Inferring Trip Purposes from Travel Smart Card Data

Nilufer Sari Aslam

SpaceTimeLab for Big Data Analytics

Department of Civil, Environmental and Geomatic Engineering

University College London

A thesis submitted for the degree of

Doctor of Philosophy in Geographic Information Science (GIS)

May 2022

STATEMENT OF ORIGINALITY

I, Nilufer Sari Aslam, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis

Signed:

Date:

ABSTRACT

Understanding human mobility, activities and trip purpose is essential for transport planning and commercial services in urban environments. Conventionally, household travel surveys have collected such information, but these have relatively small sample sizes with low update frequencies leading to an incomplete and inaccurate picture of overall urban mobility. Recently, big data sources such as Smart Card Data (SCD) have provided alternative sources from which to investigate human mobility due to their longitudinal nature, large volume and high level of spatiotemporal detail. While Primary Activities (PAs) (home and work/school for adults/students) are relatively easy to identify from SCD, Secondary Activities (SAs) (the rest of a person's activities) and trip purposes are difficult to understand directly given only limited information provided by SCD, e.g., the tap-in/out station and time.

The aim of this thesis is to investigate activities and trip purposes from large SCD by making use of survey data and Point-of-Interest (POI) data using both heuristic and machine learning approaches. The heuristic approach uses a set of rules/algorithms to identify PAs and SAs based upon journey counts, visit frequency, activity duration and direction information (from and to) extracted from SCD combined with spatiotemporal attributes of POIs. The machine learning approach, however, uses a deep learning framework, ActivityNET, to derive the trip purpose of both PAs and SAs from labelled SCD data contributed by volunteers (survey SCD, or SSCD) in one step. The proposed models are validated using London Travel Demand Survey (LTDS) data and SSCD, achieving higher detection accuracy than benchmark methods and the deep learning methods are found to be more accurate than the heuristic approach.

Using volunteers' survey data (SSCD) with their associated SCD is innovative in this study which overcomes the challenges in using machine learning for trip purpose detection and validation. The study has developed a cost-effective framework to use big data sources (SCD and POIs) for urban mobility analysis, which has strong policy implications. In addition, the study provides new ideas for future applications to help or eliminate conventional household surveys for travel demand analysis.

IMPACT STATEMENT

Information regarding trip purposes is essential to engage in planning, development and performance evaluations of public transit networks and services. The scope of the extant research includes myriad applications, such as transportation, urban mobility, retail analysis, environment and public health. Traditionally, researchers have relied on household surveys to collect information regarding trip purposes and estimate current and future demand to support the enhancement of the transportation system under various socio-economic scenarios and land use structures. Seeking to capitalise on the availability of big data sources and rapid improvements in computational power, this thesis designs a systematic framework to develop state-of-the-art trip purpose inference models using big data sources for public transport networks.

This research fills several critical gaps in the literature from a theoretical point of view. First, under the heuristic approach, the proposed models, i.e., heuristic primary (PAs) and secondary activity (SAs) identification models use the characteristics of smart card data (SCD). This part of the research also improves efforts to identify the types of SAs by combining SCD with land use points of interest (POIs) to infer trip purposes in large cities. The proposed PAs and SAs identification models that benefited transport research have been published in the journal Geo-spatial Information Science (2019) and Annals of GIS (2021). Second, under the machine learning (ML) approach, the proposed ActivityNET framework has contributed to academic literature using a deep learning model to infer trip purposes from SCD, capturing uncertainty and spatial dependencies. The framework has been published in the journal Geo-spatial Information Science (2021). Moreover, to overcome the current challenges of using an ML approach to detect and validate trip purposes, this study relies on a new data source - survey smart card data (SSCD) from volunteers. As the first survey data from SCD, SSCD facilitates the model's evaluation due to its longitudinal similarity to the source SCD.

Regarding applications of this research, it has the potential to benefit the fields of transport and urban planning. For example, developers may build the proposed methodologies into a software programme to accurately and automatically predict trip purposes using SSCD collected through web- or phone-based online applications. Such intelligent trip purpose prediction applications could enhance travel demand surveys and thus support the work of transport planners.

ACKNOWLEDGEMENTS

This PhD research is funded by the Economic and Social Research Council (Grant Number: 1477365), cooperated by the Consumer Data Research Centre. It would not have been possible to do without the support, guidance and company that I have welcomed people around me.

First and foremost, I have to thank my first supervisor, Prof Tao Cheng, for all the guidance and support. Tao inspired me with new ideas in my research. Her motivation and wise supervision helped me focus on my work during this journey. Her encouragement at work and in life will always stay in my memories. Besides, I thank my second supervisor, Prof James Cheshire, for his valuable comments, suggestions and discussions for my research.

In addition, I would also like to thank Dr Cheng Zhong and Prof Alison Heppenstall for attending as my PhD Viva examiners. They created a friendly environment during the Viva and provided thoughtful remarks and recommendations, which assisted me in improving my thesis.

Other special thanks go to my colleagues/friends at SpaceTimeLab: Dr James Haworth, Dr Sarah Wise, Dr Aldo Lipani, Dr Huanfa Chen, Dr Juntao Lai, Dr Di Zhu, Dr Yang Zhang, Dr Mohamed R Ibrahim, Davide Di Gioia, Tongxin Chen and Guangsheng Dong, including my friends Dr Athanasios Bantis, Selda Gundogan and Astrid Caballero. Your friendships and support during this research are unforgettable.

Most importantly, this journey would not have been possible without the support of my family, Kamran, Alina and Eren Aslam, including Nurcan, Necdet, Kubra and Ugur Sari. Thank you for being always with me during this journey.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AFCs:	Automatic Fare Collection Systems
ANN:	Artificial Neural Network
AW:	After-work activities
BW:	Before-work activities
DL:	Deep Learning
D/P:	Drop-offs/Pick-ups activities
ENT:	Entertainment activities
EAT:	Eating activities
H:	Home activities
LTDS:	London Travel Demand Surveys
ML:	Machine Learning
MD:	Midday activities
OD:	Origin–Destination
PAs:	Primary Activities (home and work/school for adults/students)
POI:	Point-of-Interest
PTW:	Part-time activities
SC:	Smart Card
SCD:	Smart Card Data
SSCD:	Survey Smart Card Data
SAs:	Secondary Activities (SAs) (the rest of a person’s activities)
SHO:	Shopping activities
UD:	Undefined activities
W/S:	Work/Study activities for adults/students, respectively

GLOSSARY OF TERMS

Concepts	Definition
Activity	An activity is time spent between two consecutive trips as obtained from SCD.
Activity-based model	Activity-based models are travel demand models for the transport network, which drive demands from individuals' daily activities/patterns. The relevant information, e.g., when, where, how long and mode of transport for activities, was extracted from travel surveys in activity-based models.
ActivityNET	ActivityNET is a framework that uses large data sources, i.e., SCD and land use data from Foursquare POIs, along with deep learning techniques to predict trip purposes for public transport.
Alighting/Exit/Destination station of the trips	Leaving public transport at a stop or station
Artificial Neural Network (ANN)/ Neural Networks (NN)	Artificial Neural Networks (ANNs) or Neural Networks (NN) are computational models in ML techniques with a novel structure inspired by biological nervous systems, such as the brain.
Automatic Fare Collection Systems (AFCs)	Automatic Fare Collection (AFC) Systems are digitised ticketing systems (with reusable cards) operating RF-ID (Radio Frequency ID) to replace paper tickets in public transport.
Boarding/Enter/Origin station of the trips	Entering a public transport at a stop or station
Deep Learning (DL)	Deep learning (DL) is a part of ML techniques established on ANN with three or more layers.
Euclidean distance	Measuring distance using a straight line between two spatial points.
London Travel Demand Survey (LTDS)	Household surveys were carried out by Transport for London (TfL).

Machine Learning (ML)	ML is a technique by which machines can learn from data without using simple or complex rules.
Point-Of-Interest (POI)	A POI is a geographic location that may be useful or provide interesting information for people, e.g., a restaurant, hotel, or tourist attraction.
Primary Activities (PAs)	Activities at key (anchor) stations in public transport • home (H) • work/school for adults/students (W)
Secondary Activities (SAs)	The rest of the activities at a station in public transport • before-work (BW) • midday (MD) • after-work (AW) • undefined (UD) activities
Segregation	The unequal distribution of the different locations/activity types in the urban environment includes residential areas, workplaces and other locations.
Smart Card (SC)	A plastic card that contains a chip to store information.
Smart Card Data (SCD)	Storing individuals' travel information in public transport.
Tap-in	SC transaction upon entering public transport
Tap-out	SC transaction upon exiting public transport
Trip/Journey	A one-way journey from one station (origin) to another (destination) using the public transport network in an individual's daily travel.
Trip purposes/ Activity types	Trip purposes are derived information from extracted activities, either data mining or processing steps using the characteristics of SCD or additional information, i.e., land use attributes in the proximity of start/end stations • Entertainment activities (ENT) • Eating activities (EAT) • Shopping activities (SHO)

	• Drop-offs/Pick-ups activities (D/P) • Part-time activities (PTW) • Other activities (O)
Urban mobility	Moving people from one station (origin) to another station (destination) using the public transport network in urban areas.

Chapter 1

Introduction

1 INTRODUCTION

1.1 Background and Motivation

Automated fare collection systems (AFCs), passively collect large Smart Card Data (SCD), offer a valuable source of information on human mobility in cities, in that passengers' daily tap-in/-out transactions in public transport networks provide spatial (i.e., start/end locations) and temporal (i.e., start/end times) details on an unprecedented scale (Pelletier, Trepanier and Morency, 2011). Longitudinal SCD on a large scale are thus considered a valuable source of big data. Data mining methods of classification and clustering are well suited to the analysis of such large mobility datasets. Therefore, transport and urban planners and researchers have devoted considerable attention to the processing and analysis of SCD (Pelletier, Trepanier and Morency, 2011).

There are comprehensive benefits to using large transit datasets in public transport research, as such datasets are relevant to a variety of topics, including transit service quality and reliability (Uniman et al., 2010; Bagherian et al., 2016), route choice and demand modelling (Chu and Chapleau, 2008; Viggiano et al., 2017), mobility analyses of travel patterns (Liu et al., 2009; Yuan et al., 2013; Sari Aslam, 2015; Zhong et al., 2016; Ma et al., 2017; Sari Aslam, Cheng and Cheshire, 2018; Sari Aslam and Cheng, 2018; Zhang, Cheng and Sari Aslam, 2019; Yang et al., 2019; Zhao, Koutsopoulos and Zhao, 2020; Liao, 2021), Origin–Destination (OD) estimation (Cui, 2006; Seaborn et al., 2009; Nassir et al., 2011; Munizaga and Palma, 2012; Padinjarapat and Mathew, 2013; Alexander et al., 2015; Alsger, 2016; Kumar, Khani and He, 2018; Hussain, Bhaskar and Chung, 2021), activity identification (Bouman et al., 2013; Nassir, Hickman and Ma, 2015; Goulet-Langlois, Koutsopoulos and Zhao, 2016; Ectors et al., 2017; Zhi et al., 2017; Ordóñez Medina, 2018) and trip purpose identification (Devillaine, Munizaga and Trepanier, 2012; Lee and Hickman, 2014; Kusakabe and Asakura, 2014; Zou et al., 2016; Han and Sohn, 2016; Alsger et al., 2018; E. Kim, Y. Kim and D. Kim, 2020; Faroqi and Mesbah, 2021).

Understanding the reasons for passengers' trips or their trip purposes is essential for planning, evaluating and developing public transit networks and services (Faroqi, Mesbahs and Kim, 2018). Besides, existing research on trip purposes has provided valuable insight for many applications and analysed urban mobility and people flows for city planners (Yang et al., 2019) and consumer behaviour for commercial establishments (Longley, Cheshire and Singleton, 2018). Moreover, scholars have developed environmental measurements to inform transport infrastructure and developments (Zheng et al., 2014), while policy- and decision-makers have conducted public health investigations, e.g., regarding the impact of COVID-19 lockdowns and ways to improve transport services during the post-pandemic recovery period (Caicedo, Walker and González, 2021).

Typically, researchers have collected information on trip purposes via household surveys. However, in neglecting to utilise big data sources such as SCD, the research in trip purposes has missed an opportunity to reveal individuals' spatiotemporal activity patterns as a sequence of activity locations, activity start and end times, activity durations and indications of land use in the proximity of the alighting or boarding station, which can be further explored with data mining techniques to determine travellers' trip purposes from SCD (Faroqi, Mesbah and Kim, 2018). Recognising this potential, this study aims to understand not only where and when but also why individuals move in urban environments. In this way, the study opens new opportunities for future research in urban environments.

This thesis, therefore, explores the ways in which SCD can be used to infer trip purposes in the context of urban mobility by applying data mining techniques to public transit networks. The study relies on two approaches to infer trip purposes: 1) the heuristic approach, which utilises the characteristics of SCD as prior information to inform what-if scenarios and 2) the ML approach, which identifies patterns in large SCD without relying on pre-defined rules. Further, this study collects data from a novel source – SSCD from volunteers – to overcome the current challenges of using ML to detect and validate trip purposes.

Following Section 1.2 examines the research challenges and gaps in the extant literature. Section 1.3 then illustrates the aims and objectives of this research to resolve the existing research gaps. Finally, Section 1.4 describes the overall thesis structure.

1.2 Challenges and Gaps

Recent advancements in information and communication technologies have highlighted opportunities for incorporating new mobility data, namely SCD, to understand individuals' activity types. SCD provides valuable details on public transport networks due to their longitudinal nature and high spatial and temporal resolution, allowing to address transport and urban issues. On the other hand, modelling attempts increase methodological challenges due to the unlabelled nature of passively obtained mobility data. Moreover, trip purposes and demographic characteristics are missing information (Li et al., 2018) and enrichment is required for trip purposes due to the limited spatial information from SCD, as only the tap-in/-out stations are available.

Household surveys have dominated research in terms of travel behaviour attributes (e.g., trip purpose, transportation mode, etc.) and demographic characteristics (e.g., income, gender, age etc.). While attempting to represent an in-depth view of individual's mobility behaviour in the transport network, some studies use travel surveys to enrich SCD and overcome the limited spatial information it provides (Chakirov and Erath, 2012; Medina and Erath, 2013; Lee and Hickman, 2014; Amaya, Cruzat and Munizaga, 2018; Alsger et al., 2018; E. Kim, Y. Kim and D. Kim, 2020; Faroqi and Mesbah, 2021). However, relying on conventional household surveys to inform trip purpose models may produce several additional limitations. First, survey data are limited by infrequent updates (usually once each year for travel surveys) and small sample sizes (e.g., data for a single day). Inferences of activity types/trip purposes based upon one-day survey data (Chakirov and Erath, 2012; Amaya, Cruzat and Munizaga, 2018; Alsger et al., 2018) may not be sufficiently generalisable to represent the full range of activities of the whole population. Second, surveys may be limited to identifying transport users' home and work locations – as is the case for the London Travel Demand Survey (LTDS) – which may not fully account for the rest of their activities (i.e., SAs, such as eating and entertainment). Third, the mismatch period between surveys and SCD may result in biases and decrease model accuracy. Finally, such survey data may not be available for many cities, especially in developing countries (Amaya, Cruzat and Munizaga, 2018). Thus, there is a need to investigate the scope and limitations of emerging data sources to investigate trip purposes from SCD, which will open up new opportunities in travel demand research (Anda, Erath and Fourie, 2017; Liao, 2021).

While some studies identified primary locations and activities from SCD (Devillaine, Munizaga and Trepanier, 2012; Hasan et al., 2012; Wei, Liu and Sigler, 2016; Han and Sohn, 2016; Yuan, Winter and Wang, 2019; Zhao, Koutsopoulos and Zhao, 2020), 'other' activities, or secondary activities (SAs) are rarely explored in transport research

(Alsger et al. 2018; E. Kim, Y. Kim and D. Kim, 2020, Faroqi and Mesbah, 2021). Thus, a more accurate representation of SAs needs to be explored using SCD due to the limited activity types and low accuracy under the existing approaches.

With this in mind, from a methodological perspective, studies inferring trip purposes from SCD relied primarily on a heuristic approach based on prior information regarding commuters' lifestyles. When researchers enriched SCD with land use attributes, the heuristic approach demonstrated a limited capability to represent various activities such as SAs and could not achieve high accuracy due to the spatial complexity involved with this combination of data. To handle such complexity and uncertainties, therefore, scholars must employ deep learning (DL) techniques (Anda, Erath and Fourie, 2017), which are flexible enough to capture uncertainties with accurate outcomes in complex cities (Xiao, Juan and Zhang, 2016; Anda, Erath and Fourie, 2017). In addition, DL models can use the input data without feature engineering and automatically capture the latent features to represent underlying patterns in the data and generate the desired output. Nevertheless, efforts to adopt an ML approach face some limitations, which include but are not limited to the following:

1) First, the noise in unprocessed smart card data requires pre-processing steps before applying prediction models to achieve high accuracy (Dacheng et al. 2018; Zhang et al. 2020).
2) Second, aggregated input features per user from a large volume of travel data, such as average travel duration and average departure times of first and last trips (Goulet-Langlois, Koutsopoulos and Zhao, 2016; Han and Sohn, 2016; Zhang, Cheng and Sari Aslam, 2019), may not accurately represent activity points. Thus, a disaggregated level of analysis to infer trip purposes is necessary.
3) New data sources need to be evaluated to overcome the current challenges when using ML to detect and validate trip purposes.

Placing uncertainty related to data limitations and modelling processes, there is a requirement for a modelling approach inferring trip purposes from SCD (Anda, Erath and Fourie, 2017; Faroqi, Mesbahs and Kim, 2018; Zannat and Choudhury 2019). Thus, this study conceptualises and develops two methodological frameworks to infer trip purposes from SCD. Both methods are valuable for many transport- and urban-related applications due to their ability to identify PAs and SAs. In addition, the use of collected

SSCD is innovative due to its longitudinal similarity to SCD, which will result in more accurate outcomes using the ML approach.

1.3 Research Aims and Objectives

This research aims to develop innovative frameworks that utilise large and longitudinal individual smart card (SC) datasets to investigate public transport users' trip purposes to help in planning, evaluating and developing public transport. Its findings will inform practical, large-scale applications, such as research into human mobility, travel behaviour and transport planning. The study focuses on the travel and associated activities of individuals from SCD to achieve the following research objectives:

Objective 1: Investigate the scope and limitations of emerging machine automated big data sources as an alternative to travel surveys to estimate travel demand.

Objective 2: Investigating mainly 'other' activities, or SAs, using the characteristics of SCD with the help of land use attributes, i.e., POIs.

Objective 3: Develop a trip purpose inference framework using the heuristic and ML approaches.

Objective 4: To overcome the limitation of the ML approach, collect a new data source, i.e., survey SCD (SSCD) to investigate trip purposes using SCD.

Objective 5: Apply the proposed models and methodologies to various scenarios in a large city, i.e., London.

1.4 Thesis Organisation

This section presents the organisation of the thesis, which comprises eight main chapters. Each chapter contributes to the aim of the thesis by building upon previous and current research methods to develop the proposed frameworks and identify future research directions that utilise SCD to infer trip purposes. The details of the chapters are as follows:

The introduction, Chapter 1, summarises the challenges and gaps in the existing literature. In doing so, it defines the research aims, objectives and organisation of this dissertation.

Chapter 2 reviews the literature on inferring trip purposes from SCD. It begins by identifying the relevant research domains as well as conventional and current possible data sources for obtaining trip purposes. Then the chapter systematically reviews the trip purposes literature to identify and summarise the gaps in the prevailing methodologies, including input features and methods from SCD.

Following the literature review, Chapter 3 describes the methodologies proposed by this study to fill the research gaps, including the thesis methodological framework, data collection, database structure and the proposed methods for inferring users' trip purposes from SCD.

Chapter 4 describes the data sources used in this study. First, Section 4.1 details the study area, London's public transport network, which includes the city's train and tube systems. Then, Section 4.2 examines the potential data sources and their characteristics as well as the advantages and disadvantages of the data sources employed in this study.

Chapter 5 goes on to explain the data pre-processing steps – for example, combining SCD with land use attributes, i.e., POIs – which are necessary to extract the characteristics of SCD for use in the study's subsequent analyses. Then, the chapter defines the relevant characteristics of the data sources, i.e., journey count, start and end stations, visit frequency, activity duration, direction (from/to), opening and closing hours, activity type and the number of check-ins.

Targeting the prevailing gaps in the existing methods for inferring trip purposes using SCD, Chapter 6 presents frameworks of activity identification from SCD using a heuristic approach, i.e., identification of PAs and SAs. The chapter applies the proposed framework to a case study in London and then validates and compares the heuristic model to the benchmark models.

Chapter 7 presents a new framework to improve the accuracy of trip purpose identification using an ML approach. The chapter focuses on investigating PAs and SAs in a single model using SSCD and POIs. The chapter concludes by presenting, validating and discussing the results as well as the scope and limitations of the data sources and proposed frameworks.

Finally, Chapter 8 summarises the conclusions and contributions of this thesis, aligning them to the study's aim and objectives. The study's limitations and directions for future research follow by the concluding remarks

Chapter 2

Literature Review

2 LITERATURE REVIEW[1]

This chapter starts with the relevant research domains to highlight the motivation of the study in Section 2.1. Section 2.2 briefly explains conventional data sources and new data sources used to gather trip purposes. Section 2.3 systematically reviews the literature on trip purpose inference from Smart Card Data (SCD) to represent input features and enrichment of SCD. Section 2.4 focuses on the applied methods, i.e., heuristic, ML-based and statistical approaches. The gaps in the literature were then highlighted in Section 2.5. Finally, Section 2.6 briefly summarises the content of this chapter.

