ActivityNET: Neural networks to predict trip purposes in public transport from individual smart card data and POIs.

Nilufer Sari Aslam [a], Mohamed R. Ibrahim [a], Tao Cheng [a],
Huanfa Chen [b], Yang Zhang [a]

[a] SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering,

University College London (UCL), London, UK; [b] Centre for Advanced Spatial Analysis
(CASA), University College London, UK

n.aslam.11@ucl.ac.uk [a] (corresponding author), mohamed.ibrahim.17@ucl.ac.uk [a],
tao.cheng@ucl.ac.uk [a], huanfa.chen@ucl.ac.uk [b], yang.zhang.16@ucl.ac.uk [a]

Abstract

Predicting trip purpose from comprehensive and continuous smart card data is beneficial for transport and city planners investigating travel behaviours, and mobility research in urban areas. Here we propose a framework, ActivityNET, using machine learning (ML) algorithms to predict passengers' trip purpose from smart card data and Points-Of-Interest (POIs). The feasibility of the framework is demonstrated in two phases. Phase I focuses on extracting activities from individuals' daily travel patterns from smart card data and combining them with POIs using the proposed "activity-POIs consolidation algorithm". Meaningful match helps to generate an understanding of human mobility and urban flows in cities. Phase II feeds the extracted features into an artificial neural network (ANN) with multiple scenarios and predict trip purpose under primary activities (home and work) and secondary activities (entertainment, eating, shopping, child drop-offs/pick-ups and part-time work) with high accuracy. As a case study, the proposed ActivityNET framework is applied in Greater London and illustrates a robust competence to predict trip purpose. The promising outcomes demonstrate that the cost-effective framework offers valuable insight into transport planning.

KEYWORDS: Trip purpose prediction, Smart card data, POIs, neural networks, machine learning.

1. Introduction

Activity-based models aim to predict travel demand using trip purposes to understand and plan the transport network usage under different socio-economic scenarios and land-use structures.

Transport planning with such models rely on travel surveys, which are relatively small sample sizes, are expensive to obtain and have relatively low update frequencies (collected only one-day). Therefore, they are prone to bias when estimating travel demand for the whole population (Yang *et al.*, 2019). On the other hand, collecting smart card data has shown great potential for investigating passengers' daily activities at an unprecedented scale such as a much larger population and a longer period of data collection (Anda, Erath and Fourie, 2017). In addition, smart card data reveal an individual's spatial-temporal activity pattern as a sequence of activity locations, activity start and end time, duration of the activity and land-use in the close proximity of alighting or boarding station (Faroqi, Mesbah and Kim, 2018) which could be further explored to derive the trip purpose of the travellers (Sari Aslam and Cheng, 2018; Sari Aslam *et al.*, 2020).

Trip purpose is essential for planning purposes, performance evaluation and the development of public transit networks and services (Faroqi, Mesbah and Kim, 2018). The scope of the research expands to consumer behaviour for commercial establishments (Longley, Cheshire and Singleton, 2018), urban mobility, and people flows for city planners (Yang *et al.*, 2019), the aspiration of the quality life for economists (Nakamura *et al.*, 2016), and public health for policy and decision-makers, e.g. the spread of COVID-19 (Ibrahim *et al.*, 2020). Thus, longitudinal smart card data with volume and details need to be investigated for trip purposes such as home, work, entertainment, eating, shopping, drop-offs/pick-ups, and part-time work activities. However, the majority of the trip purpose identification models from smart card data are focused on only primary activities such as home and work/school (for adults and students, respectively) (Chakirov and Erath, 2012; Devillaine, Munizaga and Trépanier, 2012; Zou *et al.*, 2016; Yang *et al.*, 2019; Sari Aslam, Cheng and Cheshire, 2019) but rarely secondary activities (Alsger *et al.*, 2018; Sari Aslam *et al.*, 2020). The reason is that the handcrafted rules and number of constraints are limited and reduce the ability to identify trip purposes with high accuracy, specifically for secondary activities (Xiao, Juan and Zhang, 2016; Anda, Erath and Fourie, 2017), which are complex compared to regular commuters' activities. Therefore, there is a need to investigate trip purposes using data-driven ML approaches, which are flexible enough to capture complex information about trip purposes without any defined rules. Besides, they are capable of handling a non-linear problem with high accuracy (Xiao, Juan and Zhang, 2016; Anda, Erath and Fourie, 2017).

Although ML methods focused on clustering and classification of trips, passengers, and stations to investigate travel patterns and behaviours from smart card data (Faroqi, Mesbah and Kim, 2018), trip purposes hardly investigated from individuals' activities (Lee and Hickman, 2014; Kusakabe and Asakura, 2014; Han and Sohn, 2016). The reason is that, first, the model performance is low due to noise in unprocessed smart card data, requiring pre-processing steps before applying prediction models to achieve high accuracy (Dacheng *et al.*, 2018; Zhang *et al.*, 2020). Second, aggregated input features per user from a large volume of travel data, such as average travel duration, and average departure time of the first/last trips (Goulet-langlois, Koutsopoulos and Zhao, 2016; Han and Sohn, 2016), may not accurately represent activity points and incorporate cumulative errors in the model. Third, how robustly smart card data ought to be enriched by other data sources such as POIs to represent the semantic interpretations of activities (Yang *et al.*, 2019).

Therefore, in this study, we propose using the ActivityNET framework to predict passengers' trip purposes, for each activity per individual from their smart card data. The feasible framework

includes the following: *The first phase of the study* focuses on extracting activities from the travel dataset and combining spatial and temporal attributes of longitudinal smart card data and POIs using an "activity-POIs consolidation algorithm". This part of the study offers an understanding of human mobility and urban flows from two big data sources in cities. In addition, the combined dataset provides input features under three sub-groups, such as activity characteristics (activity start and end time, activity duration), day characteristics, and land-use characteristics. *The second phase of the study* uses input features with multiple scenarios and predict trip purposes with ANN under primary (home and work) and secondary activities (entertainment, eating, shopping, child drop-offs/pick-ups, and part-time work activities) with high accuracy.

The contributions of this study can be summarised as follows:

- Proposed "activity-POIs consolidation algorithm" aims to explore how two large longitudinal datasets such as smart card data and POIs ought to be combined for trip purpose prediction.

- Proposed ActivityNET framework uses multiple scenarios and predicts trip purposes for primary and secondary activities using ML algorithms with high precision.

