins, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, New York, NY, USA, pp. 295–304. https://doi.org/10.1145/2684822.2685287.

Zhou, Y., Yuan, Q., Yang, C., Wang, Y., 2021. Who you are determines how you travel: Clustering human activity patterns with a Markov-chain-based mixture model. Travel Behav Soc 24, 102–112. https://doi.org/10.1016/J.TBS.2021.03.005.