2.1 Relevant research domains

Trip purpose – in other words, 'the reason for a trip' – is a key attribute in transport research used for planning purposes, performance evaluation and the development of public transit networks and services (Faroqi, Mesbah and Kim, 2018). In addition, such information provides valuable insight for many applications, including human mobility (Sari Aslam, Cheng and Cheshire, 2018; Yang et al., 2019), urban planning (Delhoum et al., 2020), environment assessment (Beckx et al., 2013), retail analysis (Lambiri, Faggian and Wrigley, 2017) and public health, such as for COVID-19 (Caicedo, Walker and González, 2021). Therefore, in this section, a preliminary discussion of possible applications highlights the motivations of this work from the perspective of practical use, mainly in transport and urban planning-related fields, as detailed below.

2.1.1 Transportation

This study benefits from individuals' daily lifestyles, including the time, duration and location of activities, to design transport network strategies and model transit usage, demand and capacities (Goulet-Langlois, 2016; Alsger et al., 2018). For instance, Ali, Kim and Lee (2016) focused on activity characteristics from SCD as inputs to produce microsimulation travel demand models using MATSim and Delhoum et al. (2020) used

[1] Part of this chapter has been presented in the following publications:
[1] N Sari Aslam, T Cheng, J Cheshire 2019. A high-precision heuristic model to detect home and work locations from SCD. *Geo-spatial Information Science* 22 (1), 1-11.
[2] N Sari Aslam, D Zhu, T Cheng, MR Ibrahim, Y Zhang 2020. Semantic enrichment of secondary activities using SCD and points of interests: a case study in London. *Annals of GIS*, 1-13.
[3] N Sari Aslam, MR Ibrahim, T Cheng, H Chen, Y Zhang 2021. "ActivityNET: Neural Networks to Predict Public Transport Trip Purposes from Individual Smart Card Data and POIs." *Geo-spatial Information Science* 24 (4): 711–721.

an activity-based modelling approach with data on trips and trip purposes from surveys to design the layout of shops and activities in a new district to find possible current and future demand in an urban planning project. Even though trip purposes are investigated through trip-based or activity-based travel demand models from conventional travel surveys in transport research (Castiglione, Bradley and Gliebe, 2015), there is a need to investigate trip purposes from big data sources for public transport and services (Anda, Erath and Fourie, 2017; Kumar, Khani and He, 2018; Zannat and Choudhury, 2019).

2.1.2 Human mobility

Human mobility is one of the interdisciplinary subjects to investigate people's movement in cities. Data for human mobility research can be gathered or collected from sources such as mobile phone calls (Gong, Yamamoto and Morikava, 2016), social media (Hasan and Ukkusuri, 2015), SCD (Sari Aslam, Cheshire and Cheng, 2015; Sari Aslam, Cheng and Cheshire, 2018) and used for transport and urban planning purposes (Yang, Zhao and Lu, 2016).

Moreover, human mobility research also questions 'why, when, where and how people move in cities, similar to transport research from surveys (Anda, Erath and Fourie, 2017). Using emerging data sources, i.e., SCD, GPS and social media data, due to overlapping mobility traces opens up new opportunities with unique features such as large volume, easy access and small cost to investigate why individuals move in urban environments. However, their potential scope and limitations for further application, i.e., trip purposes, are not fully utilized (Liao, 2021). In addition, trip purposes from mobility patterns using SCD are limited based on activity types such as home and work/study (Long, Zhang and Cui, 2012; Hasan et al., 2012; Huang and Tan, 2014; Zhong et al., 2016; Zhong et al., 2016; Yang, Zhao and Lu, 2016; Mahrsi et al., 2017; Ma et al., 2017; Qi et al., 2018; Sari Aslam and Cheng, 2018; Yang et al., 2019; Arriagada et al., 2022).

2.1.3 Retail analysis

SCD provide valuable information for identifying travel habits and activities that can help commercial organisations to build their products and brands. An individual-based detailed analysis of SCD can illustrate users' movements at stations with time and duration, which can help certain types of retailers, such as small shops, dry cleaners, or small coffee shops, to consider the commercial spaces located around stations (Goulet-Langlois, 2016). For instance, Faroqi, Mesbah and Kim (2019) proposed trip-based (the maximum number of trips) and passenger-based models (the maximum number of passengers) for behavioural advertising in the public transit network after inferring trip purposes using the start time of the activity, activity duration and land use types in close proximity. The main contribution of their study highlights the importance of trip purposes, which connect behavioural advertising techniques and public transport to effective advertising. Another study is proposed by Trasberg, Soundararaj and Cheshire (2021), using big data sources to represent individuals' footfall movement for retailers. Their study collects raw signals from smartphones to investigate participants' activities without compromising individuals' privacy. Both recent studies show that investigating individuals' activities, in other words, the reason for human movements, through emerging big data sources provides a valuable insight for commercial organisations, which supports the motivation of this study.

2.1.4 Environment assessment

Travel demand studies have considered environmental measures and policies to manage transport infrastructure and developments in rapid urbanisation and industrialisation contexts. Similar measurements are also under consideration by governmental officials aiming to improve the quality of life in urban settings. Therefore, spatiotemporal transport data have contributed to environmental monitoring and assessment (Zheng et al., 2014). For example, Perchoux et al. (2019) used trip purposes as a measure to examine and correlate walking age groups and combine this information to see the environmental influence of walking in an urban setting. Hosseinzadeh and Baghbani (2020) focused on walking trips in the city of Rasht, Iran. First, traffic analysis zones were created in the city and then direction (from/to) information between zones considering four trip purposes was investigated. They found that walking trips were influenced by the density and diversity in urban settings and their findings were discussed in the context of how they could contribute to transport policy, urban planning and public health studies. However, these studies are based upon surveys and need to move forward using big data sources to help transport and urban-related studies under interdisciplinary thinking.

2.1.5 Public health

Public health presents another opportunity for public transport research application. For instance, Ibrahim et al. (2020) applied a variational-LSTM autoencoder to forecast the spread of coronavirus for every country using different measurements, such as the status of public transport (i.e., open or closed), rate of infection of the disease, urban population and population density. Bhatt et al. (2017) used a Gaussian process to map the risk of the disease, which provided significant support for planning toward global health goals. Moreover, Ray et al. (2017) developed kernel conditional density estimation to present a predictive map of disease using the spatiotemporal correlation of the data. One of the more recent studies (Caicedo, Walker and González, 2021) has investigated mobility (transit) data and highlighted that the SA (rest of the activity apart from home and work/study) locations are visited greater than PA locations during the COVID-19 pandemic, depending on passengers' socioeconomic status. The contribution of their work aims to highlight the impact of the COVID-19 lockdowns and guide transport and urban planners to improve transport and services during the recovery period. Besides, the recent studies highlight the importance of trip purpose inference not only for transport and urban-related studies but also for public health studies.

2.2 Data sources

This section provides brief information about inferring trip purposes from conventional travel surveys to new data sources such as smart card data. In addition, the scope and limitations of other data sources such as GPS, smartphones, mobile networks and social media data are discussed to overcome the limitation of SCD.

2.2.1 Travel surveys

Trip purpose has conventionally been collected through travel surveys aiming to gather households' travel information, including socioeconomic and demographic characteristics, to inform transport planning (Ortúzar and Willumsen, 2011; Pelletier, Trepanier and Morency, 2011). The surveys are divided into two sections, as shown in Figure 2.1. The first section focuses on household-level demographic information, such as gender, age, income and house ownership/tenure. The second section gathers individual-level information by asking respondents to list household members over five years of age. This section also contains two sub-categories: the first focuses on socioeconomic characteristics, including work status and general travel information, such as frequency and mode of public transport usage and public transportation passes held. The second asks for ‘travel diaries’ (trip sheets or travel logs) of detailed travel information.

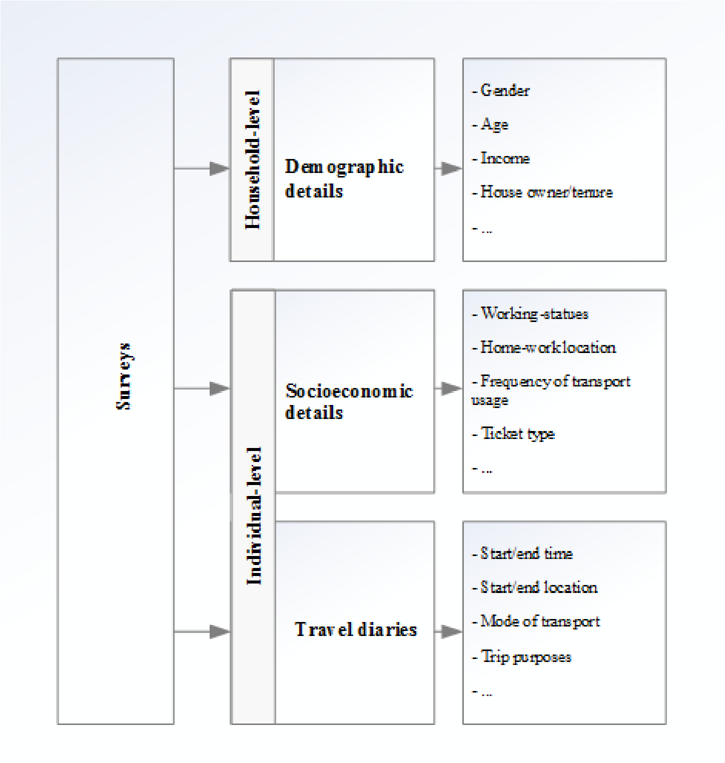


Figure 2.1 Travel/household surveys and characteristics.

Figure 2.1 further illustrates that trip purpose falls under the travel diary section of the surveys and is collected from survey respondents using the same or a similar template as is presented in Table 2.1

Aside from the example below, several types of surveys of different scopes and natures are explored to investigate travel behaviour. For instance, the National Travel Survey (NTS) – a household survey – has been conducted face-to-face by the UK Department for Transport (DfT) every year in England since 1988 (DfT, 2019). The NTS collects a total of seven days of travel data in a written travel diary and approximately 7000 households with 16,000 individuals from all age groups, including children, provide their travel information for calibration and validation purposes to resolve transport issues. Some of the collected data attributes are the mode of travel, start time of the activity, day of the week and travel purpose by age, gender and region, among many others.

Table 2.1 An example of a ‘travel diary’ survey sheet

Day of travel	Time		Location		Transport (Mode)	Trip purposes	Comments
	Start	End	From	To			

Another survey is London Travel Demand Survey (LTDS), collected face-to-face by Transport for London [TfL] annually in London. Eight thousand randomly selected households take part in the survey, including demographic, socioeconomic and travel-related information from a minimum of four days of travel data for the whole population [of the London] (more than 9 million) (TfL, 2011). Some of the collected data attributes are travel mode, time, purpose, origin and destination of each trip stage, ticket type (student or adult), price and zone.

Though travel surveys provide rich information about travel choices (mode of transport), travel characteristics (age, income, etc.), demographic details and trip purposes (Wermuth, Sommer and Kreitz, 2003), they tend to include the following limitations:

- Mismatched time period: surveys are expensive to conduct; therefore, they are usually collected over one day and used to extrapolate the use of transport during the entire year.
- Inadequate sample size with low update frequencies: surveys are usually carried out over one day, resulting in a limited sample size used to describe the entire population.
- Error-prone: household surveys are mainly comprised of a list of questions, which is labour intensive for both those conducting and responding to the survey. Sometimes, people may misinterpret the question or the person conducting the survey may make mistakes in collecting, storing, or delivering the information.

Due to the aforementioned limitations, this study has explored new data sources for travel behaviour and mobility research in detail, as described in the following sections.

2.2.2 Automatic fare collection systems

Automatic Fare Collection Systems (AFCs) are offered as an alternative to replacing conventional ticketing services in public transport (Pelletier, Trepanier and Morency, 2011). The system automates ticketing with a reusable card and reduces the cost of travel and user time during boarding. AFC systems data store records of individuals' trips with personal information such as name, age, date of birth, email address and the cost of tickets (Pelletier, Trepanier and Morency, 2011). AFCs are suitable for exploring public transport-related issues such as transit service quality and reliability (Uniman et al., 2010; Bagherian et al., 2016), route choice and demand modelling (Chu and Chapleau, 2008; Viggiano et al., 2017), mobility analyses of travel patterns (Liu et al., 2009; Yuan et al., 2013; Zhong et al., 2015; Zhong et al., 2016; Ma et al., 2017; Sari Aslam and Cheng, 2018; Zhang, Cheng and Sari Aslam, 2019; Yang et al., 2019; Zhao, Koutsopoulos and Zhao, 2020a; Liao, 2021), Origin–Destination (OD) estimation (Cui, 2006; Seaborn et al., 2009; Nassir et al., 2011; Munizaga and Palma, 2012; Padinjarapat and Mathew, 2013; Alexander et al., 2015; Alsger, 2016; Kumar, Khani and He, 2018; Hussain, Bhaskar and Chung, 2021), activity identification (Bouman et al., 2013; Nassir, Hickman and Ma, 2015; Goulet-Langlois, Koutsopoulos and Zhao, 2016; Ectors et al., 2017; Zhi et al., 2017; Ordóñez Medina, 2018) and trip purpose identification (Devillaine, Munizaga and Trepanier, 2012; Chakirov and Erath, 2012; Lee and Hickman, 2014; Kusakabe and Asakura, 2014; Zou et al., 2016; Alsger et al., 2018; E. Kim, Y. Kim and D. Kim, 2020; Faroqi and Mesbah, 2021).

2.2.2.1 Transport smart card data

Automatic data collection systems collect transport data – SCD – using SCs. SCs are portable and durable devices that store and process data for identification, authorisation and payment in many applications, including banking, retail, health care, parking transactions/payments, government and human resources and public transport. They are classified as either contact-based (e.g., the Oyster card) or contactless (e.g., some credit or debit cards), depending on the signal frequency and data transmission capabilities of the implanted chip (Pelletier, Trepanier and Morency, 2011).

Some implementations of SC payment systems within transportation networks around the world include UPass in South Korea, which was introduced in 1996 (the first such

SC system), Octopus in Hong Kong (the second-oldest), the EZ-link card in Singapore and the Presto card in Ontario, Canada.

SCD are considered the best possible complementary data source for household surveys and it is widely believed that they may reduce the need for travel surveys or replace them entirely (Alsger et al., 2018). This is because they produce large quantities of very detailed transaction data that can be of great use to transportation planners, from day-to-day operations to long-term strategic planning. As smart cards are associated with individual users, they present a unique opportunity to log each journey and capture relatively detailed spatial and temporal attributes such as origin and destination stations and trip duration. Furthermore, they also help streamline revenue collection by facilitating, for example, long-term cost reduction, flexibility in pricing options and the ability to share information with other parties. Thus, SCD provide new opportunities for gathering detailed, individual travel information without any surplus cost (Bagchi and White, 2005; Roth et al., 2011; Hasan et al., 2012a). The advantages of SCD are as follows:

- Longitudinal SCD provide rich daily information for individuals, unlike the small multiday samples from surveys that are used to represent individuals' yearly travel needs.
- SCD is collected by automatic data collection systems, which provide a high-quality and automated data-capturing process compared to more labour-intensive manual surveys.
- SCD are cost-effective data sources that include individuals' spatiotemporal daily travel information to inform transport and urban planning.
- SCD can be combined with other data sources – such as detailed transit operational data, traffic network data and land use data – to plan transit systems/operations or activity locations.

Although SCD have a wide range of positive characteristics, they also have some limitations, such as low spatial and temporal resolution, a lack of sociodemographic information due to privacy concerns and a lack of trip purpose information, including limited spatial information, i.e., tap-in/-out station. That means data represent only individuals who used the transport network without the first and last mile of the journeys made by walking, cycling, or private transport modes such as cars. However, trip purposes can be inferred using data mining techniques with the help of auxiliary data sources, such as GPS-based location data and geo-located location data (Devillaine, Munizaga and Trepanier, 2012; Kuhlman, 2015). Thus next section will review other data sources to help the limitation of SCD.

2.2.3 Other data sources

To overcome the limitation of SCD, other data sources are investigated in this section for the following reasons: 1) for validation purposes to evaluate proposed models and 2) for enrichment purposes to mitigate data limitations inherent in SCD. Thus GPS-based data sources (i.e., global positioning systems [GPS], smartphone and mobile data) and Location Based Social Network [LBSN] data sources (i.e., Twitter, Foursquare data) are compared to SCD in terms of scope and limitations with the help of Table 2.2.

GPS is a navigational system that provides spatiotemporal information, including individuals' dates, times and geographical coordinates. GPS passive data collection does not require any input from individuals and, therefore, cannot provide travel-related information, such as trip purpose or the number of boarding/alighting stations (Transport Systems Catapult, 2017). Although trip purpose has been inferred by GPS-based travel data with additional data sources using GIS technologies, inferred information still needs to be validated by individuals or through travel surveys (Transport Systems Catapult, 2017). In addition, GPS devices need to be charged and carried by the individual all the time to produce a comprehensive dataset (Nguyen et al., 2020).

The second and third potential data sources are derived from mobile phones, which are considered under two categories, i.e., smartphone sensor-based data and cellular network-based data (Wang, He and Leung, 2018) used for public transport studies. Smartphone sensor-based data are mobile phones with advanced operating systems that combine normal phone features such as phone calls, texts and others such as individuals' spatial movements, e.g., location data. Smartphones collect a large amount of location-based data throughout daily life and provide information for the whole population without requiring input from individuals'. In addition, this large data source can also be combined with other big data sources for further development. However, smartphone data cannot provide trip purpose information, travel mode, or boarding and alighting stations, among other details (Yang et al., 2021). The limitations may increase depending on each device's battery life and Wi-Fi connection and individuals' privacy concerns (some people may switch off their phone or their phone's location services if they do not wish to share their location) (Wang, He and Leung, 2018; Transport Systems Catapult, 2017). On the other hand, cellular network-based data collected by telecom companies, call data record (CDR) data, contains a set of phone activities such as phone calls or Internet access, along with the location and time information of cell towers feeding the call. CDR data collect information from users once they are moving within the coverage area, either while using (active) or not using (passive) their phones. In this data source, mobility information is available from mobile network data without any additional data collection processes (Wang, He and Leung, 2018). Therefore, they can offer valuable insights for transport planning and modelling research (Zannat and Choudhury, 2019). However, trip purposes, modes of transport, boarding/alighting stations, etc. and demographic attributes (e.g., age) are restricted due to privacy concerns by mobile phone providers (Transport Systems Catapult, 2017).

The last alternative data source is Location Based Social Network (LBSN) data available from entities such as Twitter[2], Foursquare data [3] etc., which is available at a small cost (Rashidi et al., 2017) and can be used for enrichment purposes to aid public transport research (Zannat and Choudhury, 2019)**.** While attributes such as trip purposes, travel modes and boarding/alighting stations of the journeys are not available, they could be extracted using advanced skills and technologies (Rashidi et al., 2017, Transport Systems Catapult, 2017). However, LBSN data contains biases (Rashidi et al., 2017), such as sample bias – in which some locations (e.g., for eating and shopping) have more numerous check-ins compared to other locations (e.g., home and work) – and demographic bias, meaning that younger groups between 15 and 30 years of age may be over-represented compared to older age groups (Longley and Adnan 2016).

2.2.4 Summary

Trip purposes are gathered from conventional surveys, which have a relatively small sample size (only a one-day travel diary) and are used for estimating travel demand for the whole population. Such surveys are expensive and time-consuming. Gathering SCD sources is significantly more efficient in terms of both cost and time due to the automated nature of these systems (Zannat and Choudhury, 2019; Pelletier, Trépanier and Morency 2011). In addition, they are usually available for a much larger population and for a

[2] http://twitter.com
[3] http://foursquare.com

longer period, which assists in understanding mobility behaviours and travel flows. (Bagchi and White 2005). Despite the wide range of impressive features, SCD present several challenges, such as: estimating a commuter's destination if public transport does not ask for alighting information (Gordon et al. 2013), making demographic predictions if socio-demographic information is not accessible due to privacy concerns (Zhang, Cheng and Aslam, 2019; Zhang, Sari Aslam and Cheng, 2020) and detecting activities in order to estimate a trip's purpose by linking smart card data with auxiliary data sources (Devillaine, Munizaga and Trepanier, 2012b; Kuhlman, 2015; Sari Aslam and Cheng, 2018; Yang et al., 2019). Thus, other possible data sources, such as GPS, smartphones, mobile phones (CDR) and LBSM, have been investigated, which offer many distinctive features and benefits (Chen et al., 2016).

Although there are opportunities to use new big data sources, challenges are inevitable. These may involve issues with data collection and pre-processing, or with analysis of these data sources due to mining large data with various strategies for transport modelling. These efforts may face other technical challenges such as computational efficiency, data processing, integration, evaluation and validation, as well as user privacy. For instance, one of the major challenges is the data gap (e.g., mismatch period of data, errors and missing information) between available data sources, which causes biases in the results (Zannat and Choudhury, 2019). Second, the absence of some of the details for individual-level analysis, e.g., demographic details, may result in biases in transport research due to vital input features for some of the traditional models, such as discrete and route choice models, being unavailable from big data sources (Arriagada et al., 2022). Third, using LBSM as auxiliary data to enrich location information from SCD may involve contribution and demographic biases. Fourth, considering travel surveys or censuses, which are updated once a year or every ten years, respectively, may result in biases validating trip purpose inference models. Last, none of the other data sources, i.e., GPS, smartphones, CDR and LBSM, provides trip purpose information directly and all require user validation of the labelling process (Transport Systems Catapult, 2017).