- The trip purpose prediction model, ActivityNET, is a cost-effective method using smart card and POIs to help transport and urban planning.

The next section of the paper presents the data and methods with the logical framework. The following section (section 3) offers the result of the methodology under a case study in London. Finally, discussion, and conclusions of the work are presented under section 4 and 5, respectively

2. Data and Methods

2.1. Dataset

2.1.1. Travel Survey (Oyster card) data

Smart card data provided by Transport for London (TfL) called Oyster card, which is a payment method for public transport when a passenger taps in/out at a station in London. Automatically recorded Oyster card data have attributes such as boarding and alighting time, boarding and alighting station, and transport mode without trip purpose information. In addition, TfL allows each individual to download their travel data (minimum of two months). To investigate trip purposes from smart card data, the total of 19792 trip records (9116 activity/data points) is collected for this study. Trip purposes are labelled by volunteers under seven sub-categories such as home (3994 data points), work (2006 data points), entertainment (555 data points), eating (687 data points), shopping (818 data points), child drop-offs/pick-ups (629 data points) and part-time work activities (427 data points). Besides, 5387 and 3729 of the activity points are represented as female, and male volunteers, respectively. The details of the data are subdivided into four income bands; 2486 data points represent no income, 1657 of them with earnings below £25,000, 2901 activities are between £25,000 and £40,000, and 2072 of the data with earnings of more than £40,000. Similarly, the collected data are divided into three groups based on the ages of the

participants: 3867 data points are less than 30 years old, 3453 are between 30 and 40, and 1796 activities are more than 40 years old. As of last, under occupation group, 4972 of the activities are titled as professional, and 4144 of them are students. At the end of the data collection and processing section, individual data is anonymised under GDPR rules (ICO, 2018).

2.1.2. Foursquare data

Points of interest data were collected using the Foursquare Location API. The total number of user-count (the number of users) and check-ins (the number of check-ins by users) are 38,232,516 and 81,328,352, respectively in London. There are two ways to use POIs in this study. The first one is the working hours of POIs during the weekdays and weekends, the second one is the classification of location category under seven sub-types such as home, work, entertainment, eating, shopping, outdoors & recreation, and travel & transport, as shown in table 1. The details of the POI data and data pre-processing steps are illustrated in section 2.2.1.2.

Table 1: Activity types from foursquare data.

Activity types	Activity location type
Home	Residential building (apartment/condo), housing developments
Work	Government buildings, library, post office, schools, colleges and universities, warehouse etc.
Entertainment	Art, pub, nightclub, arcade, theatre, entertainment, club, bar, concert hall, opera house, casino, etc.
Eating	Coffee shop, sandwich, pizza, bakery, burger, restaurant, steakhouse, breakfast, etc.
Shopping	Supermarket, store, pharmacy, mall, boutique, farmers market, food & drink shop, bookstore etc.
Outdoors & recreation	Outdoor & recreation (park, playground, recreation centre, ski areas, etc.)
Travel & transport	Travel & transport (hotels, bus stops, tube stations, bike rental/bike share, airports etc.)

2.2. Methods

The proposed ActivityNET framework in Figure 1 predicts trip purpose in two phases. Phase I, data pre-processing, shows how two large data sources, namely smart card data and POIs, are combined using the proposed "activity-POIs consolidation algorithm" after extracting activities. Thus, the location information, e.g. station name, from travel data, can be enhanced by dynamic

socio-economic land-use attributes. Phase II, extracted spatial-temporal features are selected with multiple scenarios and passed into the model to predict trip purposes within sub-categories, e.g., home, work, entertainment, eating, shopping, child drop-offs/pick-ups, and part-time work activities. Hence, the reason for the trips is investigated, revealing why people spent their spare time within the city using smart card data with the help of POIs.
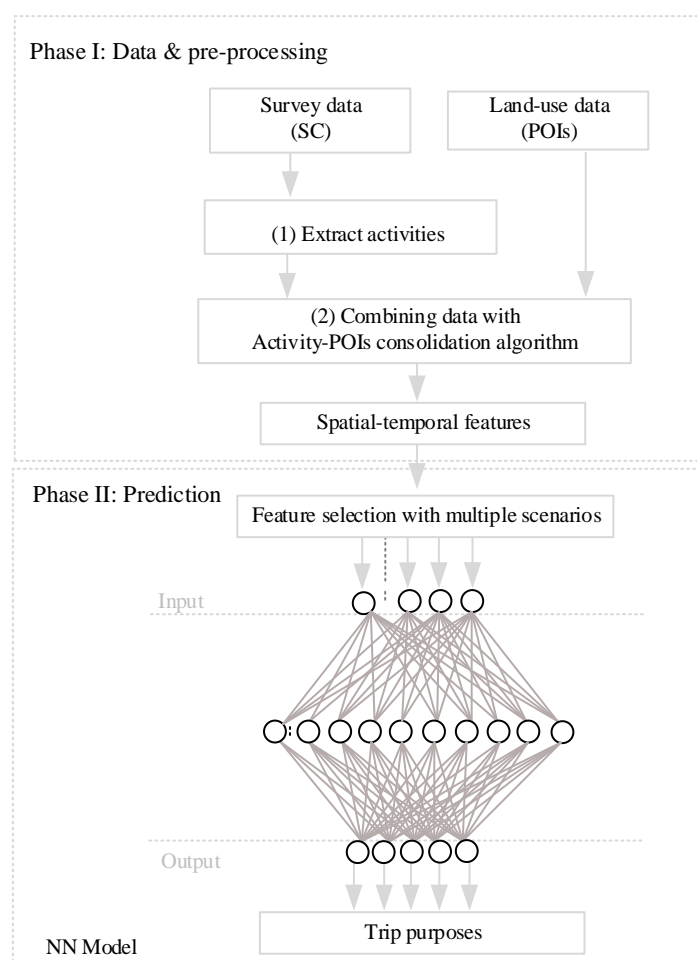
Figure 1: The logical flow of ActivityNET framework (SC and NN refer to smart card data and neural networks, respectively)

2.2.1. Phase I: Data pre-processing

This section aims to increase the accuracy of the large travel datasets while cleaning SC data such as excluding single trips and the records with missing information. Thus, 1060 single trips are excluded from the total of 19792 trip records. The reason is that single trips do not have sufficient information to present activities. Besides, 499 missing trips, e.g. alighting time or station, boarding time or station, are also excluded, which create uncertainty to extract activities (Chakirov and Erath, 2012). After data cleaning section is complete, data is ready to extract activities from travel data and combine with POIs. Hence, the meaningful match provides high accuracy in the prediction model to explore trip purposes from travel data (Anda, Erath and Fourie, 2017; Dacheng *et al.*, 2018).

2.2.1.1. Extract activities

The definition of a trip is a one-way journey from one stop to another stop. An activity is the time duration between two consecutive trips such as the alighting station of the first trip and the boarding stations of the second trip. There is a sequence of activities in a day per individuals with their characteristics such as start-end time of the activity, the location of the activity, the day of the activity, which can be used to infer trip purposes.

Trip purpose (the reason for the trip) is to find an answer "why has an activity happened in a specific location and time?" To achieve this, the location of the transit data need to be enriched using other data sources, e.g. land-use information. Then it is possible to infer trip purposes using the characteristics of activities from smart card data and the type of activities such as home, work/school, entertainment, eating, shopping and some other type of activities from POIs (Faroqi, Mesbah and Kim, 2018).

The assumptions of activity extraction are applied in this stage (Sari Aslam *et al.*2020) using transfer time and walking distance between public transit stops, which were assumed to be 15 min (Transport for London TfL, 2019) and 800 m (RTPI, 2018; Alsger *et al.*, 2018; Sari Aslam *et al.*, 2020), respectively. The resulting dataset consists of 18,232 trip records, which means 9,116 data points (activities) from smart card data.

2.2.1.2. Combining both datasets using activity-POIs consolidation algorithm

POIs from Twitter or Foursquare data have been used to investigate trip purposes, human mobility and urban flows to generate an understanding of transport and urban planning in cities (Chaniotakis, Antoniou and Pereira., 2016; Rashidi *et al.*, 2017). To infer activities from transit data, the highest probability of activity types are used from POIs (Gong, L., X. Liu, L. Wu, 2016; Alsger *et al.*, 2018; Sari Aslam *et al.*, 2020). However, this section explains how both large datasets such as smart card data and land-use information (POIs) ought to be combined in details. Moreover, this study uses the relevant POIs with aggregated counts values as input data for the ML algorithm to predict trip purposes using the labelled data.

Before combining working hours of POIs and type of POIs using POI ID, stage 1 in figure 2, working hours of POIs are merged from seven working days to weekdays and weekends. Because the working hours of each POIs have some variation on different days. Thus, the earliest and the latest working hours for each POIs are used for weekdays and weekends in this section. For instance, if opening hours and closing hours of a coffee shop are 10:00 to 15:00 on Saturdays and 12:00 to 16:00 on Sundays, the opening and closing hours for the place are 10:00 and 16:00 for weekends. The similar condition is also applied for the weekdays, and each location has four records as opening and closing hours of weekdays and weekends. After that, smart card data and POI datasets (opening and closing hours of POIs & type of POIs) are combined using location/station ID (Sn). Second, in stage 2a, the idea is presented how each activity from smart card data needs to be checked step by step for each POI from foursquare data. The reason is that each station has different numbers of activities (Ai) and POIs, which are in walking distance. In addition, if there are POIs in catchment areas, e.g. between station 210 and 211, both stations have the POIs from the catchment area in their list to be checked. In stage 2b, proposed "activity-POIs consolidation algorithm" is presented how relevant POIs are filtered for each activity and

aggregated (sums) the counts in activity types. First, selecting a station and activity in that station, needs to be checked "do we have POIs at the station in walking distance?" If it is yes, select a POI for that activity. Then the activity-POI match is tested against two conditions: "the start time of the activity ≥ the opening time of POIs (for weekdays/weekends) & the end time of the activity ≤ the closing time of POIs (for weekdays/weekends)". If the conditions are met, the number of user-count are saved for the first activity. The algorithm then moves to the next possible POI for the same activity. Once all possible POIs have been checked, the counts are aggregated in each activity type and, the process is repeated for the next activity. When all activities have been tested, the process is repeated for the next station. At the end of the process, as an output (stage 2c), each activity is represented by different aggregated (sum) values for each of the POI categories, e.g. home, work, entertainment, eating, shopping, outdoor & recreational, and travel and transport activities. As a result, the characteristics of land-use information using the counts of POIs is assigned to each activity as different weights.
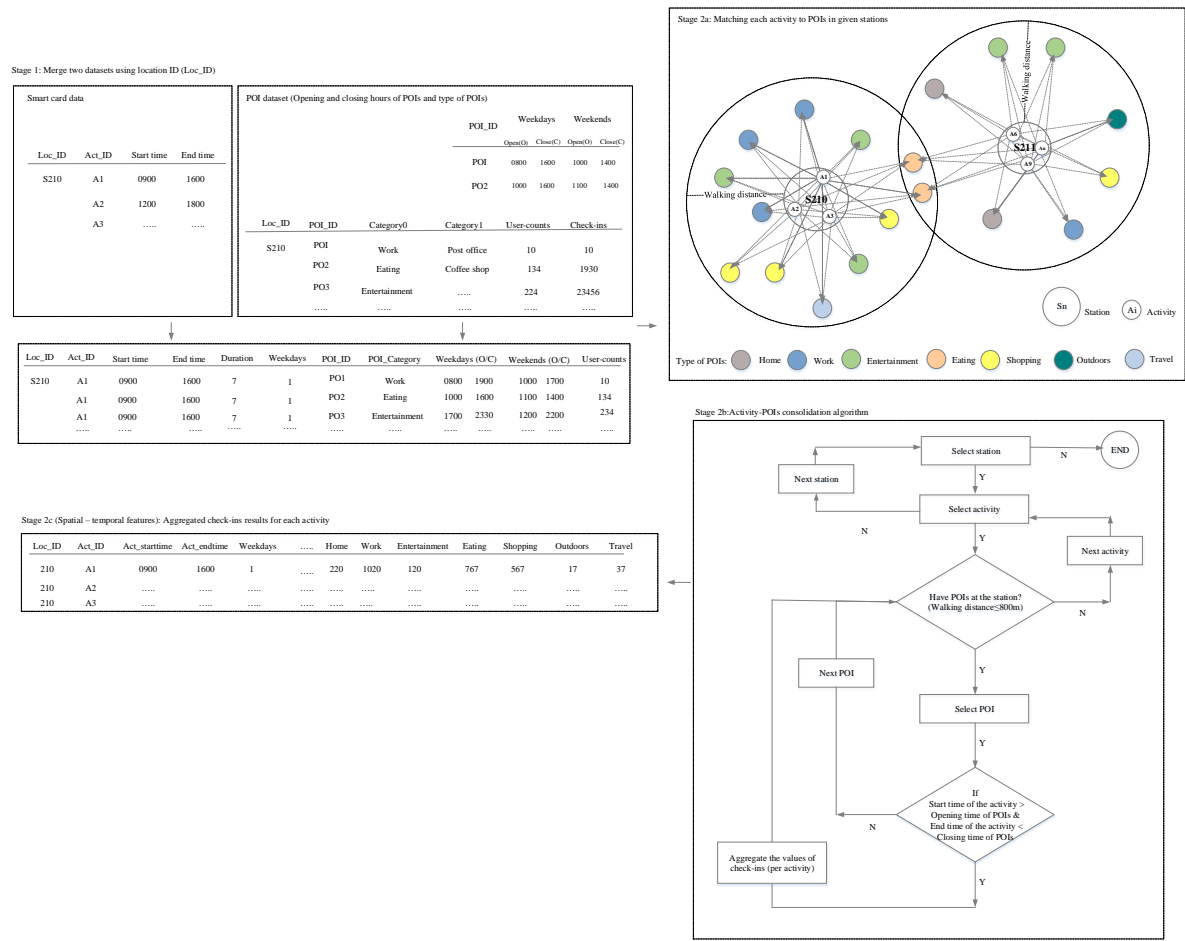
Figure 2: How both datasets are combined in two stages. 1) Combining both datasets using location ID, 2) using "activity-POIs consolidation algorithm" to filter relevant POIs for each activity and aggregate (sums) user-counts based on activity types. Sn and An refer to stations and activities, respectively.