Table 2.2 The comparison of SCD with other possible data sources

Attributes	Smart card data	GPS data	Mobile data		Geo-located data (Twitter/Foursquare)
			Smartphone data	Call data record (CDR)	
Origin (journey starting point)/ Destination (journey ending point)	√	√	Can be calculated using location-based information	Data processing is required to derive the information	If the user tweets frequently with location information
Time left from departed location to origin station/arrived from destination station to arrival location	-	Can be inferred with pre-processing steps	Can be inferred with location information	Can be inferred from the timestamp of speaking time on the phone and cell tower	- Unless the information is available in the text
Purpose of trip	Can be inferred, but cannot be validated	Can be inferred	Can be inferred	Can be inferred, but cannot be validated	Can be inferred if mentioned in the text
Distance travelled	Distance travelled by public transport	Can be detected once the location is changed	Can be found and inferred using location information	Can be detected once the location is changed	Can be found from tweets and their timestamps
Mode of transport	√	Some level of information might be inferred by heuristic rules with speed, arrival time	Can be inferred by heuristic rules with GPS locations, speed, arrival time	Some level of information might be inferred from the spatial pattern, speed and distance	- Unless the information is available in the text

2.3 Reviewing trip purpose inference from smart card data

A systematic review approach is adopted using PRISMA guidelines (PRISMA, 2015) for trip purpose inference. First, all manuscripts to date related to trip purposes are gathered. The search takes into account books, book chapters, peer-reviewed journal articles, conference proceedings and technical reports using four search engines: Google Scholar, Scopus, Microsoft Academic and Web of Science. The combination of keywords, as shown by a function using Boolean operations, are ‘activity types’ or ‘trip purposes’ and ‘public transport’; ‘activity inference’ or ‘trip purposes’ and ‘enrichment’ and ‘SCD’; ‘trip purposes’ and ‘SCD’ or ‘social media’ or ‘GPS’.

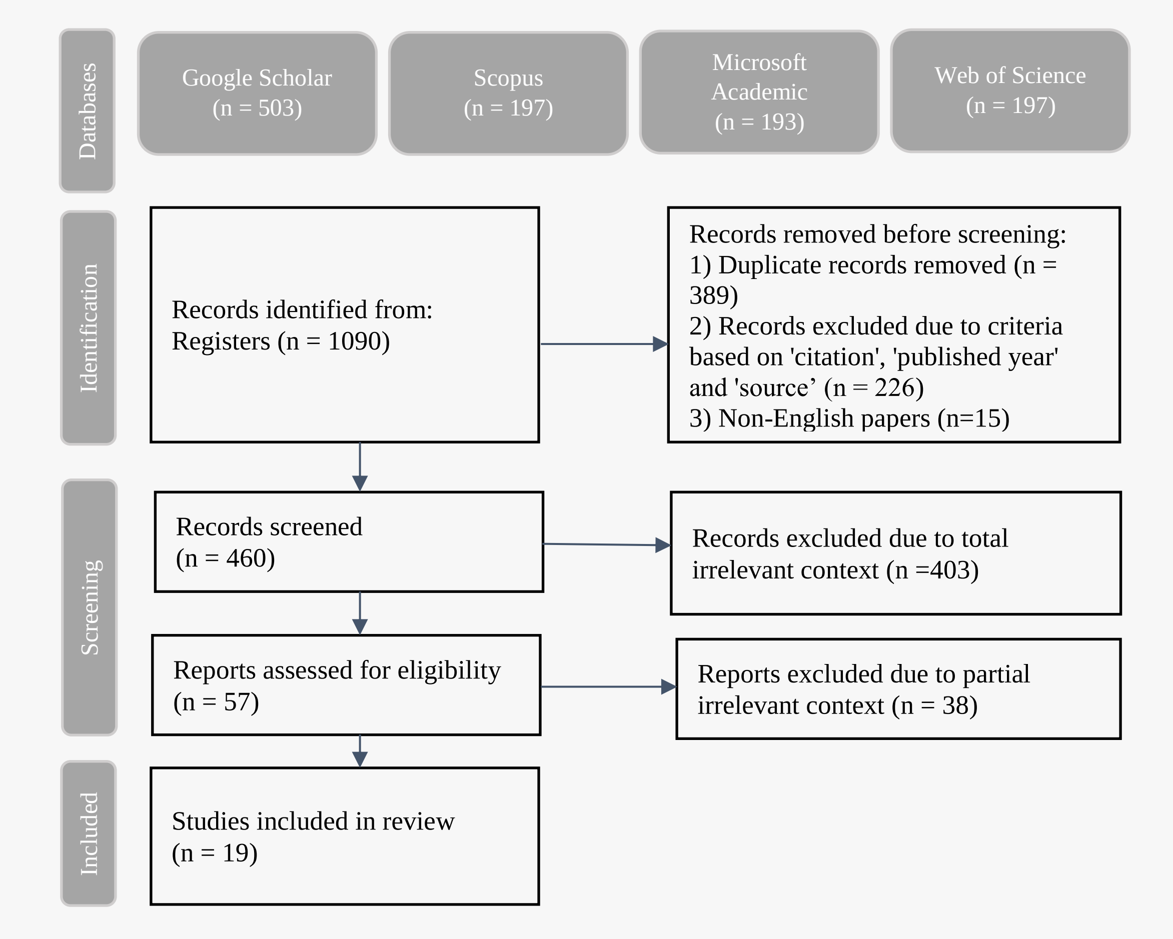

Databases
Google Scholar (n = 503)
Scopus (n = 197)
Microsoft Academic (n = 193)
Web of Science (n = 197)
Identification
Records identified from: Registers (n = 1090)
Records removed before screening:
1) Duplicate records removed (n = 389)
2) Records excluded due to criteria based on 'citation', 'published year' and 'source’ (n = 226)
3) Non-English papers (n=15)
Screening
Records screened (n = 460)
Records excluded due to total irrelevant context (n =403)
Reports assessed for eligibility (n = 57)
Reports excluded due to partial irrelevant context (n = 38)
Included
Studies included in review (n = 19)

Figure 2.2 Flow chart for the systematic review.

After entering these combinations of search terms, Figure 2.2 illustrates the results obtained from the search engines listed above, totalling 503, 197, 193 and 297, respectively. After combining all results (1090 research items), duplicate entities are removed and the sample size decreases to 701. Filters are then applied using ‘citation’, ‘published year’ and ‘source’ attributes. If the year of publication was before 2015, the citation is absent, or no source information is available, the study is removed, as are non-English papers (15). The total number of manuscripts is thereby reduced to 460. Then, the screening process begins and manuscripts are checked by title, keywords and abstract. After this step, the total number of relevant publications is 57. After reviewing each manuscript, some of the papers are found to partially involve activity patterns and mobility analysis using individuals’ activity types and locations from SCD, but not specifically trip purposes. Therefore, 19 papers are considered relevant due to their clear methodologies, which can contribute to future research on trip purpose inference from SCD.

2.3.1 Input features: Activities/trip end stations

From an SCD perspective, a trip is defined as a one-way journey from one station (origin) to another station (destination) using the public transport network in an individual’s daily travel. Besides, the definition of an activity is time spent between two consecutive trips as obtained from SCD. Furthermore, trip purposes are derived from extracted activities, either data mining or processing steps using the characteristics of SCD or additional information, i.e. land use attributes in the proximity of start/end stations (Faroqi, Mesbah and Kim, 2018). Therefore, ‘activity type(s)’, ‘activity inference’ and ‘trip purposes’ are considered to have the same meaning, but ‘activities’ in this research. Two types of activities are identified to infer trip purposes using the characteristics of SCD: primary activities (PAs), which occur at home and work or school (for adults and students, respectively) (Section 6.1) and secondary activities (SAs), which include all other activities (such as eating, shopping and entertainment) outside of PAs (Section 6.2).

There is a close relationship between trips and activities because they share the same spatial point in the transport network at the same time. Figure 2.3 illustrates this relationship at the individual level in terms of space and time, where (Si) is the spatial location of station number (i). Tj denotes a one-way journey (j) from one station to another. Temporally, an activity (A) is time (ti) spent between two consecutive trips (Tj and Tj + 1). The start time and location of the trip are the end time and location of the previous activity and the end time and location of the trip are the start time and location of the next activity. Thus, researchers have selected input features either related to ‘activity’ or ‘activity and trip’ to build their models, except Hasan et al. (2012) used only the frequency parameter.

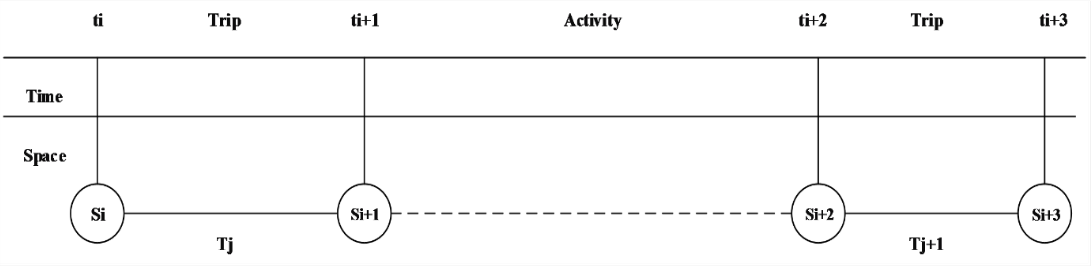


Figure 2.3 Schematic diagram of the data generation process for an activity.

There are four characteristics of an ‘activity’: ‘activity duration’, ‘start/end time of the activity (time of day)’, ‘day of the week’ and ‘location of the activity’. The most widely used characteristics in research on trip purpose inference are ‘activity duration’ and ‘start time of the activity’ (Zou et al., 2016; E. Kim, Y. Kim and D. Kim, 2020; Li et al., 2015; Faroqi and Mesbah, 2021; Devillaine, Munizaga and Trépanier, 2012; Han and Sohn, 2016; Du, 2019; Chakirov and Erath, 2012; Ordóñez Medina, 2018), except for home activities. The reason for this might be to eliminate confusion about the duration between home and work activities in trip/activity chains. In addition, this combination is changed to ‘start time of the trip’ and ‘activity duration’ by some researchers (Lee and Hickman, 2014; Wei, Liu and Sigler, 2015; Alsger et al., 2018). The reason for this might be based on the available data types in their research, e.g., some surveys have only start time and location of the trips and trip purposes. In that case, the difference reflects the length of travel time between ‘start time of the trip’ and ‘start time of the activity’ per individual.

2.3.2 Enriching smart card data with other data sources

Identified station-based activities need to be enriched with auxiliary information such as land use data, POIs, or surveys. These can provide information about the types of activities performed (Noulas et al., 2015; Gong et al., 2016) and thus facilitate activity prediction and activity pattern classification (Hasan and Ukkusuri, 2014). Thus, PAs and especially SAs from large spatiotemporal transport data need auxiliary information to infer trip purposes.

The enrichment part of the study is mainly considered ‘frequency’ (an indicator of the regularity of the visit/trip or activity) as a characteristic/parameter of SCD. That means if the home or work locations/activities are defined with what-if scenarios using activity- and trip-related characteristics, frequency parameter helps in the decision process to determine the conditions based on the location either from SCD alone (Yuan, Winter and Wang, 2019) or other enriching data sources, such as land use POIs.

There are three types of enrichment in the existing literature investigating trip purposes based on SCD: 1) enriching SCD with household or travel surveys (Chapleau, Trépanier and Chu, 2008; Lee and Hickman, 2014; Zou et al., 2016; Alsger et al., 2018; Faroqi and Mesbah, 2021). 2) enriching the station information with land use attributes (Lee and Hickman, 2014; Alsger et al., 2018) and 3) enriching activities with geo-location social media data, e.g., from POIs (Yang et al., 2019).

Survey data have been used to enrich SCD in the literature (Lee and Hickman, 2014; Alsger et al., 2018; Faroqi and Mesbah, 2021), for example, through the distribution of extracted activity types with temporal characteristics from survey data to determine the possible activity types around alighting stops based on the temporal probabilities from SCD (Alsger et al., 2018). The main limitation of this methodology was that the representations of activity types were based upon one-day survey data (Chakirov and Erath, 2012; Amaya, Cruzat and Munizaga, 2018; Alsger et al., 2018), which may not be adequate for representing all activity types for the whole population from SCD. In addition, these methodologies are limited once such surveys are not available for the urban settings.

On the other hand, land use data, without a temporal variable, have also been used to infer trip purposes (Lee and Hickman, 2014; Alsger et al., 2018). However, many urban points of interest (establishments) have multiple functions with different opening and closing hours. Furthermore, the land use data used in the research treat all geographical points as equal without applying a weighting factor such as popularity. These limitations, multiplied across thousands of locations, represent a major source of bias when analysing and inferring individual activities (Lai, 2018). Lastly, enriching SCD with POIs from social media data, such as from Foursquare, can provide opportunities to infer trip purposes using the opening and closing hours of given locations, check-ins and activity types (Yang et al., 2019). However, POIs from social media may over-represent some locations, with substantial counts in restaurants or shopping centres compared to workplaces (Rashidi et al., 2017). In addition, demographic biases in the dataset are inevitable, given that Foursquare is used more by younger groups (e.g., those under 30 years old) than older groups in the city (Longley and Adnan, 2016).

Another challenge presents itself in determining station-based trip purposes from mixed land use types as opposed to single land use types. If cities are well segregated in terms of residential and work areas – as is the case in Beijing and Brisbane, for example – then researchers can use this information as ground truth in their model and make decisions accordingly. For instance, if the destination of the trip is a residential area or work area, then activities are assigned based on location (Wei, Liu and Sigler, 2015; Wang et al., 2017; Ordóñez Medina, 2018). On the other hand, some studies suffer from the distribution of highly mixed land use types– as is the case, for example, in London, New York and Tokyo. This results in a lack of clear evidence from the land use data regarding whether the location is dominated by residential, work, or other activities. In that case, some researchers do not use land use information and focus on only the characteristics gleaned from SCD (Hasan et al., 2012). However, a model or framework using only SCD without any enrichment from land use or other data sources may work for PAs but not for SAs due to inconsistent travel patterns with rich classifications of activity types such as eating, shopping and entertainment.

2.4 Methods for trip purposes

This section reviews the literature investigating trip purposes from a methodological point of view, focusing on three categories of methodologies: heuristic (rule-based), machine learning (ML) and statistical. Each method is explained briefly in an order based on popularity in the existing literature. The heuristic approach is most commonly used to investigate trip purposes. Therefore, it is presented first, followed by the ML and statistical approaches. In addition, the relevant works from the literature are summarised in Table 2.3.

2.4.1 Heuristic (rule-based) approach

A heuristic technique is used when an optimal solution is impractical or impossible. This technique provides self-discovery for the problem, which requires prior domain knowledge using the characteristics of the data. The method results in immediate findings with a satisfactory solution and aids in the decision-making process by using a set of automated rules, which decreases the labour required compared to manual exploration (Yu et al., 2020). The technique has been widely used in different areas, such as cognitive mapping, philosophy, law and artificial intelligence (Pearl, 1983; Yu et al., 2020), as well as transport and urban research – e.g., OD estimation (Alsger, 2016), travel behaviour analysis (Agard, Morency and Trépanier, 2006), mobility research (Hasan et al., 2012), location choice modelling (Wang et al., 2017), activity pattern analysis (Zhou et al., 2021) and trip purpose inference (Zou et al., 2016).

The reason for the prevalent research interest in the heuristic approach is that large SCD with spatiotemporal details have prior information but labelled information. Thus, some of the methods are impossible to apply straight away, e.g., discrete choice models – a model uses a set of two or more discrete options (i.e. distinct and separable; mutually exclusive) to describe the most prefered choices, e.g., location, activities – or supervised ML methods. Therefore the heuristic approach has mainly been used to identify PAs from SCD to observe the longitudinal behaviours of the data in the current literature. On the other hand, there are challenges to the heuristic approach that are as follows: 1) The performance of the method relies on the quality of the data (Ross, Wei and Ohno-Machado, 2014); therefore, pre-processing steps are essential before applying this method. 2) The selection of the parameters and rules and tuning parameters (Turchin, Gal and Wasserkrug, 2009) has to be checked, especially once a new dataset has arrived in the system or cloud, to achieve better performance (Neill, McFowland and Zheng, 2013). 3) The information/outcome captured from this approach only represents the defined rules. The model does not have the flexibility to find new patterns in a large dataset. However, the model is fast in disaggregate analysis and can provide a valuable outcome once the rules are widely agreed upon.

The following researchers have investigated PAs in different cities with their scope and limitations as follows.

Chapleau, Trepanier and Chu (2008) enriched SCD using primary anchor locations (study, work and residential locations) to infer trip purposes on the bus network in Gatineau, Canada, using one month of data. Card types were used with appropriate anchor points, such that student cards were paired with study points and adults' cards with work or residential points. Spatial (distance from boarding and alighting points to nearest anchor point) and temporal (time of the trip) characteristics with frequency parameters were further investigated using multiday transit data to make transit planning more efficient in the study area. However, presenting the aggregated results without accuracy makes the study impossible to compare with any other studies. In addition, the study needs features such as anchor locations from surveys and only identified the limited PAs, i.e., work/study activities.

Of note among heuristic research is the study of Devillaine, Munizaga and Trépanier (2012), in which the researchers focused on the temporal attributes of activities and card types of SCD. The study only detected PAs, such as home and work (adults) or school (students) from Santiago, Chile and Gatineau, Canada. Home activities were identified as the destination of the last journey of the day and work activities were defined based on the work culture of each city: those which lasted longer than 2 hours (weekday) in Santiago and longer than 5 hours (weekday) in Gatineau for adult-registered cards. As a limitation, the study's first step uses the card type information from SCD, which is not available for many cities, including London.

Hasan et al. (2012) also proposed a simple mobility model for predicting home and work locations and activities using the frequencies of places visited in the city by individual users. According to their study, the most frequently visited places were classified as home locations, whereas the second most visited locations concentrated around city centres were identified as work locations. The study is limited in terms of identifying PAs without relying on surveys.

Wei, Liu and Sigler (2015) studied home and work locations and activities from five days of SCD in Brisbane, Australia. The study extracted journey-to-work patterns from SCD using three rules: 1) alighting time is between 6 am and 10 am, 2) activity duration, or alighting time to next boarding time, is more than four hours and 3) frequency for rules 1 and 2 must be repeated four or more times in five weekdays. Once journey-to-work patterns were identified, the origin location/station was established as the home location and the destination location was considered the work location. Validation was achieved based on land use attributes. As a result, work locations were identified correctly in city centres. However, the identified home locations failed to point to residential areas in Brisbane. According to commuter patterns from transport data, the authors suggested that some of the residential areas may not have been residential areas anymore at the time of the study.

A more recent study by Zou et al. (2016) extended the assumptions of Barry et al. (2002) to identify PAs. Their study proposed a centre point–based detection algorithm to detect home locations. The algorithm used the shortest path algorithm between the location of the first trip and the location of the last trip for the selected day. If the shortest distance was less than the defined walking distance around the station, the centre point is captured as the home location for the passenger and moved to the next passenger. Once all passengers are checked, the next day is used for determination/validation purposes. Thus, the algorithm was able to accurately identify 88.7 per cent of passengers' home locations by mining one week of the SCD. Then the same data were analysed for travel information, including temporal attributes (activity duration and boarding time), spatial attributes (identified home location), maximum travel time and travel regularity, which were combined in a rule-based approach. Identifying work/study locations using card types (adult, student, employee and senior cards, one-trip cards and cards for working) and detailed surveys may have provided an accurate outcome in Beijing, but this may be due to its residential areas being segregated from the centre of the city. The study's methodology may not provide similar results for cities like London due to their unclear segregation between home and work locations. Moreover, the level of detail seen in Beijing's card types and surveys may not be available for other studies.

Wang et al. (2017) focused on one category of SAs – after-work activities – from public transport data in Shanghai, China. The study also identified home and work locations/activities. Home locations/activities were initially defined as the boarding station of the first trip and frequency and land use were then used to decide whether the location was a home location. If the location was in a residential area, then it was called a home location. Work locations were identified as places where more than six hours were spent on activities during the day. If work activities/locations appeared more than one time, the distance from the work location to the home location was multiplied by the frequency, as suggested by Alexander et al. (2015). The location with the largest value was then defined as the work location. After identifying home and work locations, they, i.e., home and work were excluded from the dataset. The rest of the activities were called SAs and their study mainly focused on identifying after-work activities, which were assumed (using a separate model from the one used to determine PAs) to occur from 16:00 to midnight. However, each person's sequence of activities differs daily and extracting them with fixed temporal attributes may have overlooked some of the SAs. Their contribution to the literature was mainly on SAs without relying on surveys.

Amaya, Cruzat and Munizaga (2018) proposed another home location identifier using card type and frequency data. First, they extracted regular users who made more than one transaction per day for five consecutive days, as suggested by Lee and Hickman (2011). The study focused on activities and transactions from 4 am to midday to decide the centre of gravity of the coordinates. For instance, if there were three tube stations used frequently for morning trips, their mid-coordinates were checked based on two distance measures. The first measure, the maximum distance from the centre of gravity to the first transaction, was considered 1000 m. The second measure was walking distance, defined as 500 m. If the maximum distance was lower than the walking distance, then the centre of gravity was accepted as the home location for the study. The maximum and walkable distances were decided based on the responses of 55 users in survey data. Although the study was applied in a well-segregated city in terms of residential and work locations (Santiago, Chile), only 70 per cent of home locations were correctly identified from the small survey data.

Alternatively, Alsger et al. (2018) incorporated PAs and SAs in a single model using a wide array of auxiliary data sources, such as an OD survey, land use data, household travel surveys and weather reports, for a rule-based approach. They extracted the distribution of trip purpose types over the temporal attributes from household surveys, then the temporal probabilities were considered according to the feasible activity types around the alighting stops. Although Alsger et al. (2018) shed light on inferring PAs and SAs, the study's methods do not apply to large SCD where the detailed surveys and auxiliary attributes are not available. In addition, while their work was successful for inferring PAs, it failed to achieve high accuracy for less frequent OD trips such as shopping and recreational activities.