Note that the steps mentioned in this section may create memory issues such as "out of memory" due to the size of the data. The processing steps need to be executed in a distributed way using

parallel machines. Data processing packages such as Pandas in python are designed to work in a single machine, and their memory allowance has a low threshold. Therefore, PySpark is used to be able to carry out the analysis in this section.

As a result, combined input features are presented with details as temporal features (activity characteristics, day characteristics) and spatial features (land-use characteristics) in table 2.

Table 2: Input features to identify trip purposes.

Category	Feature	Definition
Trip purposes	TRP_PURP	The labelled activities for the reason of the trip
Activity characteristics	ACT_DUR	Duration of the activity (hrs)
	ACT_ST_TIME	The start time of the activity in 24hrs
	ACT_EN_TIME	The end time of the activity in 24hrs
Day characteristics	Weekdays	The activity has happened on Weekdays (1)/Otherwise (0)
Land-use characteristics	HOM	Aggregated user-counts of POIs for home locations
	WOR	Aggregated user-counts of POIs for work locations
	ENT	Aggregated user-counts of POIs for entertainment locations
	EAT	Aggregated user-counts of POIs for eating locations
	SHO	Aggregated user-counts of POIs for shop locations
	SCH	Aggregated user-counts of POIs for school locations
	REC	Aggregated user-counts of POIs for outdoors & recreation
	TPO	Aggregated user-counts of POIs for transport stations & hubs

2.2.2. Phase II: Prediction of trip purposes

This section shows the structure of the model, training the model using input features, and the prediction of trip purposes using the trained model illustrated under "phase II" in figure 1.

2.2.2.1. The structure of the artificial neural network (ANN) with multiple scenarios

The artificial neural network is applied for predictive analysis to classify multi-class trip purposes using its non-linear pattern classification capabilities. The reason is that neural networks are capable of handling dimensionality of the problem using spatial dependencies in a large dataset with high accuracy and low computing time (Xiao, Juan and Zhang, 2016; Ibrahim *et al.*, 2019) compared to statistical models, which are parametric and struggle from high computational complexity in large-scale scenarios. On the other hand, standard ML methods are narrow in architecture that cannot comprehensively handle non-linear large spatial-temporal data with high dimensionality.

The details of the structure of the model illustrated in figure 3 are provided in the following subsections;

1. Input layer: The first layer of neural networks transfers the information from input features using the same dimensionality. Due to class imbalance issues (see section 2.1.1), (1) random over-sampling technique that duplicates data points randomly in the minority classes and (2) random

under-sampling technique that removes data points from majority classes randomly (Brownlee, 2020c), are compared to (3) unchanged values in this section. In addition, the dimensionality of the layer is increased and decreased including (input dimension = 12, with POIs) and excluding of spatial features (input dimension = 4, without POIS) to evaluate overall accuracy with different scenarios in the model (section 3.2).

2. Hidden layers: These layers process the information from the input layer to output layers. In this section, the number of neurons and functions needs to be investigated. Even though there is no rule of thumb to choose the number of layers in neural network (Goodfellow, Bengio and Courville, 2017), two hidden layers are processed the transformation, one with 100 and one with 60 units, which are activated using the Rectified Linear Unit (ReLU) (Glorot and Bordes, 2011) to increase the nonlinearity of the model and improve the performance of the unites (Dahl, Sainath and Hinton, 2013).

Dropout technique (Hinton *et al.*, 2012) is considered after hidden layers with 0.5 to reduce overfitting. The cross-entropy loss was applied to the model as the training objective function. The model is compiled using the stochastic gradient descent Adam optimiser (Kingma and Ba, 2015) to minimise the loss function with an initial learning rate of 0.001. Different values of mini-batch gradient descents with different possible epochs are also investigated, and the best accuracy is attained using a batch size of 64 with 700 epochs during the training process.
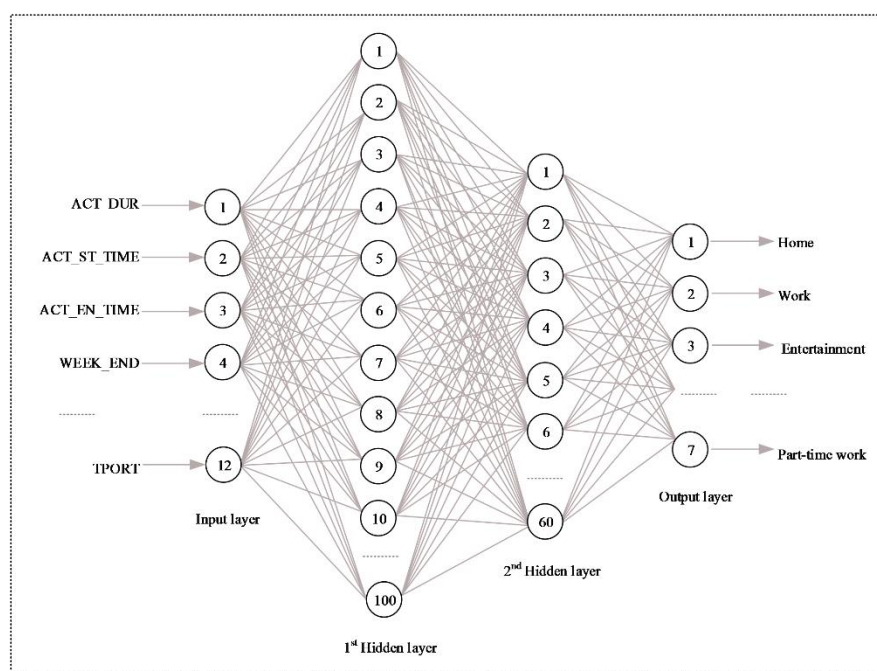
Figure 3: The structure of the ANN model for the study. (12 neurons in the input layer, seven classes as trip purposes in the output layer)

Hyper-parameters such as the number of neurons, drop rate, optimisers, activation functions, loss functions are tuned to decide the best possible parameters in the model using grid search techniques (one parameter is changed once others are unchanged) (Brownlee, 2020b)

3. Output softmax layer: The output layer is activated using the softmax as the last activation function to distribute the probability throughout each output class. The result of the given input feature is presented as the high probability value for predicting the output class.

As a result, the proposed model is trained with 70% of the data (training data) and tested with the rest of the dataset (30%, testing data) or other datasets.

2.2.2.2. Evaluating and validating the model performance

Validation of the model is crucial for the study, and the model evaluation is illustrated under two sub-sections. The first approach of evaluating model performance is achieved under three sub-categories (1) evaluating the model performance with three measures presented such as precision, recall, and F1-score (Brownlee, 2020a), (2) plotting the confusion matrix to illustrate the prediction performance for each class independently, and (3), comparing the effectiveness of the model to other baseline models using cross-validation.

The second approach of the validation focusses on the comparison of the accuracy obtained from the highest probability of land-use information, which has been widely used in literature (Gong, L., X. Liu, L. Wu, 2016; Alsger *et al.*, 2018; Sari Aslam *et al.*, 2020). Thus, after phase 1, we have inferred the activities from smart card data using the highest probability of POIs as a benchmark model and compared the results with the labelled survey data. The validation of activity type has been calculated as follows:

$$V_{A_T} = \frac{CA_T}{TA_{T_n}} * 100$$

Where A_T is activity type such as home, work, etc., V_{A_T} is the percentage of validated activity type, CA_T is the correctly identified activity points from labelled data using the highest probability of land-use (POIs) values and TA_{T_n} is the total number of activity type in n, which means user-count in this study. Hence, CA_T is normalised based on the available data from user-counts. Note that n also can be the number of check-ins from foursquare POIs. As a result, the accuracy for each activity type is presented in section 3.2.2.

3. Results

3.1. The result of the multiple scenarios for input features to predict trip purposes

The classification methods have the potential to examine trip purpose within travel data (Kuhlman, 2015; Alsger *et al.*, 2018). However, the representation of trip purposes in each class with the different number of data points may create class imbalance issues in ML approach (Brownlee, 2020c). For instance, almost 60% of the activities in the survey data are primary activities, and 40 % are secondary activities, which reveals that the count of each secondary activity is much lower than the count of each primary activity. Therefore, random over and under-sampling techniques are compared to unchanged values of each class to evaluate overall accuracy. In addition, the classification accuracy using different scenarios such as including and excluding land-use

attributes (with/without POIs, respectively) are also evaluated in this stage to obtain the best possible model performance. According to the results in figure 4, using random under-sampling techniques with POIs achieved an overall accuracy of almost 95%. Conversely, without POIs this number decreases almost 7% for an overall accuracy of 88%. Over-sampling techniques with POIs increased to the accuracy of the model by almost 96%, and without POIs, the accuracy was 89%. Finally, without balancing any classes (9,116 data points), the overall accuracy was 89% and 83% with and without POIs, respectively.
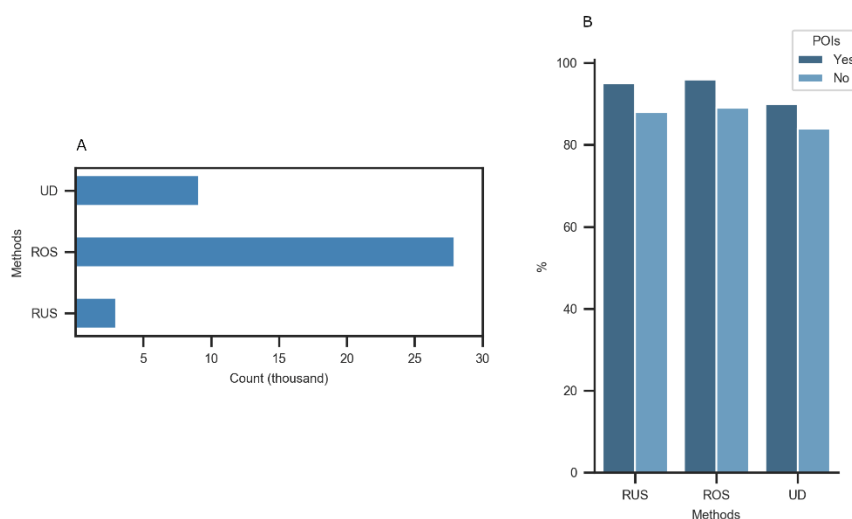
Figure 4: The representation of the data points in each method (A) and the results of overall prediction with/without POIs using unchanged data (UD), random under- and over-sampling (RUS and ROS, respectively) techniques (B).

As a result of this section, training speed using the under-sampling technique with 2,989 data points has a lower impact compared to the over-sampling technique with 27,958 data points. In addition, there is consistent 6 to 7% accuracy difference using each model with and without POIs shown in figure 4B. Therefore, the rest of the analysis is presented using random under-sampling with POIs.

3.2. The results of the validation process

There are two types of the validation process in this section, the first one is the model evaluation from test data, and the second one is the accuracy comparison from the benchmark model using the highest probability of land-use information such as POIs.

3.2.1. Evaluating the model performance

This section presents the performance of prediction under three sub-sections. First, we evaluate the models using three performance metrics in each class, such as precision, recall, and F1-score (Brownlee, 2020a). The best results in precision, recall, and F1 were attained for work activities (primary activities) and child drop-offs/pick-ups and part-time work activities (secondary activities) presented in table 3.