Yuan, Winter and Wang (2019) also described home and work locations and activities from a heuristic approach. They separated the methodology between flat-fare SC systems (e.g., London's transit network and Melbourne's transit system) and distance-based systems (e.g., Beijing's and Singapore's transit systems). According to their study, home locations and activities were determined using the following steps: first, the first boarding station (or the last alighting stop from the previous day if it is applicable within a reasonable distance) was checked. If this station repeatedly appeared on different days, then it was identified as a home location. On the other hand, work locations and activities were defined by activity durations of more than six hours at stations that appeared on multiple days. The main difference between the flat-fare and distance-based methods mainly appeared in the approach for home locations using either ‘the first boarding stop’ or ‘the first boarding stop and the last alighting stop’, respectively. On the other hand, a six-hour interval, from boarding to boarding for flat-fare or alighting to boarding for distance-based methods, was considered to define work locations. However, boarding to boarding trips may include transfer times, such as 50 min for London (Seaborn et al., 2009) and 60 min for Beijing (Ma et al., 2013). The reason for using a threshold of six hours for both flat-fare and distance-based SC systems is because 96 per cent of the Beijing household travel survey respondents claimed to work over six hours at a time in 2005. However, such information may not apply to or be available for other study areas. Besides, according to their model, if the home or work locations appear more than once, the intersection area between boarding to boarding or boarding to alighting stations are checked using buffers. Then the most frequent location is assigned as home and work locations. Even though the study area is considered well-segregated in terms of home and work locations, the obtained accuracy was low for home and work locations using two weeks of Beijing SCD from 2017.

In summary, PAs models depend on features such as card types (Chapleau, Trepanier and Chu, 2008; Devillaine, Munizaga and Trépanier; Zou et al., 2016), an activity duration threshold (Yuan, Winter and Wang, 2019), POI locations, i.e., home, work, study locations (Zou et al., 2016; Alsger et al., 2018) from surveys. Thus, there is a need to identify PAs using only the characteristics of SCD without relying on surveys.

Even though the heuristic approach is also used for inferring trip purposes from other data sources such as GPS and mobile phone data (Wolf, Guensler and Bachman, 2001; Wolf et al., 2004; Stopher, FitzGerald and Zhang, 2008; Chen et al., 2010; Shen and Stopher, 2013; Hossain and Habib, 2021), the summarised information in Table 2.3 is presented from SCD, outlining the following for each study: data, study area and scope; input features; methods; the contribution to literature and accuracy of the study; and limitations and validation.

Table 2.3 A summary of the selected papers on trip purposes from SCD.

Heuristic approach					
Study's authors:	Data, study area and scope:	Input features:	Description of the method:	Contribution to literature and accuracy:	Limitations:

Chapleau, Trepanier and Chu (2008)	SCD, 785,383 transactions, no users;
	Canada;
	Work/study activities
		Activity duration, card type and POI	Considered 300m buffers around alighting/boarding stops to discover POIs. Detected major trip generators and assigned users to those points.	N/A	The method needs survey(s)
					Limited PAs (only work/study)
					No SAs

Devillaine, Munizaga and Trepanier (2012)	SCD, 38 million transactions, 3 million users (one week) in Santiago, 45 million transactions, 186,000 users (nine years) in Gatineau; Santiago, Chile and Gatineau, Quebec, Canada; Home, work/study and others	Activity duration, card type, last transaction of the day and POI	Extracted home locations as the last transaction of the day and the first transaction of the next day. Extracted work/study locations using cardholder status and activity duration	Compared defined activity types between two cities. The study considers more planning purposes than trip purposes. N/A	The method does NOT need survey(s) Limited activity types (only PAs) No SAs
Hasan et al. (2012)	SCD, 1000 users, no transaction numbers; London; Home, work/study and others	Only frequency	Classified most frequently visited places as home and second-most-visited locations as work. Other locations were checked based on an agent-based model with simple scenarios.	Defined a simple heuristic model for PAs. Others (SAs) were located in the centre of the city N/A	The method does NOT need survey(s) Limited activity types (only PAs) No SAs

Chakirov and Erath (2012)	SCD (five working days), no users and transactions, survey data (7936 transactions), land use data (Master Plan 2008); Singapore; Home, work/study and others	Activity duration for work activities, activity start time and land use information	A discrete choice model was used for home and work/study activities from survey data start time, land use and activity duration. As a heuristic approach, activities longer than six hours were identified as work activities.	Temporal variables were important for obtaining PAs. Spatial attributes from land use data did not have a substantial effect on the model. SCD and survey data were limited (five and one working day, respectively). N/A	The method needs survey(s) Limited activity types (only PAs) No SAs
Wei, Liu and Sigler (2016)	SCD (five weekdays), no users or transactions, land use; Brisbane, Australia; Home and work	Alighting time, activity duration and frequency of trip	Journey-to-work patterns were extracted based on morning peak and duration (more than 4 hours) and compared to land use distribution to highlight home and work locations.	Workplaces identified from SCD were reliable, but residential areas were not. SCD and survey data were limited. N/A	The method does NOT need survey(s) Limited activity types (only PAs) No SAs
Zou et al. (2016)	SCD (one week), 6.92 million transactions, 4.2 million users; Beijing, China; Home, work/study and others	Card type, activity duration, the start time of the activity, land use types and home locations	Identified homes using a centre point–based algorithm. Then, estimated home locations were used to identify work/study	Eighty-eight per cent of passengers' home locations and four types of trip purposes could be detected effectively.	The method needs survey(s) Limited activity types (only PAs) No SAs
Wang et al. (2017)	SCD, 3000 users Shanghai, China Home, work and after-work activities	Activity duration, first trips, land use, frequency and distance	Defined home as the first trip and work as more than 6 hours of activities. Frequently defined home and work locations were	Went one step further to understand secondary locations/activities (i.e., after-work locations/activities).	The method does NOT need survey(s)

			checked based on land use, distance and frequency	N/A	PAs and limited SAs (after-work activities)
Amaya, Cruzat and Munizaga (2018)	SCD (3,288,464 transactions, 55 users), surveys (358 records); Santiago, Chile; Home only	Card type and frequency	Checked all morning trips and defined a centre-of-gravity coordinate for a home area. Then, defined a residential zone using max and walking distance from the central point.	Seventy per cent correct estimates for home locations/activities from survey data. The proposed methodology is suitable for cities with clearly segregated home and work areas.	The method needs survey(s) Limited PAs (only home) No SAs
Alsger et al. (2018)	HTS 2600 survey trips, land use, the South East Queensland Strategic Transport Model (SEQSTM), General Transit Feed Specification (GTFS) data, OD survey data, SCD; Brisbane, Australia; home, work/study, shopping, recreational	Land use types around the alighting/end stop, start time of the trip and activity duration	Extracted the distribution of trip purpose types over the temporal attributes from HTS data, then considered the temporal probabilities according to the feasible activity types around the alighting stops.	Used a single model for PAs and SAs. Ninety-two per cent accuracy for work, 96 per cent accuracy for home activities. Identification of shopping and education trips improved after applying temporal attributes compared to spatial attributes.	The method needs survey(s) PAs (high accuracy) and SAs (low accuracy)
Yuan, Winter and Wang (2019)	SCD (two weeks), 7,283,866 transactions, 250,000 users; Beijing, China; Home, work and others	Activity duration, start location, frequency and boarding/start locations	Defined home as first boarding stop and work as activity duration more than six hours appearing in more than one day. If home or work appears in more than one location, check the	Proposed a methodology for flat-fare and distance-based SC systems. Sixty-nine per cent inference rate of residential locations and	The method does NOT need survey(s) Limited activity types (only PAs) No SAs

			intersection area using a buffer, including land use	more than 72 per cent inference rate of workplace locations were identified.	
ML approach					
Medina and Erath (2013)	SCD (7 days), no transactions, no users; travel diary survey, data is 1 per cent of SCD, building information data sources; Singapore; Work location/activities only	Activity duration and start time	k-means clustering was applied to survey data to gather information Then building information such as parcel size and footprints were combined for work capacities.	Identified work activities and estimated work capacities. Because SCD and land use attributes were used, the study also highlighted potential locations where users go/visit. N/A	The method needs survey(s) Limited PAs (only work) No SAs

Lee and Hickman (2014)	SCD (five working days), 300 users, no transactions;
	Minnesota, USA;
	Work/study and other activities
		Trip frequency, activity duration, card type, transaction time and land use	Rule-based work trips were defined as trips that took place in the morning peak and had return trips in the evening peak from SCD. A decision tree model was developed to label transactions by specific trip purposes.	Included a wide range of features.
				Unclear accuracy for work and study activities.
				N/A
					The method needs survey(s)
					
					Limited PAs (work/study)
					No SAs

Ordóñez Medina (2018)	SCD (7 days), 20,856,442 transactions, 274,005 users; travel diary survey data is 1 per cent of SCD, building information data sources; Singapore;	Start time of trips, trip travel time, locations of ODs and card type	Split survey data (i.e., work and study) and then extracted information using discrete choice models. Used the parameters to create an SCD vector to feed into DBSCAN and presented the 17 most	SCD were enriched by surveys (obtained home/work/study/others from surveys). Highlighted work/study activity patterns from seven days' worth of SCD.	The method needs survey(s) Limited activity types (only PAs) No SAs

	Home, work, study and others		popular work and study patterns.	N/A	
E. Kim, Y. Kim and D. Kim (2020)	SCD (full dataset included 77,904 users and 117,608 transactions; 10 per cent of SCD was used because of the scale of the survey's data); Korea; commute, business, leisure, return home	Activity duration, trip sequence, departure time based on three temporal windows (a.m./inter/p.m.) , travel time, age of traveller, land use (H/W/O area ratio) and bus/train stops	Survey data were used for feature importance using permutation-based variable importance, H- statistic–based variable interaction and accumulated local effect (ALE). Then, random forest (RF) was applied and compared to baseline methods – i.e., MNL (multinomial logit), NB (naïve Bayes) and GBM (gradient-boosting machine).	SCD were enriched by surveys. Eighty-three per cent overall accuracy was achieved. Temporal features (activity duration, departure time at the origin) were the dominant features. It was suggested that data collection methods need to be revisited to capture rich trip purposes (e.g., the rest of the SAs). Validation was based on SCD and survey data.	The method needs survey(s) PAs and limited activity types in SAs (i.e., leisure)
Faroqi and Mesbah (2021)	SCD (SCDset, 128,977 users with 282,453 transactions/trips), Survey data (1233 users with 2523 trips); Brisbane, Australia; Work, home, education shopping, recreational	End time of the trip (start time of the activity) and the time gap between trips (activity duration)	Trip sequences were created using the activity start time and activity duration. The similarity was measured (Jaccard) using users' sequences. Agglomerative hierarchical clustering (AHC) was applied to survey data and the best clusters were checked using the silhouette coefficient.	Enriched SCD from surveys. Accuracy was improved as compared to Alsger et al., 2018. 99 per cent accuracy for home; 93 per cent accuracy for work; 93 per cent accuracy for education; 94 per cent accuracy for shopping; and 95 per	The method needs survey(s) PAs and SAs

			Then, the method was applied to SCD.	cent accuracy for recreational activities from survey data Limited survey data (one-day data) to infer clusters.	
Probabilistic/Statistical approach					
Kusakabe and Asakura (2014)	SCD (20 months) 7,074,768 trips 553,259 travellers. HTS (survey) (1,586 trips, 1576 travellers/users; Commuting to work/school, leisure, business, returning home; Osaka, Japan	Arrival time (trip end time/start time of activity) and duration of activity	Proposed a data fusion approach using integrated SCD from household travel survey. Common data attributes of the SC and survey data were used in the NB formula for each trip purpose.	Overall estimation was 86.2 per cent in the survey data. The highest percentage (92.1 per cent) was captured of commuting trips compared to 74.2 per cent of leisure and business trips and 84.5 per cent of trips returning home.	Only one metro station was investigated. The method needs survey(s) PAs and limited activity types in SAs (i.e., leisure)
Li et al. (2015)	SCD (3 months); survey and land use; Home and work, others; Singapore	Card type, duration, activity start time and frequency Spectral analysis technique combined with heuristic	Using a rank aggregation technique and spectral analysis, derived a location ranking list in addition to identifying periodic travel patterns.	Ninety-five per cent accuracy for home locations was obtained against the city's urban planning dataset (survey data).	The method does NOT need survey(s) Limited activity types (only PAs) No SAs
Han and Sohn (2016)	SCD only one day (306,766 trips) Seoul, Korea	Start and duration times of activity and four land use	Proposed a continuous hidden Markov model (CHMM) using emissions probability to find eight	A CHMM does not require an extra survey to obtain labelled data for training.	The method does NOT need survey(s)

	Home and out-of-home activities, i.e., work	characteristics around activity locations	clusters interpreted as patterns for home and out-of-home activities.	The processing cost of SCD is uncertain. N/A	Limited activity types (only PAs) No SAs
Zhao, Koutsopoulos and Zhao (2020)	SCD 3,339,187 activity transactions from 20,667 users; London, UK Home, work, others	Start time of day, start day of the week and duration of each activity	Selected features were considered for a Latent Dirichlet Allocation (LDA) location model. To handle high-dimensional spatiotemporal information: spatial dependencies were ignored	Obtained higher accuracy than baseline models. Adding more activities (i.e., one before/one after activities) would have increased the accuracy, but a large number of latent activities may limit the interpretability of the results.	The method does NOT need survey(s) Limited activity types (only PAs) No SAs

2.4.2 Machine learning approach

After the heuristic approach, the ML approach is the second most popular technique for trip purpose inference from SCD in literature. There are two major types of ML techniques: ‘supervised’ and ‘unsupervised’ learning algorithms. Supervised learning techniques are called classification or labelling algorithms requiring a learning process from training (labelled) samples or ground truth data. Unsupervised learning techniques are named clustering algorithms and do not require any prior knowledge to infer or learn the underlying structure of the data from labelled datasets or ground truth data.

Clustering spatial and temporal attributes of SCD, such as passengers (Morency, Trepanier and Agard, 2007; Ma et al., 2013), stops and stations (Morency, Trepanier and Agard, 2006; Cats et al., 2015; Cats, Wang and Zhao, 2015; Cardell-Oliver and Povey, 2018) and activity patterns (Goulet-Langlois, Koutsopoulos and Zhao, 2016; Zhou et al., 2021) has been studied to improve understanding of travel patterns and behaviours (Faroqi, Mesbah and Kim, 2018). However, the ML approach needs attention inferring trip purposes from SCD (Anda, Erath and Fourie, 2017). The following subsections will review some of the ML methods used for trip purpose inference from SCD and other data sources in the existing literature.

2.4.2.1 Bayes classifiers

Naive Bayes classifiers are a probabilistic machine learning model based on Bayes' theorem, which calculates conditional probabilities. Bayes classifiers are easy to implement. However, the model assumes that features in the model are independent and have an equal effect on the output. Therefore, the model performs with low accuracy for SCD.

In literature, Kusakabe and Asakura (2014) suggested using the joint attributes from the person trip survey data and SCD to detect trip purposes. The naïve Bayes classifier was used in only one station in Osaka, Japan, for about 20 months. The methods of this study focused on estimating PAs and leisure activities and were validated with 86.2 per cent overall accuracy and 58.9 per cent accuracy for leisure activities from their model. The contribution of this study is for the PAs, whereas the model demonstrates low accuracy for leisure activities(SAs).

2.4.2.2 K-means

One of the simple ML techniques, K-means, is used for trip purpose inference. K-means is an unsupervised machine learning algorithm that minimises the sum of distances between the data points and their respective cluster centroid. Even though K-means are operated quite often due to the easy implementation, the drawbacks of the algorithm are as follows: First, there is a need to choose the number of clusters (k) manually. Second, the algorithm assumes that clusters are spherical clusters and each cluster has equal numbers for observations. Third, the algorithm has a sensitivity to scale. For instance, changing the scale of the dataset will change the results. Fourth, the algorithm is sensitive to the order of the values, which means that different orders in the dataset may provide different results. Fifth, the performance of the algorithm is very low once categorical and numerical variables are used in the models. Last, outliers need to be found and excluded from the dataset; otherwise, K-means create another cluster around those outliers.

For instance, Medina and Erath (2013) used a clustering algorithm – k-means – with activity start time and duration to detect only work locations under 10 clusters. Their study focussed on the study proposed by Chakirov and Erath (2012) due to overlapping study areas in Singapore. Chakirov and Erath (2012) detected work activities using a discrete choice model with activity start time, duration and land use data and estimated by using the Household Interview Travel Survey 2008, with a 97.5 per cent success rate. However, once they applied the same method to SCD, only 30 per cent of work locations/activities were identified. Therefore, Medina and Erath (2013) used Chakirov and Erath (2012)'s method to understand clusters. The main contribution of the study proposed by Medina and Erath (2013) focused on estimating the capacity of workplaces using building information data sources, i.e. building footprints, floor area and did not present the accuracy of identified work locations, which was the drawback of the method from trip purpose perspective. In addition, their study did not look at home locations from PAs.

2.4.2.3 DBSCAN clustering

DBSCAN is a density-based clustering algorithm that groups dense data points in a cluster. The unsupervised algorithm is robust to outliers and can find non-linearly separable clusters. The algorithm doesn’t need to find the number of clusters as a priori like in K-Means. However, the density of data points is important for this algorithm. Similar density points or large differences in densities create bias in the model. In addition, the algorithm works with Euclidean distance. Thus, data in high-dimensional spaces cannot be transferred in low-dimensional settings.

For instance, Ordóñez Medina (2018) presented a method to differentiate work and study activities from SCD. To achieve this, survey data were divided based on card type, i.e., work and study and then information and parameters were extracted using discrete choice models. The extracted parameters were used to create a 14-dimensional vector which was then fed to DBSCAN to present the 17 most popular weekly work and study patterns over seven days. The study was limited based on the two activity types (work and study) and its dependence on surveys. Few cities in developing countries have regular surveys (e.g., China) (Ordóñez Medina, 2018; Yang et al., 2019), so these methods may not be applicable on a broad scale. In addition, the contribution of the study is based on only work and study activities from SCD and does not cover the rest of PAs, i.e., home activities or SAs.

2.4.2.4 Hierarchical clustering

Hierarchical clustering (HC) is an unsupervised learning algorithm that groups similar objects into clusters. There are two types of hierarchical clustering. The first one is an agglomerative clustering (bottom-up approach), which starts from each cluster and merges the clusters to move up the hierarchy. The second one is a divisive clustering (top-down approach), which begins from all observations and splits them up to move down the hierarchy.

The HC algorithm is easy to implement and understand due to the dendrogram visualisation. However, the dendrogram representation implicates misinterpretation in large data sources with arbitrary decisions. In addition, the algorithm is not efficient for working with categorical and numerical values.

One of the recent studies in trip purposes inference is proposed by Faroqi and Mesbah (2021) from activity sequences using the agglomerative hierarchical clustering (AHC) algorithm. In their model, trip sequences were first created using the activity start time and activity duration. Then, the similarity was measured (Jaccard [index]) from individuals' daily sequences. Next, agglomerative hierarchical clustering (AHC) was applied to survey data and the best clusters were checked using the silhouette coefficient.

93 per cent of work activities, 99 per cent of home activities, 93 per cent of education activities, 94 per cent of shopping activities and 95 per cent of recreational activities were identified correctly from survey data. The same practice was applied to one-day SCD to represent the application of the model. Although the authors liked the idea of using unsupervised ML on a large dataset, the proposed model used limited features, i.e., only activity duration and activity start time and could not provide a solution without single-day survey data available. The model could be built upon by adding additional features, such as land use attributes due to the study area, Brisbane, having well-segregated residential and work areas (Wei, Liu and Sigler, 2015). While the accuracy in this model depends on the available survey data, not one-day SCD, the model has achieved higher accuracy than one of the latest trip purpose models proposed by Alsger et al. (2018) in the same city.

Other data sources are also used with this method to infer trip purposes. For instance, Liao, Fox and Kautz, (2007) applied hierarchical conditional random fields to find significant locations and activities using GPS data. Even though their results provide high accuracy based on manual labelling, the data size,i.e., four participants in seven days, is small compared to other studies with GPS.

2.4.2.5 Decision Tree

A Decision Tree (DT) is a supervised machine learning algorithm that supplies a useful structure to evaluate the possible options. The algorithm presents clear visualisation of the output, which can be decoded by humans easily. The algorithm can handle working with both numerical and categorical variables. DT can handle non-linear parameters unless there is a high non-linearity based on independent variables. In addition, the algorithm is robust, working with outliers. However, the algorithm cannot work well with noise and large datasets.

Within this mind, Lee and Hickman (2014) proposed another PA identification framework using decision tree classification called the trip purpose assignment process. The approach uses information from SCD such as card type, activity duration, activity start time, activity location and frequency to classify individuals using heuristic rules. Then, the extracted characteristics – such as card types to define users (adult or student), activity durations of less or more than 9 hours, AM/PM peaks and activity locations (downtown of the city and other locations) – were used for k-means to explain four types of clusters using decision tree methods. The training and testing results were then compared to household survey data. However, the study was applied to a small set of transit users over only five working days and required different classification methods to see the accuracy of baseline models, which the authors suggested as future work. Even though the study used spatial and temporal variables in their model, the study is limited to identifying only PAs.

DT is also used to infer trip purposes from other data sources. For instance, Deng and Ji (2010) presented an alternative method using GPS data to derive travel activities. A decision tree with adaptive boosting employed attributes such as timestamps, land use, type of trips, demographic and socioeconomic characteristics of users to construct six purposes (work, school, pick up/ drop off, shopping/recreation, business and others) with an 87 per cent overall accuracy rate.

2.4.2.6 Random forest (RF)

RF is a supervised machine learning algorithm based on the ensemble learning technique. The algorithm constructs many trees on the subset of the data and contains the output of all the trees. Thus, the model performs with higher accuracy compared to the decision tree. RF also works well with numerical and categorical variables. RF is stable once new data points arrive and even impact one of the trees in the model. In addition, the algorithm is less affected by noise and robust working with outliers. However, RF needs computational resources due to working with lots of trees. In addition, training the model is much longer than the DT.

One of the more recent works in trip purposes inference is proposed by E. Kim, Y. Kim and D. Kim (2020). Household travel survey data, including four activity types – commuting, returning home, business and leisure – were used to train the random forest model and obtained 83 per cent overall accuracy from the validation dataset. The main contribution of the study was feature selection using permutation-based variable importance, H statistic–based variable interaction and accumulated local effect (ALE). In addition, the work was concluded that temporal features had a more significant effect on inferring trip purposes than spatial features. Even though the details from the survey data (one day) were limited, including activity types, the study covers PAs and one of SAs, i.e., leisure.

Due to the limited number of studies inferring trip purposes from SCD, the rest of the latest works listed here are based on GPS data (Feng and Timmermans, 2013; Montini et al., 2014; Meng et al., 2017; Ermagun et al., 2017; Yazdizadeh et al. 2019; Gao, Molloy and Axhausen, 2021; Yang et al., 2021).

2.4.2.7 Neural Networks (NNs)

Neural Networks (NNs), or Artificial Neural Networks (ANNs), are supervised computational models and a subset of machine learning (ML). They are inspired by the human brain. The novel of the model is the structure of the network, which is framed by many nodes (neurons). Each node constructs signals using a mathematical function that is a multiple linear regression into a nonlinear activation function presented in independent layers, i.e., an input layer, a hidden layer and an output layer. The input layer contains the features and the hidden layer contains an arbitrary number of nodes greater than the input nodes. As a classifier, information entered from the input layer passes through the network from one layer to another and reaches the output layer.

ANNs have complexity after having two or more hidden layers called deep networks. In other words, deep learning (DL) directs to ANN with complex multilayers. DL models have several benefits compared to standard ML techniques: First, DL models can use the input data without feature engineering. The model can automatically capture the latent features to represent underlying patterns in the data and generate the desired output, which is called end-to-end learning. In addition, DL models have the capability to capture nonlinear relationships using nonlinear activation functions to detect all possible interactions between dependent and independent variables.

Even though NNs are used to investigate the issues in public transport (Padinjarapat and Mathew, 2013; Dacheng et al., 2018; Assemi et al., 2020), inferring trip purposes with the help of NNs is rarely used by GPS data, whereas no studies were found with SCD. For instance, Xiao, Juan and Zhang (2016) derived variables from GPS data (duration of the activity, mode of travel, day of the week as weekday or weekend) with land use characteristics and sociodemographic characteristics from smartphone survey data. After applying ANN, 96 per cent of overall accuracy was obtained using a ground truth dataset, which may not be readily available for other studies. Another study is proposed by Cui et al. (2018), predicting current and next trip purposes by matching Google POIs with historical Twitter data. In this study, the Bayesian neural network (BNN) has achieved improved accuracy for eating, recreation and shopping activities.

2.4.2.8 Summary

It has been suggested that more studies should use unsupervised learning to investigate trip purposes from SCD (Faroqi, Mesbah and Kim, 2018). However, trip purposes present more of a classification problem than a clustering problem (Xiao, Juan and Zhang 2016; Alsger et al. 2018; Nguyen et al. 2020; E. Kim, Y. Kim and D. Kim 2020). Clustering of spatial and temporal attributes of SCD cannot reveal the derived demand to investigate trip purposes (Kuhlman, 2015; Alsger et al., 2018).

Even though deep learning has been suggested for trip purpose inference from big data sources, i.e., SCD to improve public transit networks (Anda, Erath and Fourie, 2017), the current literature using SCD is limited due to methodological challenges, i.e., class imbalance, overfitting problems and data challenges, i.e., labelled data.

2.4.3 Statistical approach

Statistical methods are used to analyse events, data and problems based on statistically significant patterns or information using necessary mathematical formulas. The methods are easy to apply in small datasets. On the other hand, probabilistic methods can analyse a problem based on a probability distribution over the entire dataset and extract suitable events, which are dynamic (Yu et al., 2020). However, the accuracy is always based on the representativeness of the available dataset. Therefore, the estimated results may not be correct if the sample data do not cover all scenarios (e.g., activity types). Furthermore, these models are often unsuccessful in working with complex problems and incapable of handling large-scale scenarios with many dimensions. Therefore, statistical models are usually successful for small datasets or with aggregated values.

Several early studies proposed probabilistic models as alternatives to the heuristic approach to activity identification. Chakirov and Erath (2012) carried out one such study to detect home- and work-related activities based on SCD for public transport in Singapore. Although their study used a heuristic approach to identify work locations based on duration, it mainly focused on the discrete choice model with activity duration, start time and land use to distinguish home- and work-related activities. The duration of work activities was defined as 7 to 11 hours and their start time was between 6:00 and

11:00. For home activities, the duration was defined as 12–13 hours, while the start time for the vast majority of them was between 16:00 and 23:00. Then, the discrete choice model was applied to detect home and work activities using activity start time, duration and land use data, i.e., Household Interview Travel Survey 2008, with a 97.5 per cent success rate. However, when the same method was applied to SCD, only 30 per cent of work locations/activities were identified. Another limitation was that the size of the data, which only covered five working days, may have decreased the accuracy due to the lack of regularity.

Another probabilistic model approach was proposed by Li et al. (2015), who identified the most likely home and work pairs from SCD based on duration and frequency in Singapore. First, children, students and elders' travel records are excluded due to irrelevant or less relevant demographic groups for home and work pairs. Second, interchange stations were considered irrelevant or less relevant locations for home and work pairs and were excluded from the dataset. Using a rank aggregation technique and spectral analysis, the authors derived a comprehensive location ranking list in addition to identifying periodic travel patterns. After applying the model, the home location results were validated against the city’s urban planning dataset.

Han and Sohn (2016) also presented a probabilistic model derived from an unsupervised learning approach to identify activities from SC transactions. The proposed continuous hidden Markov model (CHMM) used emissions probability to find eight clusters interpreted as patterns for home and out-of-home activities. An advantage of this model was its ability to find cluster membership and generate activity chains to build simulation data without survey data. Although the model presented a way to discover activities and activity patterns, the model was limited to home and out-of-home activities (work) without providing accuracy. In addition, the study was applied to only one day data and the processing cost of such large amounts of SCD is uncertain in this approach (Anda, Erath and Fourie, 2017; Zou et al., 2016).

Zhao, Koutsopoulos and Zhao (2020) is one of the latest papers to uncover activities from SCD in London. The study focused on the location, start time of the day, start day of the week and duration of each activity using a Latent Dirichlet Allocation (LDA). However, the model ignored spatial dependencies when handling dimensionality, which was its limitation.

From other data sources, trip purposes were also investigated using a statistical approach. For instance, Gong, Liu and Wu (2016) proposed a model combining Bayes' rules with a Monte Carlo simulation to infer trip purposes from taxi data using POIs in Shanghai, China. The trip purposes were categorised into nine daily activity types based on temporal regularity, spatial dependency, trip lengths and directions.

Although the statistical approach has been applied for trip purpose inference from SCD, the main challenges in this method are twofold: First, the models depend on a sample dataset (i.e., survey data) to generalise the whole dataset, which may not be available to all researchers to follow the methodologies. Moreover, even if the survey data is available, it comprises limited sample data to represent the whole population. The second challenge of this approach is that statistical models are limited to working with complex problems and are incapable of handling large-scale scenarios with many dimensions. Therefore, statistical models have usually considered aggregated values from large datasets or disaggregated values from small datasets. However, this work mainly focuses on disaggregating individuals' daily travel behaviours from large SCD to infer trip purposes using a data-driven approach.

2.5 Summary

The limitations in terms of data sources and methodologies for inferring trip purposes from SCD can be summarised as follows:

First, trip purposes are investigated through conventional household surveys, which are limited by low update frequencies and limited activity types and may not be available for many cities (Amaya, Cruzat and Munizaga, 2018). There is a need to investigate trip purposes using big data sources, i.e., SCD, for public transport networks.

Second, PAs (home and work/school for adults/students) are relatively possible to infer/identify trip purposes using the characteristics of SCD. However, determining 'other activities', or SAs, require more attention, which facilitates investigating people's use of spare time and surplus income in cities. Further research is necessary to enable SA trip purposes to be inferred, given the current challenges and limitations of SCD.

Third, there is inherent complexity in human behaviour displaying different temporal regularities, such as weekday/weekend temporal patterns or daily/weekly travel patterns per individual (Goulet-Langlois, 2016). In addition, enriching SCD with other big data sources, e.g., POIs, is necessary to infer trip purposes. Nevertheless, spatial complexity adds another limitation once data sources become large with high dimensions. Thus, there is a need to investigate learning methods in dynamic cities.

Finally, possible data sources have been reviewed from the literature. None of the data sources has fully helped SCD for validation purposes or provided a solution to investigate further trip purposes from SCD (e.g., using sophisticated methods in data mining). To surpass these limitations, there is a need to collect suitable datasets for this research to investigate trip purposes from SCD.

2.6 Chapter summary

This chapter has introduced a literature review to infer trip purposes from SCD, including relevant research domains, data sources, methods and materials. First, relevant research domains for the study were manifested in Section 2.1 to highlight the importance of the research. Then, data sources were presented in Section 2.2 to investigate trip purposes from SCD. Next, a systematic review of trip purposes from SCD was presented under three sections with input features in Section 2.3.1, enrichment of SCD in Section 2.3.2 and applied methodologies in Section 2.4, such as the heuristic approach, ML approach and statistical approach, including the scope and limitations of each method. Finally, the research gaps that were missing in the existing literature are summarised in Section 2.5. The chapter ends in Section 2.6.

Chapter 3

Methodology

3 METHODOLOGICAL FRAMEWORK

This chapter introduces the methodologies used in this study. First, Section 3.1 presents the overall workflow for inferring trip purposes from Smart Card Data (SCD). Second, Section 3.2 presents a new data source – Survey Smart Card Data (SSCD) – and the methodological approach to collecting SSCD. Then, Section 3.3 details the structure of the study's databases. Fourth, Section 3.4 explains the proposed methodologies under two approaches, i.e., the heuristic approach and the machine learning (ML) approach. Lastly, Section 3.5 summarises the content of this chapter.

3.1 Methodological framework

The methodological framework is divided into four sections, i.e., input, pre-processing, proposed models and output, as illustrated in Figure 3.1. The framework starts with 'data' as input to present the data sources and their flows into the data processing section. The study has four types of input data: 1) SCD collected from Automatic Fare Collection (AFC) systems to represent individuals' movements in transport networks without trip purposes and demographic attributes. 2) SSCD, which is SCD with additional information such as trip purposes and demographic attributes representing the survey data collected by volunteers. 3) London Travel Demand Survey (LTDS) data collected by Transport for London (TfL), including demographic details and trip purposes (e.g., home and work). 4) Land use data, i.e., POIs, collected from Foursquare representing the characteristics of land use attributes.

The second section of the methodology focuses on data pre-processing, explaining how data sources are processed and activities are extracted. The aim of data pre-processing is to improve the accuracy of the models with pre-steps, such as cleaning the dataset, excluding incomplete trips, etc., before applying the proposed methodologies. For instance, after extracting activities from SCD, or SSCD, if travel data require linkage to POIs, data need to move to 'activity-POI consolidation algorithms'. Enriched data is used for the ActivityNET framework that uses large data sources, i.e., SCD and land use data from Foursquare POIs, along with deep learning techniques to predict trip purposes for public transport. The data pre-processing steps are detailed in Chapter 5.

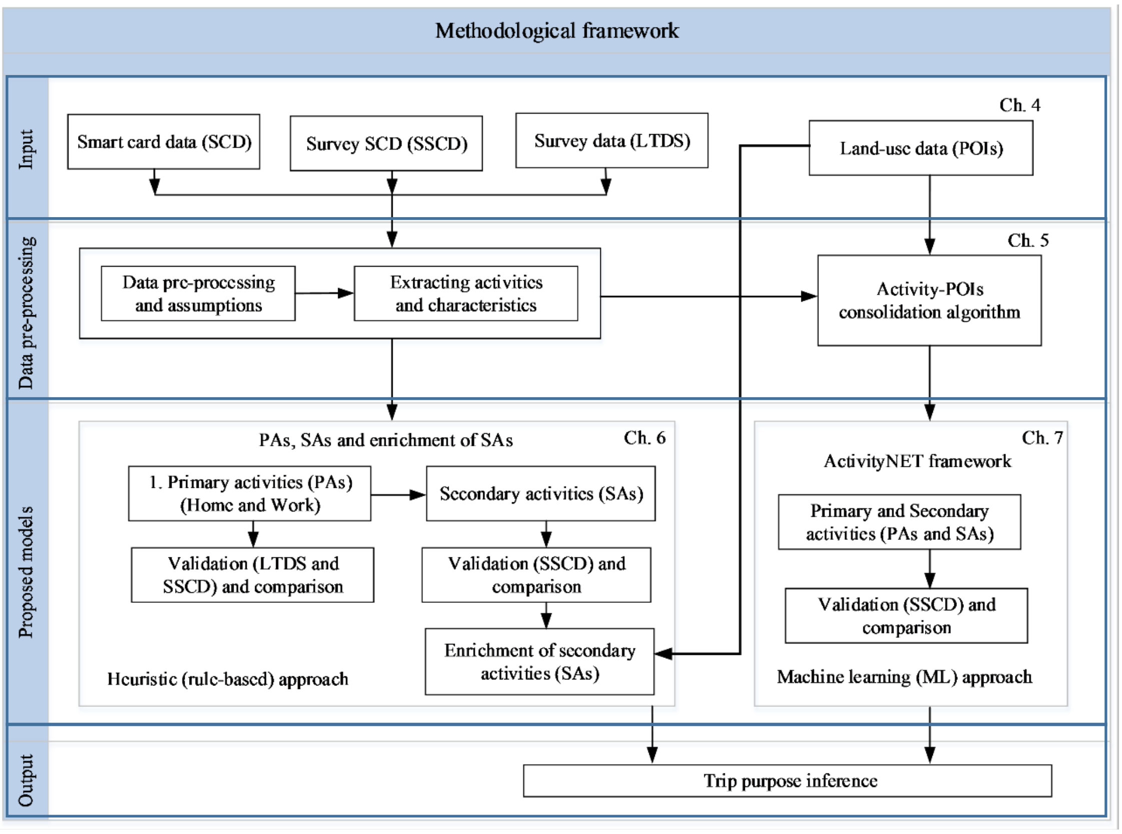

Methodological framework
Input
Data pre-processing
Proposed models
Output
Smart card data (SCD)
Survey SCD (SSCD)
Survey data (LTDS)
Ch. 4
Land-use data (POIs)
Data pre-processing and assumptions
Extracting activities and characteristics
Ch. 5
Activity-POIs consolidation algorithm
PAs, SAs and enrichment of SAs
Ch. 6
1. Primary activities (PAs) (Home and Work)
Secondary activities (SAs)
Validation (LTDS and SSCD) and comparison
Validation (SSCD) and comparison
Enrichment of secondary activities (SAs)
Heuristic (rule-based) approach
Ch. 7
ActivityNET framework
Primary and Secondary activities (PAs and SAs)
Validation (SSCD) and comparison
Machine learning (ML) approach
Trip purpose inference

Figure 3.1 Overall methodological framework.

The third section of the study proposes trip purpose identification methods from SCD using two approaches – a heuristic approach and a machine learning approach (ML), presented in Chapters 6 and 7, respectively. Figure 3.1 presents the flow of the steps visually. First, the heuristic approach identifies primary activities (PAs) and secondary activities (SAs) from pre-processed SCD. After validation and comparison of both models, station-based SAs are enriched using activity-POI consolidation algorithms. Second, the ML approach uses linked travel data, SCD or SSCD, with POIs to predict PAs and SAs – the ActivityNET model. This section also presents the validation of the ActivityNET model and its comparison with baseline models.

Finally, the methodological framework ends with trip purposes inference from SCD as an outcome of this research, gathered from the previous three steps such as input: data, data pre-processing, including proposed models using heuristic and ML approaches.

3.2 Data collection process

There are challenges (lack of labelled information) and limitations (tap-in/out stations) associated with inferring trip purposes from SCD mentioned in Section 2.2.4. To overcome these issues, data collection process is started to investigate and validate activity type information, which helps to facilitate a better understanding of trip purpose from SCD. The methodological steps during the survey data collection are explained as part of this section and SSCD and the corresponding attributes are detailed in Section 4.2.2.

SSCD were collected between 2017 and 2020. A total of eighty-four volunteers are surveyed during this period. However, the analysis carried out for this research is activity-based but individuals, which means one individual contributes approximately 120 data points/activities in the two-month dataset. The study focussed on convenience sampling, which is a non-probability sampling method and contains individuals who are easy to access for the researcher (Pro, 2021). This technique is a straightforward and inexpensive way to gather data but may not represent the whole population in the study area (Frey, 2011). Thus, convenience sampling may involve sampling bias in the study area.

During the data collection process, regular transport users are considered for the research to capture full OD information. This is mainly for train and tube trips in our study; however, the proposed methodologies are applicable to any transport modes as long as ODs are available. Different backgrounds, such as students, professionals and non-working professionals, including all genders and ages, can be captured from SSCD. A limitation of the sample data in terms of demographic attributes is that neither retirees over 55 nor people under 18 years of age were included, which means demographic bias also exists in the dataset.

The data collection process is explained by steps that reduce data collection biases and produce reliable results (Pro, 2021). For instance, to make the steps straightforward during the registration, four types of data collection templates are prepared for this study: the Oyster card template, the contactless payment card template, Google activity location data (a location history data for users who has Google account) and questionnaires to collect volunteers' demographic attributes. Each template is presented in Appendix A. Second, labelling the data points twice, i.e., one for Labelling-I and another one for Labelling-II reduces misclassification bias in the dataset (Šimundić, 2013). Besides, cross-validation has been applied to SSCD, which means dividing the proposition of the dataset into k folds and using one fold for validation and the rest for training also reduces the biases in the proposed methodology in Section 7.2.

3.2.1 Card registration and data download

Each individual is asked to register their available transport cards, such as Oyster cards and contactless cards. If the registration process is incomplete, TfL does not allow anyone to download users' transit data.

Differences between the contactless and Oyster card data appear once the individuals download their data. Thus, the downloaded data need to be restructured. As long as the person has registered their card details and credentials, i.e., name, date of birth, address, email and phone number, they can connect to and download their SCD and contactless card data. Then both data sources are combined under the same person's details after formatting one file (contactless) to another file (SCD).

The data allowance depends on the person's request. For instance, if a person requests each month's data, TfL sends the person's monthly data on a regular basis. In our dataset, eleven people had already been collecting their travel data monthly and provided more than a year of labelled transit data for this study. However, when a person downloads their travel data for the first time without any previous request, only data from the previous two months is available from the date of request.

3.2.2 Data processing and labelling

Once individuals download and save their data, the data need to be pre-processed using the steps described in Section 5.1, e.g., excluding single trips in a day and missing information such as tap-out. Then, extracted activities are labelled by volunteers using the two different classifications presented in Table 3.1: 1) Labelling-I contains home, work/study, before-work, midday, after-work, undefined and 2) Labelling-II contains home, work/study, entertainment, eating, shopping, drop-offs/pick-ups, part-time-work.

This classification was used during the labelling process because of the validation requirements of this study's proposed models. However, the methodology for data collection can employ different classifications in future research.

SCD are then labelled to mitigate the repetition of some types of activities. For example, consider an individual who provides two months of SCD, comprised of approximately 120 trip records. After processing and extracting activities, the sample data is left with fewer than 60 activity records (two consecutive trips equal an activity), which need to be labelled. The majority of the time, 60 per cent to 70 per cent of the data is labelled as PAs (home and work activities), so the rest of the activities (SAs) include around 20 to 30 activity records. Thus, the volunteers' participation time in this process was mainly for SAs.

Table 3.1 Feature labels to identify trip purposes

Class	Features	Name	Definition
Labelling-I	H	Home	Time spent at home
	W/S	Work/Study	Time spent at work/study
	BW	Before-work	Time spent at any location before going to work
	MD	Midday	Time spent at any location from work to work
	AW	After-work	Time spent at any location from work to home
	UD	Undefined	Time spent at any location out of any conditions above
Labelling-II	H	Home	Time spent at home
	W/S	Work/Study	Time spent at work/study
	ENT	Entertainment	Time spent at any location for entertainment purposes
	EAT	Eating	Time spent at any location for eating purposes
	SHO	Shopping	Time spent at any location for shopping purposes
	D/P	Drop-offs/Pick-ups	Time spent at any location for drop-offs/pick-ups
	PTW	Part-time-work	Time spent at any location for part-time work purposes
	O	Others	Time spent at any location out of any conditions above

In addition, the 'questionnaire' part of the survey, mentioned in Appendix A, provides additional information about the volunteers' travel behaviour and their demographics.

3.2.3 GDPR and challenges during the data collection process

According to nine volunteers' requests, we have applied GDPR (General Data Protection Regulation) rules and signed a document (ICO, 2018), which requires that their data be used for only this work and not for any other purposes. At the end of the data collection process, personal information is removed and all data is anonymised for further analysis.

The data collection process was challenging due to privacy issues. Two specific challenges were noted during the data collection process related to information sharing and the time required to participate.