Table 3: Prediction performance using precision, recall, and F1-score on test data.

Trip purposes	Type of activity	Precision	Recall	F1-score
Home	Primary	0.84	0.98	0.90
Work	Activities	0.99	0.97	0.98
Entertainment		0.73	0.84	0.78
Eating	Secondary	0.74	0.76	0.75
Shopping	activities	0.75	0.62	0.68
Child drop-offs/pick-ups		0.95	0.84	0.89
Part-time (PT) workers		0.89	0.81	0.85

Then, we present the confusion matrix to clarify the prediction performance for each class independently. The confusion matrix using test data in figure 5 illustrates that the probability of a correct prediction is larger than misclassification. The lowest prediction score is for shopping activities with 17% misclassified as entertainment or eating activities. The misclassification may suggest that the temporal variation in the three activities is overlapping. For example, shorter duration shopping activities might be misclassified as eating, and longer duration shopping activities might be misclassified as entertainment. The best score among primary activities is fairly close, with 99% of home and 97% of work activities correctly predicted. The best prediction of inference among secondary activities is obtained for drop-offs/pick-ups (84%) and PT-work activities (81%) as a result of regular activity patterns. The rest secondary activities present similar outcomes with high temporal stability and regularity, such as 84% of entertainment activities, 76% of eating activities.
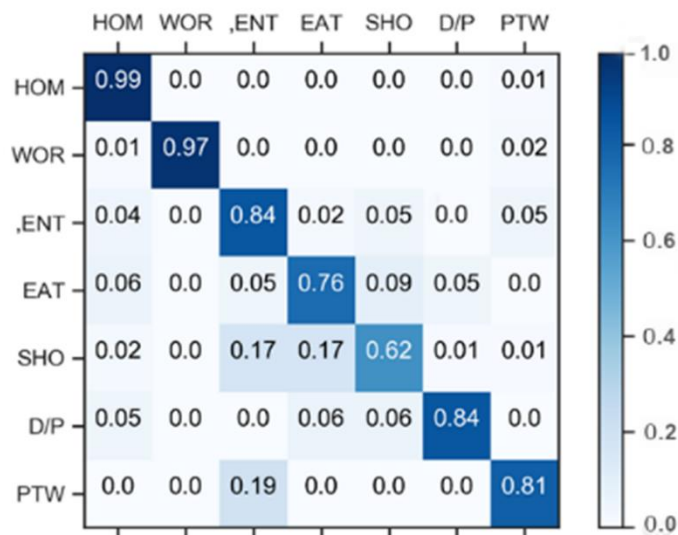
Figure 5: Inferring trip purposes using the confusion matrix with POIs. (X and y-axis are stated as predicted and actual labels, respectively.

The third one is the comparison of the model with other baseline models using 10-fold cross-validation. In this section, trip purpose prediction accuracy is compared with several baseline models such as Random Forest (RF) (Breiman, 2001), Support Vector Machine (SVM) (Cortes and Vapnik, 1995), Logistic Regression Classifier (LR) and Naïve Bayes (NB). In the existing literature, these models have been adopted for trip purpose prediction from different data sources such as GPS, phone data, but smart card data. Therefore, they are considered as baseline models to compare to the proposed model in this study.

As shown in figure 6, the original data is randomly partitioned into ten subsamples, and the highest accuracies of between 86% and 99% are achieved using neural networks with almost 13% variance. The second highest accuracy is achieved using RF with 84% to 89% with almost 5% variance. The third highest accuracy with the lowest variance is captured using SVM, with 78% to 81% accuracy. Last, LR and NB are associated with the lowest results in the analysis of cross-validation compared to other classifiers. These results support the assertion that neural networks can build computation-intensive classification with high accuracy using transport smart card data and locational POIs information with the help of data pre-processing steps.
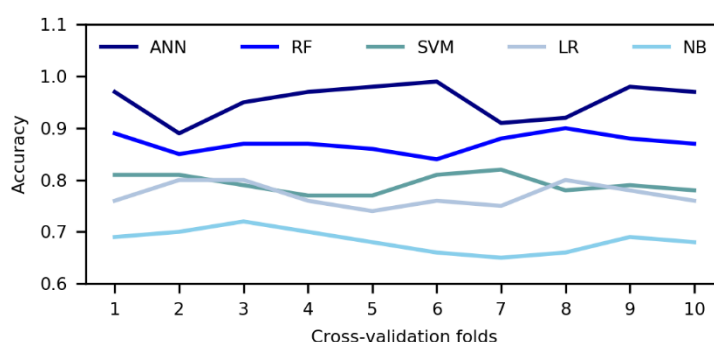
Figure 6: The accuracy of 10-fold cross-validation is presented using baseline methods.

3.2.2. Validation of the model

This section aims to compare the accuracy of the proposed framework to existing models using the highest probability of land-use information from POIs. Note that this part of the enrichment is obtained after phase 1. As a result, 51 % of work and 49 % of home activities, 44 % of entertainment, 33 % of eating, 35 % of shopping, 34 % of D/P and 39 % of PTW activities are identified as correct. As a result, the proposed ActivityNET framework demonstrates a higher success rate as compared to rule-based techniques in the literature.

The reason for the low accuracy in the heuristic approaches is that the distribution of highly mixed land-use provides lower accuracy than the distribution of single land-use such as residential or work centres, i.e. London. Besides, sophisticated techniques provide higher accuracy to predict trip purposes (Anda, Erath and Fourie, 2017).

4. Discussion

This study aims to predict trip purposes using the spatial and temporal attributes of transport data and land-use data derived from POIs with machine learning algorithms. Multiple scenarios, including spatial features with random under-sampling techniques, are helped to achieve high accuracy in the model. The model accuracy is also compared to white-box model from phase 1. As a result, the proposed framework, ActivityNET, is able to detect trip purposes using machine learning techniques as a single model with improved accuracy.