Information sharing is one of the issues for volunteers during the data collection process because people are wary of sharing data containing their home and work locations, even though SCD is based upon stations. For instance, activity types are being provided by some volunteers, but not journeys. Therefore, before the data collection process, a detailed letter is provided to the individuals explaining why this project is important, how their data is useful and, more importantly, how we anonymise their data to make their credentials (i.e., name and card details) invisible. Further, participants are given the opportunity to receive more information on the usage of their data after running our models/algorithms.

On the other hand, sharing information through Google activity location data, which provides location history for users who have Google accounts, was an issue during the data collection process. Even though GPS-based data sources offer more information, people rarely like to share such details (only four volunteers) due to privacy concerns. Besides, due to this limitation, the study mainly focuses on SSCD. Details about Google activity location data and analysis are accessible in Appendix A.

The time required to participate is another issue for volunteers because the entire process of downloading, processing and labelling data take approximately 45 minutes. Therefore, we offer to assist volunteers in data downloading and labelling to minimise their time on these processes.

3.3 Data architecture

In this section, the data architecture is illustrated in Figure 3.2 to represent how data are stored and transformed to other platforms in this study. PostgreSQL, a relational database management system, is used for storing the data in four databases for POIs, SCD, SSCD and LTDS. Python (programming language for data processing and analysis) and PySpark (Python-based API for Apache Spark for big data processing and analysis) are used to process and analyse the data. The main tools and libraries used are matplotlib, NumPy, pandas, sklearn, seaborn, Keras, TensorFlow and PySpark. Data cleaning and processing (Section 5.1) and extracting activities (Section 5.2) from the SCD and LTDS data, as well as the application of the proposed methods and algorithms, are achieved in Python using the corresponding data sources from the databases. Their assumptions and details are discussed in Chapter 5.

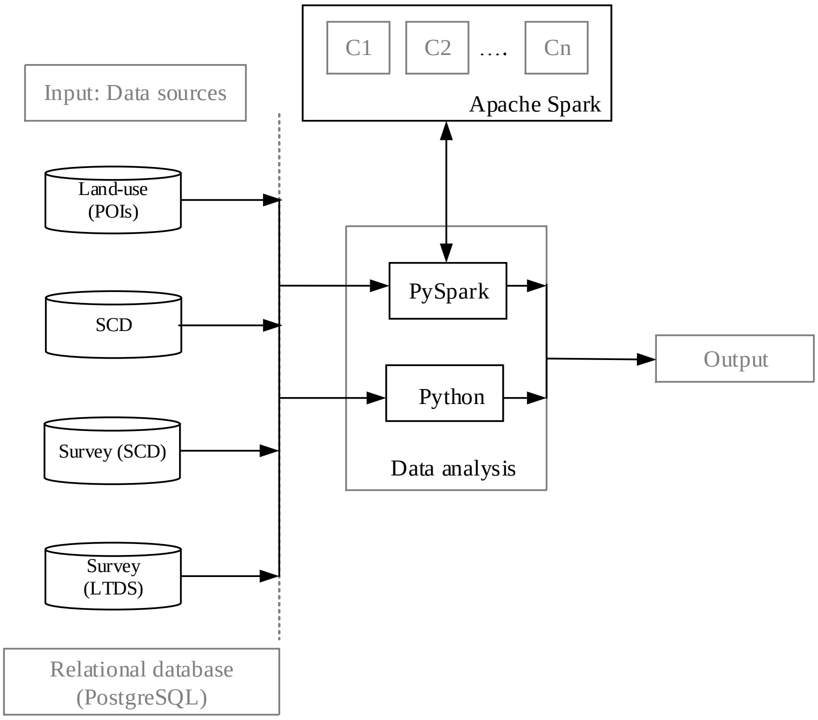


Figure 3.2 Data architecture of the research.

Combining travel data (approx. 2 million records) with POIs (approx. 81 million check-ins) increases the volume of the data. Data processing packages such as pandas in Python are designed to work on a single machine and their memory allowance has a low threshold which can create memory issues. Therefore, these processing steps are executed using the Apache PySpark framework. PySpark works as an interface for Apache Spark and is used for the activity–POIs consolidation algorithms to match each activity with relevant POIs around tube and train stations. The relevant section, the activity–POIs consolidation algorithms, is explained in more detail in Section 5.3.

The ML part of the study uses artificial neural networks (ANN) as described in Section 7.2. Though the training process using labelled data (approx. ten thousand data points) takes about 5 minutes, predicting values from the trained model (more than one hundred thousand points) takes about 30 minutes, illustrating a significant slowing of the process. Even running a grid search to decide on the best parameters for the model is a huge task for Python. To improve efficiency (i.e., running time), the methodology is also decoded to PySpark to increase the speed of the workflow. The output is obtained by either Python or PySpark following data processing and analysis, as shown in Figure 3.2.

3.4 Trip purpose identification methods

This section explains the proposed methods applied in this research to infer trip purposes from SCD in detail using two approaches, i.e., a heuristic approach and an ML approach, as follows:

3.4.1 Heuristic approach

The proposed trip purposes inference model using heuristic approach from SCD is introduced in three sections: 1) the PAs to identify primary locations and activities, 2) the SAs to identify secondary locations and activities and 3) the enrichment of SAs to present station-based enrichment, including trip purpose inference.

The heuristic activity identification models start with journey (trip) counts as an indicator of usage regularity in Figure 3.3. The PAs re-define two pragmatic assumptions proposed by Barry et al. (2002), which are used in literature for OD estimations (Cui, 2006; Trépanier, Tranchant and Chapleau., 2007; Barry, Freimer and Slavin, 2009; Nassir et al., 2011; Munizagaa and Palma, 2012; Devillaine, Munizaga and Trepanier, 2012; Munizaga et al., 2014; Alsger et al., 2015; Zou et al., 2016; Zhao et al., 2017; Alsger et al., 2018; Huang et al., 2020). The first assumption states that the majority of commuters return to the destination station of their first journey to commence their next journey. The second states that a substantial number of people finish the last journey of the day at the station where they started their first journey of the day. The re-defined assumptions for home locations are as follows: first, the origin/start station of the first journey and the destination/end station of the last journey of the day are the same or within walking distance (≤ 800 m) for each user on each day. Second, if the station has passed the first condition and appears more than the defined visit-frequency (or frequency) threshold, the station is marked as the home station for the individual. Although Zou et al. (2016) made a similar assumption, the home station/activities identification algorithm contributes to the literature by combining Barry's assumption with the frequency threshold (Hasan et al., 2012) and a distance threshold (walking distance, ≤ 800 m) to increase the accuracy of the findings. Based on the data available for analysis in the study, no assertions can be made about the home or work location assumptions for specific groups such as working parents, tourists etc.

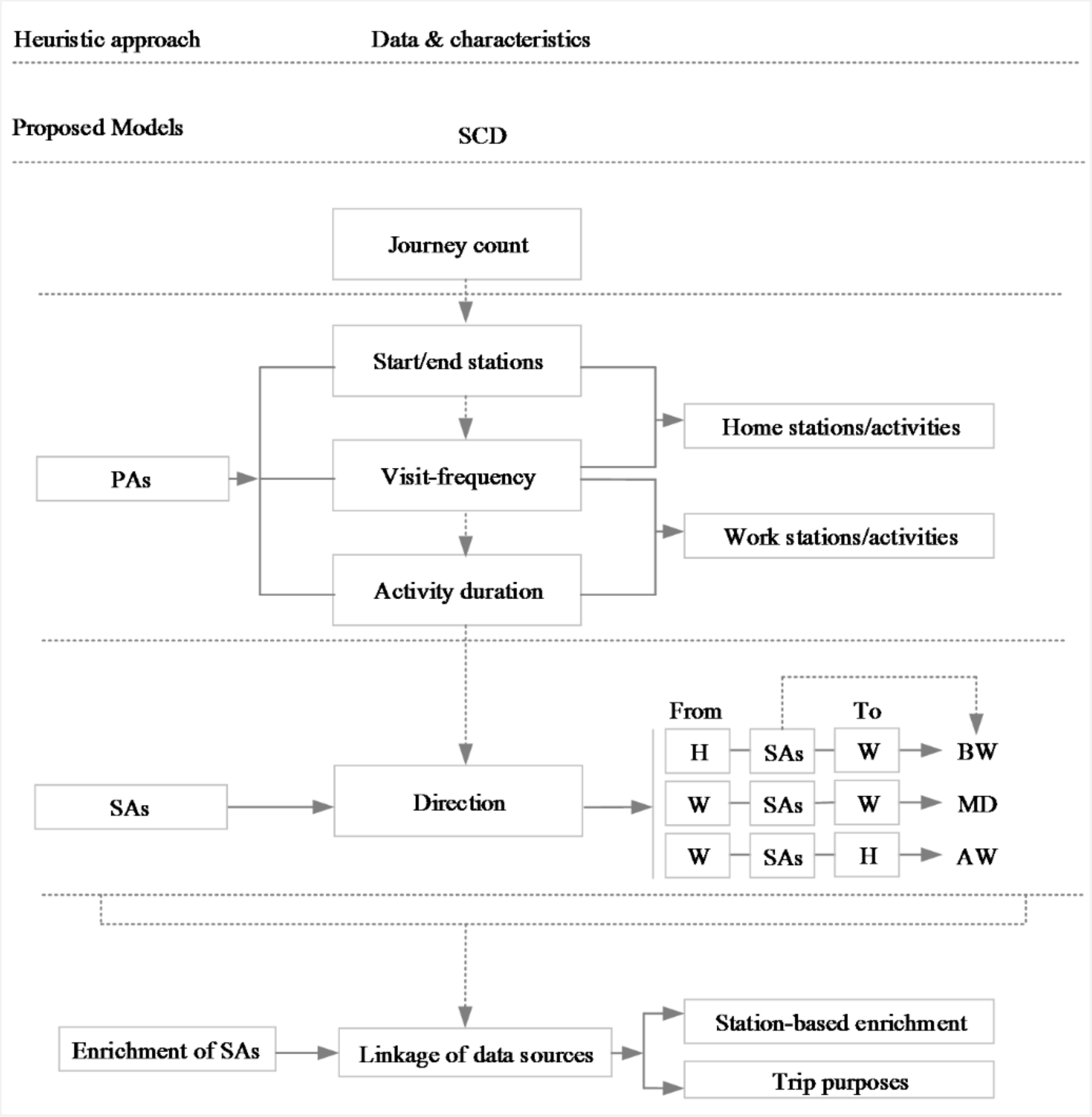


Figure 3.3 The logical flowchart of activity identification using the heuristic approach.

After identifying home stations, all consecutive journey pairs for all working days are evaluated. In the model, the destination station of the first journey and the origin station of the second journey in the journey pairs are selected. If the selected stations match or are within walking distance (≤ 800 m), the activity duration is extracted using the destination time of the first journey and the origin time of the second journey. Then, the result is compared to the defined stay-time and visit-frequency thresholds to be able to label the station/activity as the work station/activity. Thus, the work location identification algorithm contributes to the literature by combining a stay-time criterion proposed by Devillaine, Munizaga and Trépanier (2012) and visit-frequency of consecutive journeys proposed by Hasan et al. (2012). The combination of both indicators and the walking distance criterion is the basis of work location identification in this study.

After identifying PAs, the rest of the activities are further investigated using direction (from and to) information in the SAs. The from-activity location (FL) is defined as the last location before an activity, or where the individual came from. The to-activity location (TL) is defined as the next location after an activity, or where the individual is headed to next. Thus, if where a person comes from is H and where they're headed is W, then the SAs are labelled as before-work (BW); if where they came from is W and where they're headed is W, then the SAs are labelled as midday (MD); and if where they came from is W and where they're headed is H, then the SAs are labelled as after-work (AW). The rest of the activities are labelled as undefined activities (UD) in this model.

In the literature, only one type of SA (after-work activities) is investigated from public transport using time constraints (Wang et al., 2017). The first constraint is calculated as the threshold of the earliest time of departure from work and the second constraint is the finishing time of the tube lines, restricting individuals' after-work activities to before midnight. However, in this study, SAs are extracted using anchor points as well as the direction information of those locations. Therefore, the proposed algorithm can capture individual-level starting and ending working hours (flexible working hours) as a holistic picture with before-work and midday activities as captured in travel surveys (Rasouli and Timmermans, 2015).

Furthermore, the proposed models are validated under two approaches. The first one is comparing the proposed models to benchmark models from the literature. The second one is using the proposed methodology based on the ground truth data sources, i.e. surveys. However, each proposed model has a different benchmark model, i.e., Hasan et al., 2012; Wang et al., 2017 presented in Sections 6.1.2.4 and 6.2.3.2, respectively.

3.4.2 ML approach

The proposed trip purposes inference framework, ActivityNET, using ML approach from SCD contributes to the literature predicting passengers' trip purposes for each activity per individual from their SCD using deep learning (DL) method not only for PAs, but also for SAs. The proposed framework is introduced in two phases. Phase 1 focuses on extracting activities from the travel dataset and links the spatial and temporal attributes of SSCD, including the attributes of the POIs under three sub-sections – spatial information match (location of stations and POIs), temporal information match (activity start and end time and opening and closing hours of POIs) and attractiveness (the number of activities and check-in).

Phase 2 focuses on prediction under two sections – model training (Phase 2A) and model testing (Phase 2B) after splitting data into training and testing datasets. The training dataset is a subset of the labelled dataset, which is 70 per cent of the SSCD. The testing dataset is the rest of the labelled data and used as unseen data to evaluate the model further. In the model training (Phase 2A), the DL model learns the relationship between the input (features) and output (labels) variables using each data point from the training dataset. During the model training, the training loss values – the difference between predicted and actual values – are checked while tuning hyperparameters, i.e. the number of neurons (dimension of the input variables), activation functions (a mathematical function to decide which neuron passes the next layer), epochs (the number of iteration) and batch size (the number of training units utilized in one iteration) to determine the best parameters with grid search techniques (one parameter is changed while others remain unchanged) (Brownlee, 2020b). In addition, to evaluate the model performance, dropout rate (the probability of neurons ignored in each layer), early stopping (the number of training epochs before the model performance stops improving) and k-fold cross-validation are also done in this section. In the model testing (Phase 2B), the trained DL model is evaluated with test (unseen) data to measure its performance. To achieve this, predicted values from unseen data, i.e. trip purposes, are evaluated further using a confusion matrix and three performance metrics in each class –precision, recall and F1-score (Brownlee, 2020a).

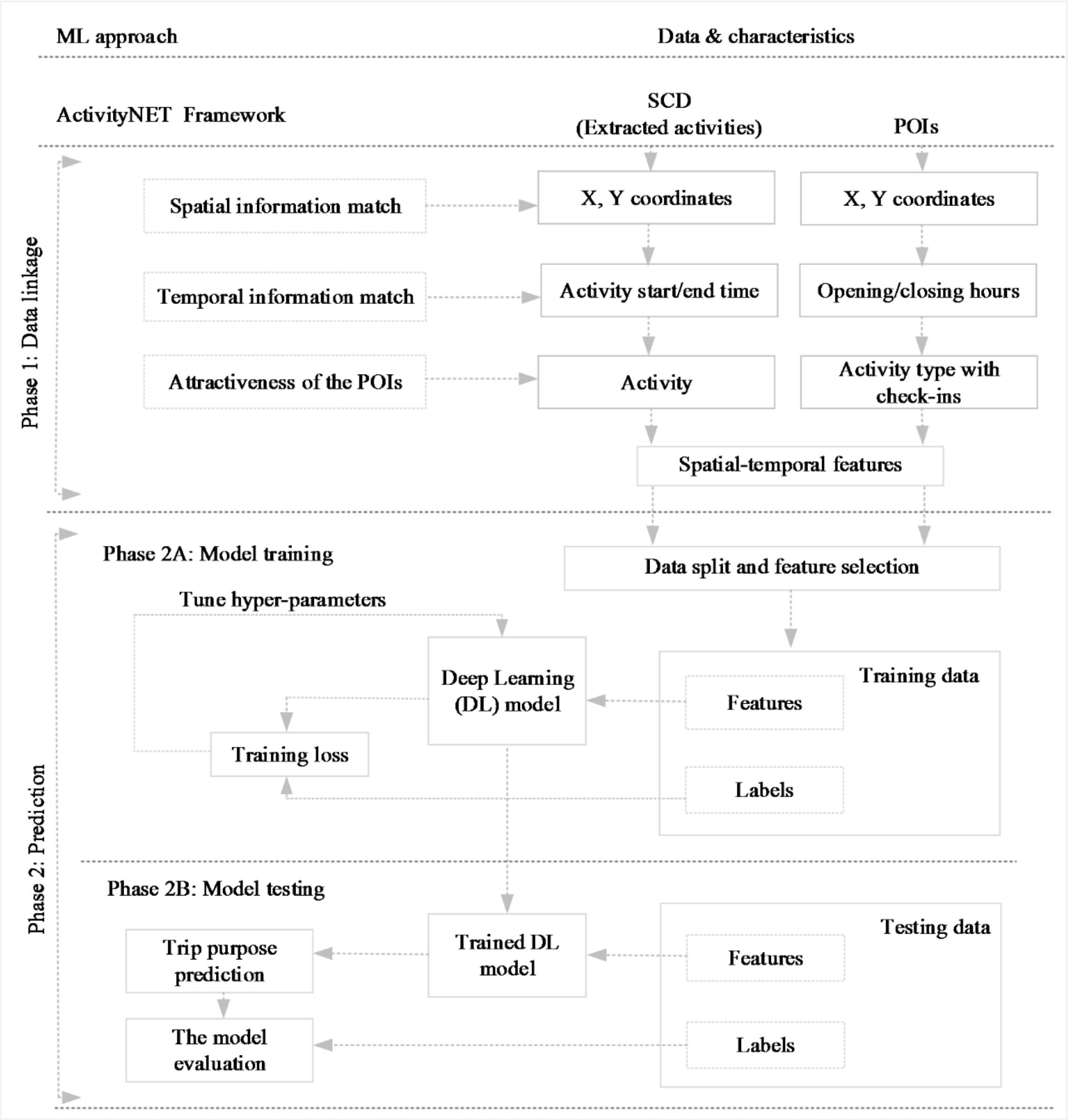


Figure 3.4 The logical flowchart of the ActivityNET framework using the ML approach.

Furthermore, the proposed model is validated under three approaches. The first one is evaluating the model from unseen (test) data using three performance metrics in each class –precision, recall and F1-score (Brownlee, 2020a), including the confusion matrix (Section 7.3.2). The second one is comparing the proposed models to benchmark models from the literature using the ground truth data sources (Section 7.3.2). The last one was comparing results based on benchmark models from literature (Alsger et al., 2018).

3.5 Chapter summary

This chapter has presented information about the methodological framework of this study in terms of various dimensions/ways. The overview of the methodological framework, first, was introduced in Section 3.1 under three steps, i.e., data, data pre-processing and the proposed frameworks. Next, new data, i.e., survey smart card data (SSCD) and the methodology of the data collection process, are briefly explained in Section 3.2, such as card registration, data downloading, data processing and data labelling, including GDPR and challenges during the study. Third, the data architecture is explained in Section 3.3 to present/highlight how data is stored and connected to each other during the study. Finally, proposed trip purposes methodologies are explained in detail using heuristic and ML approaches for this study, including the chapter summary in Section 3.4 and Section 3.5, respectively.

Chapter 4

Data Description

4 DATA DESCRIPTION

This chapter describes the study area and data sources analysed. In Section 4.1, the study area is briefly described in two sections: London as a case study (Section 4.1.1) and the public transport network in the British capital (Section 4.1.2). Three types of travel/transport data sources are introduced: London's Smart Card Data (SCD), provided by Transport for London (TfL) (Section 4.2.1); Survey Smart Card Data (SSCD), collected and labelled by volunteers for this study (Section 4.2.2); and London Travel Demand Survey (LTDS), collected and matched to relevant users' IDs (matched journeys) by TfL (Section 4.2.3). In Section 4.2.4, the land use data is presented using Point Of Interest (POIs) as auxiliary information from Foursquare data. Data challenges in Section 4.3 are discussed before summarising the chapter in Section 4.4.

4.1 Study area

4.1.1 London as a case study

London is the capital city of the UK located in South East England and has one of the most comprehensive public transport networks in the world. According to the 2021 census, the city has a population of approximately 9.2 million. London is the main transport hub in the UK, with international railway stations such as St Pancras, London Bridge and Victoria and airports including Heathrow, Stansted and Gatwick. In addition, the city embodies global connections through its finance sector for international professionals, which makes London an interesting area for study.

The study area comprises 32 London boroughs (districts) and the City of London, as is illustrated in Figure 4.1. The interior part of Greater London with 12 boroughs is called inner London and covers the City of London, Camden, Hackney, Hammersmith and Fulham, Haringey, Islington, Kensington and Chelsea, Lambeth, Lewisham, Newham, Southwark, Tower Hamlets, Wandsworth and Westminster. The rest of the London boroughs surrounding the inner London area like a ring are called outer London. Twenty boroughs in outer London are Barking and Dagenham, Barnet, Bexley, Brent, Bromley, Croydon, Ealing, Enfield, Greenwich, Harrow, Havering, Hillingdon, Hounslow, Kingston upon Thames, Merton, Redbridge, Richmond upon Thames, Sutton and Waltham Forest. Moreover, the central London area, which is in inner London, covers the City of London, including most of Westminster, Camden's inner parts, Islington, Hackney, Tower Hamlets, Southwark, Lambeth, Kensington and Chelsea and Wandsworth. The City of London, central London and inner-outer London terminologies are used in Sections 6.1 and 6.2 of the study. In addition, the map represents the postcode district level representation in London as detailed in the validation process using LTDS data in Section 6.1.2.4.

4.1.2 London's public transport network

London has one of the oldest underground transport systems and the largest urban public transport network in the world and covers 400 km with 270 stations. The underground public transport system, or the Tube, opened in 1863 between Paddington and Farringdon and was named the Metropolitan line. It was later extended to include the Hammersmith and City line (1864), District line (1868), Circle line (1884), Waterloo and City line (1898), Central line (1900), Bakerloo line (1906), Piccadilly line (1906), Northern line (1937), Victoria line (1960) and Jubilee line (1979) (Transport for London, 2021).

Today, London's public transport systems provide multi-modal interchanges between (1) London's bus service, (2) London Underground/metro service, (3) London Overground service, (4) light rail services (London Tramlink and Docklands Light Railway), (5) ferries (the River Bus) and (6) most National Rail (NR) services within the London fare zones. Transport for London (TfL), part of the Greater London Authority, is responsible for executing the Mayor's Transport Strategy, including the planning, delivery and operation of the public transport system.

Figure 4.1 illustrates the study area covering all London boroughs and the Tube network. North London is better connected than South London in terms of the transport network. The inner London areas are better connected than outer London.

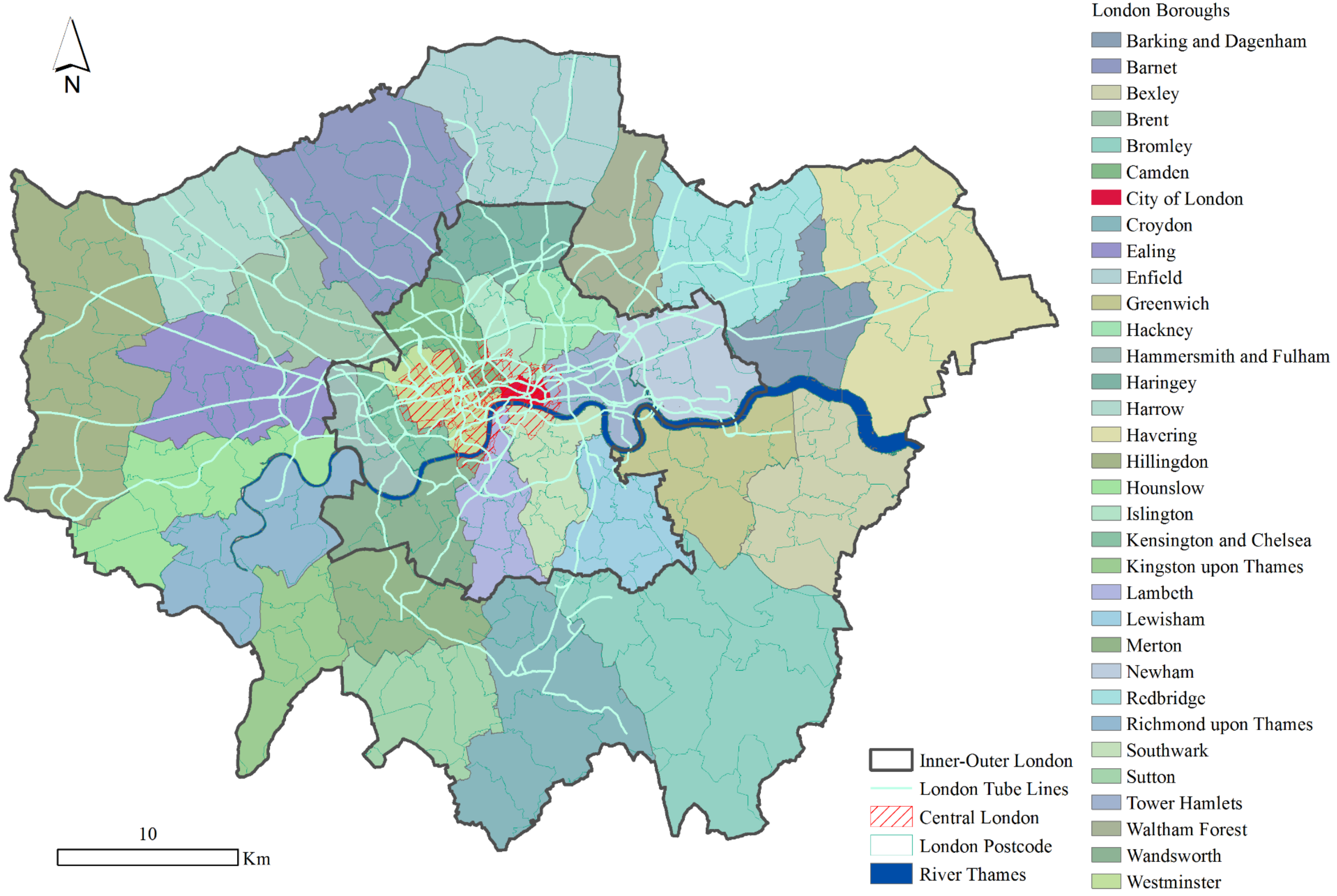


Figure 4.1 London as a study area is illustrated with its underground transport (tube) network

4.2 Data

The study has four types of input data: 1) SCD collected from Automatic Fare Collection (AFC) systems to represent individuals' movements in transport networks without trip purposes and demographic attributes. 2) SSCD, which is SCD with additional information such as trip purposes and demographic attributes representing the survey data collected by volunteers. 3) London Travel Demand Survey (LTDS) data collected by Transport for London (TfL), including demographic details and trip purposes (e.g., home and work). 4) Land use data, i.e., POIs, collected from Foursquare representing the characteristics of land use attributes. The next section will provide more information about the data landscape for this study.

4.2.1 Oyster card data (SCD)

Oyster cards are smart cards that collect large quantities of detailed transaction data on the TfL network. The cards are valid on all London public transport systems and records individuals' journey data automatically, excluding personal user details, when a passenger taps in or out at a station, e.g., tube/train stations (only tap in is required for buses and trams).

A sample of daily SCD for an individual is presented in Table 4.1. Attributes of the daily movements of individuals, such as boarding and alighting time, boarding and alighting station and transport mode, are considered for the study.

Table 4.1 Oyster card daily data for an individual (Prestige ID is a unique user identifier).

Prestige ID	Date	Boarding Time	Alighting Time	Boarding Station	Alighting Station	Mode
101519434	03/01/2014	08:36	09:15	Enfield	Oxford Circus	Underground
101519434	03/01/2014	17:21	17:56	Oxford Circus	Enfield	Underground

For this study, 10,000 individuals (anonymised by TfL) were randomly selected for the months of October and November 2013. The sample contained a total of 1,823,906 records of completed journeys by individual users out of 60,251,475 total transactions.

Although the Oyster card may be used on multiple modes of transportation across London, one of the limitations identified in the TfL dataset is the incomplete recording of trip information for bus journeys. As TfL does not currently capture the alighting information from its bus trips, bus journeys are often excluded from the trip analysis. Such journeys, however, can be included with an enhancement of the model in which missing information is identified in a sub-step of the identification process (Gordon, 2012). The second limitation is the absence of socioeconomic factors (e.g., age, career, income and home/work location) due to privacy concerns. The third limitation is the absence of trip purpose, which makes it difficult to utilise SCD (Bagchi and White, 2005).

4.2.2 Survey Oyster card data (SSCD)

A volunteer survey is conducted under three headings: 1) Labelling-I with sub-categories H (home), W/S (work/study), BW (before-work), MD (midday), AW (after-work) and U (undefined); 2) Labelling-II with sub-categories H (home), W/S (work [full-time]/study), ENT (entertainment), EAT (eating), SHO (shopping), D/P (drop-off/pick-up), PTW (part-time-work) and O (others); and 3) demographic information including age, gender and income. Thus, the survey can be used to validate the results of the models presented in Sections 6.1, 6.2 and 7. In addition, the method for data collection of SSCD is explained in Section 3.2.

SSCD were collected between 2017 and 2020. The following steps are taken during the data collection process. First, each volunteer registers with TfL and downloads their travel data (19,792 trip records). The data are cleaned and processed (Section 5.1) and 9316 activity/data points are extracted as activities. The volunteers are asked to provide two columns of trip purpose information. The first column (Labelling-I) requires them to label activities under six sub-categories: home (3994 data points), work (2006 data points), before-work (421), midday (206), after-work (2573) and undefined (166). The second column (Labelling-II) requires them to label the activities based on seven sub-categories: entertainment (555 data points), eating (687), shopping (818), child drop-offs/pick-ups (629), part-time work activities (427) and other (200 data points).

In addition, volunteers also provide demographic information – i.e., gender, income and age. Of the activity points, 5387 are from female volunteers and 3729 from male volunteers. The details of the data are further subdivided into four income bands: 2486 data points represent no income, 1657 data points represent earnings below £25,000, 2901 points represent earnings between £25,000 and £40,000 and 2072 represent earnings of more than £40,000. Similarly, data are divided into three groups based on the ages of the participants: 3867 data points are from volunteers under 30 years old, 3453 are from those between 30 and 40 and 1796 are from those over 40 years old. Under the occupation group, 4972 of the activities are titled as professional and 4144 as students. Figure 4.2 illustrates activity/data points labelled according to the Labelling-I and Labelling-II, including demographic attributes such as gender, occupation, income and age for this study. At the end of data collection and processing, individual data are anonymised under GDPR rules (ICO, 2018). More information about data collection is given in Section 3.2.

4.2.3 London Travel Demand Survey data (LTDS)

The LTDS data are based on a detailed household questionnaire focused on transit use in Greater London (TfL, 2011). London Travel Demand Survey (LTDS) is collected face-to-face by Transport for London [TfL] annually in London. The sample size of the survey is limited due to intensive labour and other costs. Eight thousand randomly selected households took part in the survey for the whole population [of the London] (more than 9 million) (TfL, 2011).

Table 4.2 An individual in LTDS data with relevant information.

Name	Label
Prestige ID	101519434
Working Status	Full-time paid employment (30+ hours a week)
Position	Employee
Mode	Underground
Home	Enfield (N21)
Sex	Female
Age	26
Work	14 Westminster (W1C)

The surveys are separated into two parts, as illustrated in Figure 2.1. The first part of the LTDS data focuses on household-level demographic information, such as gender, age,

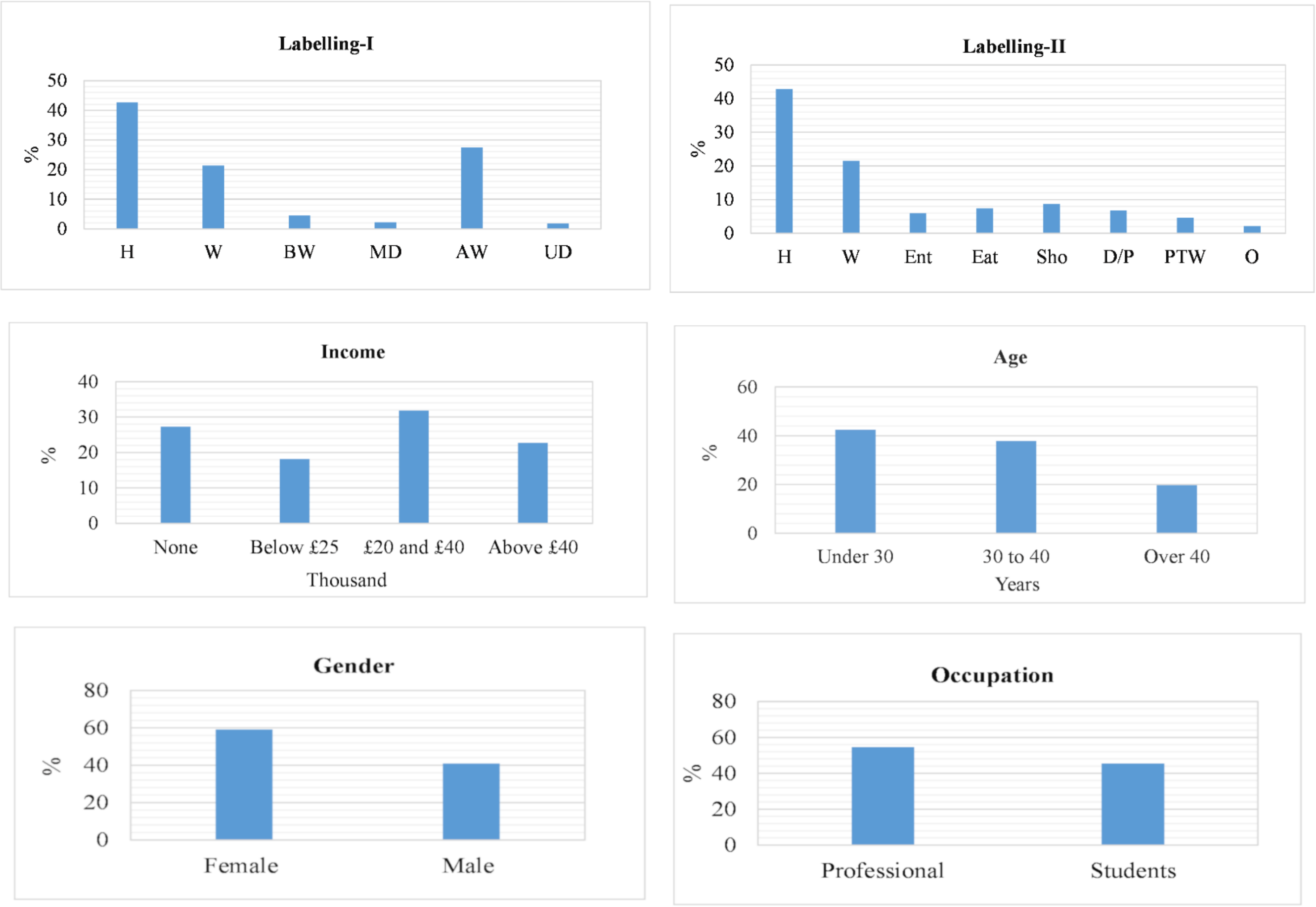


Figure 4.2 Activity points from Survey Smart Card Data (SSCD) and its details /characteristics

income and house ownership/tenure. The second part of the data focuses on individual-level information by asking respondents to list household members over five years of age, including socioeconomic characteristics such as work status, travel information, frequency and mode of public transport usage, public transportation passes and 'travel diaries' (trip sheets or travel logs), which is a minimum of four days of journey data. As an example, the characteristics of an individual's data are presented in Table 4.2, such as home, work and demographic attributes.

4.2.3.1 Matched journeys

The LTDS dataset includes the respondents' Oyster card ID numbers, enabling TfL to match the information in the questionnaire to actual journeys. In other words, travel diaries are as part of the LTDS data for the same individuals from SCD.

The matched journey data include a total of 2,718,644 records for 10,895 unique users from 2011 to 2014. Though the matched journey data cover three years, the recorded journeys are sparse. Therefore, the data used for the case study is a subset, which includes 369,745 journeys for 9479 users from January to February 2014, the period with the highest journey counts. Then, bus journeys are removed from the selected records to arrive at a total tube/train journey count of 124,031.

Matched journeys are pivotal for the validation of the proposed algorithms. In this work, the proposed home and work location algorithms (Section 6.1.2) and benchmark method (Hasan et al., 2012) are applied in matched journeys and identified home and work locations are validated from LTDS data, which is in Table 4.2 with relevant information. However, due to privacy issues, LTDS data are based on the first part of the postcode (outward code) – postcodes are divided into two sections in London– e.g., WC1E 6BT– which are outward code, i.e., an area (WC) and district (1E) and inward code, i.e., a sector (6) and unit (BT). Thus validating station-based matched journeys based on postcode districts is still limited. The detail of the validation part is in Section 6.1.2.4 and 6.1.3.2. Moreover, LTDS data provided by TfL have further limitations in terms of activity types. In other words, there are no secondary activities such as entertainment, eating and shopping available in the dataset. Therefore, carrying out research in secondary activities is impossible from LTDS data used only for validation purposes in Chapter 6.1.

4.2.4 Foursquare POIs

A POI is a geographic location that may be useful or provide interesting information for people, e.g., a restaurant, hotel, or tourist attraction. Such information has been widely used to understand land use dynamics in urban environments (Liu et al., 2015; Rashidi et al., 2017).

In this study, POIs are gathered from Foursquare data, which is a location-based social network (LSBN) data for smartphone users conveying visited places via the check-in service. Check-ins are an activity where people share their places by selecting a particular venue (e.g., a cafe or theatre) from a list of nearby locations indicated by the Foursquare application. Besides, the users are free to add a new venue to the list once they have been to an untried location. The required information for this process is the location name and tagging information from users, including comments surrounding the venue called tips in Foursquare.

Individuals' check-ins or adding a new venue permit Foursquare engineers and data scientists to confirm and alter location information, either active or passive, readings from other sources, such as GPS, Wi-Fi and Bluetooth (Sumedi and Eck, 2022). Thus, the location data evolves into reliable information and is used as ground truth data to validate social media studies (Lai, 2018).

Foursquare POIs were collected in 2018 through the Foursquare Application Programming Interface (API)[4] service. Fifty venues were allowed for each query due to the restriction by Foursquare API; thus, the study area is divided into grids, with 50 venues per grid. If the grid has more than 50 venues, it is split into smaller grids until the number of venues in each is less than 50. The venue name, category and check-ins are attributed to each venue ID (Lai, 2018). The total number of POIs and the number of check-ins are collected as 147,041 and 81,328,352 in London, respectively. The sample data can be seen in Table 4.4.

The class Name (Level 1) has used to classify the location categories into seven activity types: home, work, entertainment, eating, shopping, outdoors and recreation and travel and transport, as shown in Table 4.3. These activity types are used in Section 6.2 and Section 7.1.

[4] https://developer.foursquare.com/docs/resources/categories

Table 4.3 Foursquare POI data as an example

Characteristic	Sample data as an example
Venue ID	4ac518d2f964a5203da720e3
Venue Name	British Museum
Address	Great Russell Street, WC1B 3DG, London, United Kingdom
Lon	-0.126597231
Lat	5151886294
Class ID	4bf58dd8d48988d190941735
Class Name (Level 1)	Art and Entertainment
Class Name (Level 2)	Museums
Class Name (Level 3)	History Museum
The number of check-ins	96,141
The number of users	78,230

Many researchers have used Foursquare POI data to examine/analyse human mobility and activity patterns in urban areas (Rashidi et al., 2017) due to the following premises.

- First, the popularity of each geographic point is valuable information for inferring activities and is provided by the number of check-ins of each POI.
- Second, both check-ins and user counts are available for the geographic points, allowing for the differentiation of popularity information (check-ins) from the influences of POIs (user counts).
- Third, opening and closing hours also offer useful information for presenting the dynamics of the city with a time dimension.

Nevertheless, despite the wide range of positive applications, POIs from foursquare data have a number of limitations in terms of contribution bias, which means a small number of users are responsible for a substantial part of the check-ins, specifically for the eating and shopping activities compared to home or work activities. This creates an over-representation of some locations in cities (Rashidi et al. 2017). Besides, the data also suffer from demographic biases, which means the application is mainly popular for younger users between 15 and 30 compared to older age groups (Longley and Adnan 2016). To overcome such limitations, first, individuals' daily patterns from SCD are considered for spatial and temporal attributes of POIs. For instance, before-work activities matched with relevant POIs exclude irrelevant POIs such as nightclubs, dinner locations etc. Second, Foursquare POIs are used for the secondary activity identifications to infer trip purposes in the centre of cities. Thus, the low number of check-ins in residential areas wasn't the main concern for Section 6.2. Third, the validation or the accuracy of the trip purposes is considered based on the available number of the POIs.

Table 4.4 Activity types from Foursquare data.

Activity Types	Activity Location Type
Home	Residential building (apartment/condo), housing development, house (private)
Work	Government building, library, medical centre, office, parking, post office, radio station, recruiting agency, school, college and university, social club, TV station, warehouse, etc.
Entertainment	Art gallery, pub, nightclub, arcade, theatre, club, bar, concert hall, other nightlife, opera house, casino, event space, dance studio, etc.
Eating	Coffee shop, sandwich bar, cafe, diner, bakery, burger house, restaurant, steakhouse, breakfast bar, taco franchise, bagel shop, etc.
Shopping	Supermarket, corner store, pharmacy, mall, boutique, plaza, miscellaneous shop, farmers market, automotive shop, food and drink shop, bookstore, etc.
Outdoors and recreation	Park, playground, recreation centre, rock climbing venue, ski resort, etc.
Travel and Transport	Hotel, bus stop, tube station, bike rental/bike share, airport, etc.

POIs are also gathered in the UK from other resources. For instance, Ordnance Survey (OS) is one of the data sources which can be downloaded through Digimap[5]. Another source is OpenStreetMap (OSM),[6] which uses volunteered geographic information to create, collect and circulate geographic points representing physical features mapped by volunteers (professionals and residents). Even though both data sources (OS and OSM)

[5] https://digimap.edina.ac.uk/

[6] https://www.openstreetmap.org/

provide reliable location and classification information verified by professionals, the limitations of these data sources are as follows:

- Both data sources treat all geographic points as equal without using a weight factor such as popularity.
- There is no time attribute available in either dataset. Thus, the dynamics of the city can be presented without a time dimension despite the fact that many urban POIs have multiple functions with different opening and closing hours.

These limitations from OS and OSM multiplied across thousands of locations may represent a major source of bias. Thus, the study has focussed on Foursquare POIs when analysing and inferring individual activities.

4.3 Data Challenges[7]

This study aims to infer trip purposes from public transport networks using large data sources to reveal new insight into spatiotemporal features of urban dynamics. The study with individual-level disaggregate activities gathered from SCD, represents daily movement patterns that can support sustainable public transport services, urban infrastructures and policymakers' decision-making processes.

Traditionally, trip purposes have been notified by travel surveys. The information from surveys represents a short temporal episode in dynamic cities. Even though survey data capture demographic attributes, the sampling bias involves a small number of respondents. Thus, there has been an increased effort to understand trip purposes from the many new forms of big data sources, including smart card data and social media.