Using new big data sources such as smart card data and POIs provides an excellent opportunity to explain where, when, and why people spend their time within urban settings. Both data sources have great opportunities such as investigating human mobility, urban flow and trip purposes with some limitation. For instance, smart card data may suffer from demographic details of passengers' (Zhang, Cheng, and Sari Aslam 2019; Zhang, Sari Aslam and Cheng, 2020), recording destination information for bus users (Gordon et al. 2013), and the trip purpose of the travellers, investigated further using land-use attributes such POIs. Similarly, regardless of the wide range of positive characteristics of POIs such as quantifying the weight of the place using counts, using working hours of POIs to present dynamics of the activity patterns in cities, POIs from foursquare data may suffer from over-representing of some of the locations, which uses a small number of users with substantial counts in restaurant or shopping centres as compared to workplaces (Rashidi *et al.*, 2017). In addition, demographic biases in the dataset is an inevitable fact that the application is mainly used by younger age groups, e.g. less than 30 years old, as compared to older age groups in the cities (Longley and Adnan, 2016).

Even though the proposed framework provides high prediction accuracy compared to other ML models, trip purpose detection inherently involves uncertainty (Xiao, Juan, and Zhang, 2016; Faroqi, Mesbah and Kim, 2018) in terms of temporal and spatial similarities in the dataset. For instance, long hours of shopping activity, which may be disturbed by eating activity (drinking coffee/tea) at a location in which both shopping and eating places available. Although it is difficult to separate those activities in the daily life of individuals, survey data do not have any multiple activities as labelled data for the analysis. Therefore, we assume that this is not an issue for the proposed framework.

Moreover, this study also shows a comparison between what-if scenarios and ML approaches. The analysis demonstrates that the highest probability of activity type is depended on the distribution of land-use, which means the distribution of highly mixed land-use provides lower accuracy than the distribution of single land-use such as residential or work centres, e.g. London. In addition, the land-use information has limitations to represent primary locations. On the other hand, the complex sequential relationship between spatial and temporal features can be captured by ML approach with high accuracy to predict trip purposes. Hence, there is a potential to create individual travel diaries using the results of the ActivityNET framework as an alternative method in an urban setting for the travel demand research.

5. Conclusion

The availability of big data sources such as smart card data and POIs provide a great opportunity to produce new insights into transport demand modelling. This study aims to predict trip purposes in a feasible framework using the spatial and temporal attributes of transport data, and urban functions derived from POIs to generate an understanding of human mobility and urban flow in cities.

The proposed framework, ActivityNET, is demonstrated that neural network provides improve accuracy in trip purpose prediction. First, the framework focusses on the proposed "activity-POIs consolidation algorithm" which combines travel behaviour with socio-functional information from POIs under three sub-groups such as activity characteristics (activity start and end time, activity duration), day characteristics, and land-use characteristics. Second, the framework illustrates ANN model to predict trip purposes under primary (home and work) and secondary activities (entertainment, eating, shopping, child drop-offs/pick-ups, and part-time work activities) with high accuracy. Third, the proposed framework, ActivityNET, is applied as a case study in London and achieved 95% overall accuracy using random under-sampling techniques with POIs. In addition, high accuracy for primary activities, 99% for home and 97% for work are obtained from smart card data. Furthermore, secondary activities provide improved accuracies such as 84% of entertainment, 84% of drop-offs/pick-ups, 81% of PT-work, 76% of eating activities, and 62 % of shopping activities. Validation results using benchmark method showed that the framework is provided improved accuracy. As a result, ActivityNET offers trip purpose prediction with high accuracy for transport and urban planners. Future work will focus on creating travel diaries using the results of ActivityNET framework as a cost-effective alternative approach for travel demand research.

Data availability statement

The data that support the findings of this study can be found in these links. The first one is Travel data (Oyster card data), which can be downloaded from (https://data.gov.uk/dataset/c5b74d3f-8bf1-443c-8f2d-bd307720737f/underground-stations). Note that the user needs to label the journey data after extracting activities. Second, London stations data can be downloaded from (https://api-portal.tfl.gov.uk/). As of last, Foursquare POIs data (user-counts, opening/closing hours) can be found here (https://developer.foursquare.com/docs/places-api/endpoints/). Besides, data are available from the authors upon reasonable request (DOI: 10.5281/zenodo.4527765).

Additional information

Funding

This work was funded by the Economic and Social Research Council [1477365] as part of the Consumer Data Research Centre (CDRC) project (ES/L011840/1).

Notes on contributors

Nilufer Sari Aslam is a PhD candidate at SpaceTimeLab for Big Data Analytics, University College London, UK. Her research interests are data mining techniques, spatial-temporal big datasets, urban and transport planning.

Mohamed R Ibrahim is a PhD candidate at SpaceTimeLab for Big Data Analytics, University College London, UK. His research interests are artificial intelligence, urban modelling and geoinformatics.

Tao Cheng is a professor at SpaceTimeLab for Big Data Analytics, University College London, UK. Her research interests include space-time analytics and big data mining with applications in transport modelling.

Huanfa Chen is currently a teaching fellow in spatial data science (Centre for Advanced Spatial Analysis), University College London, UK. His current research interests include geospatial machine learning, agent-based modelling and spatial optimisation.

Yang Zhang is currently a post-doc researcher at SpaceTimeLab for Big Data Analytics, University College London, UK. Her research interests include urban computing, deep learning, and artificial intelligence.

References

Alsger, A. *et al.* (2018) 'Public transport trip purpose inference using smart card fare data ☆', *Transportation Research Part C*. Elsevier, 87, pp. 123–137. doi: 10.1016/j.trc.2017.12.016.

Anda, C., Erath, A. and Fourie, P. J. (2017) 'Transport modelling in the age of big data', *International Journal of Urban Sciences*. Taylor & Francis, 21(October), pp. 19–42. doi: 10.1080/12265934.2017.1281150.

Breiman, L. E. O. (2001) 'Random Forests', *Machine Learning*, 45, pp. 5–32.

Brownlee, J. (2020a) *How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification*, *Machine Learning Mastery*. Available at: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/.

Brownlee, J. (2020b) *How to Grid Search Hyperparameters for Deep Learning Models in Python With Keras*, *Machine Learning Mastery*. Available at: https://machinelearningmastery.com/grid-search-hyperparameters-deep-learning-models-python-keras/ (Accessed: 4 January 2021).

Brownlee, J. (2020c) *Random Oversampling and Undersampling for Imbalanced Classification*, *Machine Learning Mastery*. Available at: https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/ (Accessed: 1 October 2020).

Chakirov, A. and Erath, A. (2012) 'Activity Identification and Primary Location Modelling based on Smart Card Payment Data for Public Transport', *13th International Conference on Travel Behaviour Research Toronto*, (July).

Chaniotakis, E., Antoniou, C. and Pereira., F. (2016) 'Mapping Social Media for Transportation Studies', *IEEE Intelligent Systems*. IEEE, 31(6), pp. 64–70. doi: 10.1109/MIS.2016.98.

Cortes, C. and Vapnik, V. (1995) 'Support-Vector Networks', *Machine Leaming*, 20, pp. 273–297.

Dacheng, C. *et al.* (2018) 'Traveler Segmentation using Smart Card Data with Deep Learning on Noisy Labels', 18. doi: 10.1145.

Dahl, G. E., Sainath, T. N. and Hinton, G. E. (2013) 'Improving deep neural networks for LVCSR using rectified linear units and dropout', *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. IEEE, pp. 8609–8613. doi: 10.1109/ICASSP.2013.6639346.

Devillaine, F., Munizaga, M. and Trépanier, M. (2012) 'Detection of Activities of Public Transport Users by Analysing Smart Card Data', *Transportation Research Record: Journal of the Transportation Research Board*, 2276(3), pp. 48–55. doi: 10.3141/2276-06.

Faroqi, H., Mesbah, M. and Kim, J. (2018) 'Applications of transit smart cards beyond a fare collection tool : A literature review', *Advances in Transportation Studies*, 45(July), pp. 107–122. doi: 10.4399/978255166098.

Glorot, X. and Bordes, A. (2011) 'Deep Sparse Rectifier Neural Networks', in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 315–323.

Gong, L., X. Liu, L. Wu, Y. L. (2016) 'Inferring trip purposes and uncovering travel patterns from taxi trajectory data', *Cartography and Geographic Information Science*. Taylor & Francis, 43(2), pp. 103–114. doi: 10.1080/15230406.2015.1014424.

Goodfellow, I., Bengio, Y. and Courville, A. (2017) *Deep Learning (Adaptive Computation and Machine Learning series)*. London, England: The MIT Press. Available at: https://lccn.loc.gov/2016022992.

Goulet-langlois, G., Koutsopoulos, H. N. and Zhao, J. (2016) 'Inferring patterns in the multi-week activity sequences of public transport users', *Transportation Research Part C*. Elsevier Ltd, 64, pp. 1–16. doi: 10.1016/j.trc.2015.12.012.

Hinton, G. E. *et al.* (2012) 'Improving neural networks by preventing co-adaptation of feature detectors', in, pp. 1–18.

Ibrahim, M. *et al.* (2020) 'Variational-LSTM Autoencoder to forecast the spread of coronavirus across the globe', *MedRxiv*. doi: 10.1101/2020.04.20.20070938.

Ibrahim, M. R. *et al.* (2019) 'predictSLUMS: A new model for identifying and predicting informal settlements and slums in cities from street intersections using machine learning', *Computers, Environment and Urban Systems*, 76, pp. 31–56. doi: 10.1016/j.compenvurbsys.2019.03.005.

ICO (2018) *Guide to the General Data Protection Regulation (GDPR)*. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/.

Kingma, D. P. and Ba, J. L. (2015) 'Adam: A method for stochastic optimisation', in *ICLR*. Online, pp. 1–15.

Lee, S. G. and Hickman, M. (2014) 'Trip purpose inference using automated fare collection data', *Public Transport*, 6(1–2), pp. 1–20. doi: 10.1007/s12469-013-0077-5.

Longley, P. A. and Adnan, M. (2016) 'Geo-temporal Twitter demographics Geo-temporal Twitter demographics', *International Journal of Geographical Information Science*. Taylor & Francis, 30(2), pp. 369–389. doi: 10.1080/13658816.2015.1089441.

Longley, P., Cheshire, J. and Singleton, A. (2018) *Consumer Data Research*. London: UCL Press. Available at: https://discovery.ucl.ac.uk/id/eprint/10046615/1/Consumer-Data-Research.pdf.

Nakamura, K. *et al.* (2016) 'Failure of Transit-Oriented Development in Bangkok from a Quality of Life Perspective', *Asian Transport Studies*, 4(1), pp. 194–209. doi: 10.1590/S0100-72032012000400008.

Rashidi, T. H. *et al.* (2017) 'Exploring the capacity of social media data for modelling travel behaviour : Opportunities and challenges q', *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd, 75, pp. 197–211. doi: 10.1016/j.trc.2016.12.008.

RTPI (2018) *How Far is it Acceptable to Walk ?* Available at: https://www.rtpi.org.uk/media/2739252/wyg_gareth_pdf.pdf.

Sari Aslam, N. *et al.* (2020) 'Semantic enrichment of secondary activities using smart card data and point of interests : a case study in London', *Annals of GIS*. Taylor & Francis, pp. 1–13. doi: 10.1080/19475683.2020.1783359.

Sari Aslam, N. and Cheng, T. (2018) 'Smart Card Data and Human Mobility', in Longley, P., Cheshire, J., and Singleton, A. (eds) *Consumer Data Research*. London,UK: UCL Press, pp. 111–119. Available at: https://www.jstor.org/stable/j.ctvqhsn6.11.

Sari Aslam, N., Cheng, T. and Cheshire, J. (2019) 'A high-precision heuristic model to detect home and work locations from smart card data', *Geo-spatial Information Science*. Taylor & Francis, 22(1), pp. 1–11. doi: 10.1080/10095020.2018.1545884.

TfL (2019) *Out-of-station interchanges*. Available at: https://tfl.gov.uk/corporate/publications-and-reports/out-of-station-interchanges (Accessed: 26 September 2019).

Xiao, G., Juan, Z. and Zhang, C. (2016) 'Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimisation', *Transportation Research Part C: Emerging Technologies*. Elsevier Ltd, 71, pp. 447–463. doi: 10.1016/j.trc.2016.08.008.

Yang, Y. *et al.* (2019) 'Who , Where , Why and When ? Using Smart Card and Social Media Data to Understand Urban Mobility', *ISPRS International Journal of Geo-Information*, 8(6), p. 271. doi: 10.3390/ijgi8060271.

Zhang, Y. *et al.* (2020) 'You are how you travel : A multi-task learning framework for Geodemographic inference using transit smart card data', *Computers, Environment and Urban Systems*. Elsevier, 83(June), p. 15. doi: 10.1016/j.compenvurbsys.2020.101517.

Zou, Q. *et al.* (2016) 'Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway', *Transportation*. Springer US, (3). doi: 10.1007/s11116-016-9756-9.

Zhang, Y., Cheng, T. and Sari Aslam, N. (2019) 'Exploring the relationship between travel pattern and social - demographics using smart card data and household survey', in *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Enschede, The Netherlands, pp. 10–14.