Using new big data sources, such as smart card data and POIs from Foursquare data, 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 limitations. For instance, SCD may suffer from demographic details of passengers' (Zhang et al., 2020; Zhang, Cheng and Sari Aslam, 2019), recording destination information for bus users (Gordon et al., 2013) and the trip purpose of the travellers, investigated further using

[7] Part of this chapter has been presented in the following publications: N Sari Aslam, MR Ibrahim, T Cheng, H Chen, Y Zhang 2021. "ActivityNET: Neural Networks to Predict Public Transport Trip Purposes from Individual Smart Card Data and POIs." *Geo-spatial Information Science* 24 (4): 711–721.

land use attributes such POIs. Similarly, regardless of the wide range of positive characteristics of POIs from foursquare data, e.g. quantifying the weight of the place using check-ins, using working hours of POIs to present dynamics of the activity patterns in cities, POIs may suffer from over-representing of some of the locations, e.g. a small number of users with substantial check-ins in restaurant or shopping centers as compared to workplaces (Rashidi et al., 2017). In addition, demographic biases in the dataset are inevitable that the application is mainly used by younger age groups, e.g. less than 30 years old, compared to older age groups in the cities (Longley and Adnan, 2016).

Dealing with large data sources with different/mismatch periods is also a challenging task. In this work, available sample data from SCD (Oct and Nov 2013 with 1,823,906 journey records), SSCD (captured from Sep 2017 to Mar 2020 with 19,792 journey records) and LTDS data (Jan and Feb 2014 with 369,745 journey records) covered different periods. Land use data, which consisted of Foursquare POIs (collected from Jan to Apr 2018 with 147,041 POIs) were also in different periods. This may have caused biases and decreased the model accuracy. To overcome such challenges, the data sources used for the analysis were as follows.

i-) SCD (Oct and Nov 2013) and LTDS matched journeys data (Jan and Feb 2014) are used in Section 6.1.

The period of data sources is not an exact match but was close enough for the analysis. Home and work/study locations/activities are viewed for long-term forecasts. They are usually considered ‘mandatory activities’ and do not change drastically for the users (Castiglione, Bradley and Gliebe 2015). Nevertheless, this introduced a level of inaccuracy in matching that adversely impacted the validation accuracy.

ii-) SSCD (captured from Sep 2017 to Mar 2020) and Foursquare POIs (2018) were linked for enrichment purposes in Section 6.2 and for training the ActivityNET model in Section 7.1.

The land use pattern in urban centres like London changes over time. For instance, some businesses close down and new ones open up in an area. In the activity consolidation algorithm, data was integrated at the location level and not at the level of the exact date period. For instance, check-in aggregation was based on the day of the week and time. Even though there was little difference in the period alignment of the two datasets, POIs may not be exactly representative of the activities. However, this presented a level of inaccuracy in matching and impacted the validation accuracy.

iii-) The SCD (Oct and Nov 2013) was unlabelled data and only used for inference/prediction throughout this study (Section 7.2), not for training ActivityNET (Section 7.1) or heuristic model validation (Sections 6.1.3.2 and 6.2.3.2). Because of the significant time period difference between the SSCD, Foursquare POIs and SCD datasets, it was anticipated that the inference/prediction results would not be as accurate as the model validation results. This was one of the limitations of the study due to the constraints of the available data.

Moreover, individuals represent different numbers of activities. Imbalance activity types gathered from any data types represent issues for the modelling urban areas. To overcome this challenge, random under and over-sampling techniques are applied in the dataset in Section 7.2.

Finally, 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 may be disturbed by eating (drinking coffee/tea) at a location where both shopping and eating places are available. Although it is difficult to separate those activities in individuals' daily lives, there are no multiple activities in SSCD for the analysis. Thus, we assume this is not an issue for the proposed methodologies.

4.4 Chapter summary

This chapter describes the data sources such as SCD, SSCD, LTDS and POIs and their characteristics to be used in the following sections. First, Primary location identification, detailed in Section 6.1, considers three types of data sources for London: (1) individual Oyster card data (Section 4.2.1) from London stations using home and work location identifiers, (2) LTDS data for validation of home and work locations (Section 4.2.3) and (3) SSCD (Section 4.2.2) from volunteers, also used for validation. Second, for the secondary activity identification algorithm (Chapter 6.2), the study focuses on SCD (Section 4.2.1), enriched using Foursquare POIs (Section 4.2.4). The performance of the algorithm was validated using SSCD (Section 4.2.2) in the London case study. Third, trip purpose prediction (Chapter 7) using the ML algorithm focused on three types of data: Oyster card data (SCD) (Section 4.2.1), SSCD (Section 4.2.2) and Foursquare POIs (Section 4.2.4). Finally, data challenges are discussed with the pros and cons of the data sources for the research before summarising the chapter in Section 4.4.

Chapter 5

The Characteristics of Extracted Activities

5 THE CHARACTERISTICS OF EXTRACTED ACTIVITIES

This chapter describes activity extraction and its characteristics from transit data and combines activities with land use data (Point of interests (POIs)). Section 5.1 illustrates the data pre-processing steps and assumptions used for the three transit data sources – Smart Card Data (SCD), Survey Smart Card Data (SSCD) and London Travel Demand Survey (LTDS) (matched journeys). Section 5.2 presents the extracted activities with the definitions of characteristics from SCD, including activity start/end times and stations, visit-frequency, stay-time (activity duration) and direction (from and to). Section 5.3 explains the process of combining travel data (SCD and SSCD) with auxiliary land use information using an activity–POIs consolidation algorithm to select relevant POIs for each activity. The chapter is summarised in Section 5.4.

5.1 Data pre-processing steps/assumptions

A *trip* is defined as a one-way journey from one station (origin) to another station (destination) using the public transport network in an individual's daily travel. An *activity* is defined as a period between two consecutive trips, including the start and end stations (locations). Trip purposes, on the other hand, is derived information from extracted activities provided either data mining or processing steps using the characteristics of SCD or additional information, i.e. land use attributes in the proximity of start/end stations (Faroqi, Mesbah and Kim, 2018). Thus, before extracting any information from travel data such as SCD, SSCD and LTDS (matched journeys), cleaning and data pre-processing steps with the following assumptions are performed for each individual. First, each individual has a unique ID stored automatically by the public transport system (Bouman, Kroon and Vervest, 2013; Nassir, Hickman and Ma, 2015; Assemi et al., 2020). Second, trips with missing key data attributes, such as alighting time and stations, which represent mainly trips by bus or tram, are excluded. The reason is that the incomplete recording of trip information may create extra ambiguity in inferring trip purposes in a complex urban system. However, such journeys can be included with an enhancement of the model that estimates missing information as a sub-step rule (Gordon et al., 2013). Third, each individual per day must have a minimum of two consecutive trips to be extracted as an activity. Therefore, single trips (per day) have been considered a relevant cause of failure for further analysis and are excluded (Trépanier and Chapleau, 2006; Munizaga and Palma, 2012; Gordon et al., 2013; Ma et al., 2013; Munizaga et al., 2014; Nunes, Dias and Cunha, 2015; Alsger, 2016; Jung and Sohn, 2017; Hora et al., 2017; Kumar, Khani and He, 2018). The rest of the journeys for the same individuals are still considered for further analysis. However, alternative destination stops for single trips can be investigated individuals' travel regularity from their mobility patterns with additional assumptions (Cong, Gao and Juan, 2019; Lei et al., 2021). Fourth, extracted activities are checked based on the accepted transfer time threshold (Alsger, 2016). Details of transfer times are discussed in Section 5.1.1. Fifth, extracted activities are also checked based on the stations and walking distance. The reason is that if the end station and start station of the next trip are not the same, the activity cannot be identified. However, a different start station may be considered if within a defined walking distance from the end station. Details are given in Section 5.1.2 (Nassir et al., 2012; Gordon, 2012; Alsger, 2016).

Oyster cards with unique IDs are used on multiple modes of transportation across London. In this study, the Oyster dataset includes almost 43 per cent tube and train journeys, 54 per cent bus journeys and 1 per cent tram journeys. These percentages are slightly different in the matched journeys dataset: 33 per cent tube and train, 66 per cent bus and 1 per cent tram journeys. The survey data contain only 1 per cent bus and tram journeys due to the study being focused specifically on tube and train journeys. In addition, the impact of the single trips is based on the total number of journeys decreased almost 6 per cent, 4 per cent and 5 per cent for the SCD, LTDS and SSCD datasets, respectively.

This research mainly focussed on OD's to identify trip purposes from extracted activities, e.g., tube and train. Even though sensitivity analysis in trip-chaining assumptions is investigated in literature to estimate destination stations, e.g., for bus journeys (Trépanier et al., 2007; Chu and Chapleau, 2010; Seaborn, Attanucci and Wilson, 2009b; Wang, 2010; Munizaga and Palma, 2012; Gordon, 2012; He et al., 2015; Alsger, 2016; Hora et al., 2017; Cong, Gao and Juan, 2019; Lei et al., 2021), this idea is applied for the improvement of the activity identification in this study under two scenarios. The first scenario examines consecutive trips where the alighting station of the trip is the same as the boarding station of the next trip ($S_{Origin} = S_{Destination}$). The second scenario relaxes the station condition of the consecutive trip selection to consider the potential walking distance between two locations even if the origin and destination stations are not the same ($S_{Origin} \neq S_{Destination}$). Thus, transfer time threshold and walking distance are further investigated: i) from the literature, i.e., Sections 5.1.1 and 5.1.2 and ii) from the preliminary analysis, i.e., Section 5.1.3.

5.1.1 Transfer time threshold

Transfer time is defined as the duration between two consecutive journeys to transfer from one trip to another. Therefore, one of the challenges in differentiating activities between two consecutive trips is deciding whether a particular duration is an activity or transfer time (Nassir, Hickman and Ma, 2015; Alsger, 2016). The question has been investigated using different transfer time thresholds throughout the literature. If the duration between alighting and boarding stations is less than the presumed time threshold, the duration is labelled as a transfer. In contrast, if the duration between alighting and boarding stations is more than the presumed time threshold, then the duration is labelled as an activity. However, the challenge remains that activities occurring in a short time frame, such as buying a coffee or visiting a bank, may be mislabelled as transfers (Nassir et al., 2015).

Acceptable transfer time thresholds have been determined to fall between 20 min and 180 min depending on the combination of transportation modes, such as bus-to-bus or bus-to-underground, in different cities (Bagchi, Gleave and White, 2003; Nassir, Hickman and Ma, 2015; Gordon et al., 2012; Ma et al., 2013; Alsger et al., 2018).

For London, potential interchange time thresholds are 20 min, 40 min and 50 min for tube-to-bus, bus-to-tube and bus-to-bus interchanges, respectively (Seaborn, Attanucci and Wilson, 2009a). The time difference is double between tube-to-bus and bus-to-tube transfers because bus-to-tube transfers include the bus travel time (from bus boarding to tube boarding due to lack of bus alighting time and station data), whereas tube-to-bus transfers are identified from tube alighting to bus boarding without the bus travel time. Gordon (2012) and Zhang et al. (2020), on the other hand, determined the transfer time for London to be 45 min because they included bus trips. A transfer time of 20 min was chosen for this study, the same as that of tube-to-bus interchanges, which do not include the durations of bus trips, as proposed by Seaborn, Attanucci and Wilson (2009). Further, TfL (2019) estimated the average transfer time for the majority of central London stations to be approximately 20 min. Further analyses are carried out about transfer time in Section 5.1.3.

Table 5.1 Transfer time threshold for different cities

Literature	Transfer Time Threshold	City
Bagchi and White (2005)	30 min	London, UK
Hofmann and Mahony (2005)	90 min bus to bus	Not given
Seaborn, Attanucci and Wilson (2009)	20 min tube to bus 40 min bus to tube 50 min bus to bus	London, UK
Nassir et al. (2011)	30 to 90 min bus to bus	Minneapolis, Minnesota, USA
Gordon (2012)	45 min, including bus	London, UK
Ma et al. (2013)	60 min bus to subway	Beijing, China
Alsger (2016)	90 to 180 min	Brisbane, Australia

5.1.2 Walking distance

Many studies have considered that if the alighting activity station is the same as the station of the boarding of the next activity, there is no need for walking distance to be factored in ($S_{Origin} = S_{Destination}$). If the alighting activity station is not the same as the station of the boarding of the next activity ($S_{Origin} \neq S_{Destination}$) then walking distance is checked based on an assumed threshold in this study.

Maximum acceptable walking distances within different cities are presented in Table 5.2. Although walking distance might be affected by different factors such as the functionality of the transport network, type of city, weather, or type of person, a maximum distance is regarded to be 400 to 1100 m in various combinations of trip interchange, such as bus-to-bus or bus-to-tube, in different urban environments (Cui, 2006; Seaborn, Attanucci and Wilson, 2009; Munizaga and Palma, 2012; Gordon, 2012; He et al., 2015; Chaniotakis et al., 2016; Zhao et al., 2017; Alsger et al., 2018).

According to Gordon (2012), London's walking distance is 750 m, which captures over 94 per cent of the activities in their dataset. Wang (2010) mentioned walking distance in London is 1000 m. In addition, according to TfL (2014) and RTPI (2018), 800 m is considered the acceptable walking distance between two stations for Londoners. Further analysis are investigated for different walking distances in Section 5.1.3.

Table 5.2 Walking distance thresholds for different cities

Literature	Walking distance	City
Cui (2006)	1,110 m for bus stop	Chicago, Illinois, USA
Trépanier et al. (2006)	2000 m	Ottawa, Canada
Zhao et al. (2007)	400 m	Chicago, Illinois, USA
Wang (2010)	1000 m	London, UK
Nassir et al. (2011)	800 m	Minneapolis, Minnesota, USA
Munizaga and Palma (2012)	1000 m	Santiago, Chile
Gordon (2013)	750 m	London, UK
He et al. (2015)	400 m	Brisbane, Australia
Chaniotakis et al. (2016)	640 m	Porto, Portugal
Alsger (2016)	400, 800, 1000 m	Brisbane, Australia

5.1.3 Preliminary analysis based on walking distance and transfer time

This section aims to investigate walking distance and transfer time threshold to improve activity identification. Table 5.3 illustrates different walking distances: i) origin and destinations stations for an activity are the same ($S_{Origin} = S_{Destination}$), shown with 0 meters and ii) an activity origin and destinations stations are not the same stations ($S_{Origin} \neq S_{Destination}$), investigated based on 200, 400, 600, 800, 1000, 1200, 1400 meters using Euclidean distance (Gordon, 2012). Besides, activity times are presented as 5 min, 10 min buckets and increase cumulatively. That means 10 min activity time bucket also covers 5 min activities. For instance, 800 m walking distance, activity identification is improved from 327399 to 346307 without transfer time threshold.

However, less than 20 min activities in 800 m walking distance are 10917, considered as transfer activities and may exclude for further analysis.

In Table 5.3, none in activity times represents identified activities without a transfer time threshold. It can be seen that increasing walking distance helps to identify more activities. However, the change in terms of the count of activity identification decreases after 800 m distance.

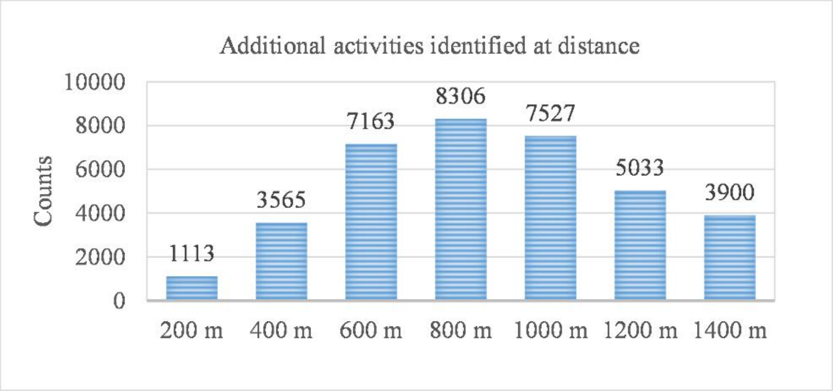
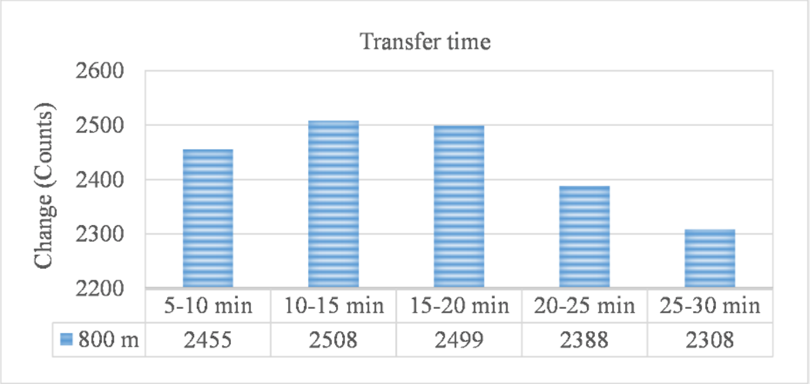


Figure 5.1 Different walking distance thresholds to improve activity identification

Further investigation in 800 m distance is illustrated in Figure 5.2 illustrates for transfer time. It can be seen that there is a close difference between 10-15 min and 15-20 min activities for 800 m distance. Such activities need to be excluded due to their short duration. Otherwise, they may decrease accuracy during the prediction processes unless there are labelled as transfer activities.



	5-10 min	10-15 min	15-20 min	20-25 min	25-30 min
▬ 800 m	2455	2508	2499	2388	2308

Figure 5.2 Different transfer time thresholds to improve activity identification

Table 5.3 The result of the walking distance and the transfer time from SCD. None in activity times represents activities without time conditions.

Activity times	Walking Distance							
	$(S_O = S_D)$.	$(S_O \neq S_D)$.						
	0 meter	200 m	400 m	600 m	800 m	1000 m	1200 m	1400 m
5 min	3377	3380	3420	3486	3495	3503	3520	3530
10 min	5749	5752	5811	5941	5950	6006	6039	6079
15 min	8187	8190	8266	8443	8458	8609	8654	8694
20 min	10478	10483	10575	10811	10917	11061	11142	11242
25 min	12695	12701	12837	13133	13345	13484	13585	13595
30 min	14835	14841	15005	15371	15653	15837	15955	16105
>30 min	312247	312353	315743	322544	330549	337879	342783	342793
None	326160	327273	330838	338001	346307	353834	358867	362767
Additional activities identified at a distance		1113	3565	7163	8306	7527	5033	3900

The results are based on the available data and may need to require new analysis to see/plot the result of walking distance and transfer time for the further analysis, such as including bus journeys or using a new sample dataset.

5.2 Extracting activities

Trips and activities are both used in the literature to investigate travel behaviour, mobility patterns and traffic volumes in urban settings (Ben-akiva, Bowman and Gopinath, 1996). However, this study mainly focuses on activities from large smart card datasets to investigate the rationale for individual movements, except for home activities from heuristic approach. The reason for focusing on activities rather than trips is threefold. First, travel is derived from demand and the reason for making a trip (the why) is the time spent on an activity rather than simply for its own consumption value in the transport network (Zhang and Levinson, 2004). Second, activities are the basis for behaviour at a certain place at a given time. It is, therefore, crucial to understand the when (start and end time of the activity), where (the location of the activity) and how (mode of the travel) of each activity. Finally, an individual’s activities connected in a sequence on a given day can reveal travel patterns. Investigating each activity in the travel pattern helps to predict travel choices for subsequent trips, providing valuable insight not only for city and transport planners but also for commercial organisations interested in brand and product placement (Goulet-Langlois, 2016).

Thus, activities are extracted for durations greater than the defined transfer time threshold and when the stations are nearer than the walking distance threshold. After applying transfer time and walking distance conditions, there is an increase of approximately 4 per cent of activity extraction from SCD, 2.4 per cent of activity extraction from matched journeys data and 1.8 per cent of activity extraction from survey Oyster card data.

5.2.1 Start and end stations

Another important indicator for this study is an individual’s start and end activity stations in a given day. The assumption proposed by Barry et al. (2002), is that most people start the first journey of the day at the station where they end their last trip of the day. Many researchers later improved this idea by using new assumptions to infer ODs from SCD (Trepanier et al., 2007; Seaborn, Attanucci and Wilson, 2009). In this study, a similar idea is adopted from Barry et al. (2002) and other researchers (Chakirov and Erath, 2012; Hasan and Ukkusuri, 2014; Li et al., 2015; Zou et al., 2016) using the characteristics of SCD to identify home locations for individuals, further detailed in Section 6.1.2.

5.2.2 Travel frequency

Commuters' regular travel patterns are an important component of travel behaviour and pattern analysis. Quantifying travel regularity for tube/train passengers daily, weekly, monthly, or even yearly is valuable information for transport planners and has been considered for OD estimation and travel behavioural pattern analysis (Zou et al., 2016; Hasan et al., 2012; Li et al., 2015; Alsger et al., 2018) and transport network demand analysis (Wang et al., 2017).

In this study, the frequency parameter is considered to be an indicator of the regularity of the activity. The level of regularity is measured for each user based on the question, 'how many times has the user visited this location?' The frequency parameter is examined based on the available data period and the acceptable threshold is selected to define home and work locations/activities as described in Sections 6.1.2.

5.2.3 Stay-time duration/activity duration

One of the essential characteristics from extracted activities is stay-time duration, defined as the activity duration between two consecutive trips. The time spent in a specific location is different based on trip purposes, such as for home, work, eating, or shopping. Using stay-time duration with other characteristics – such as activity start and end times or the type of facilities at the location – provides valuable insights for further analysis.

Stay-time duration is used in several places in this study. The first is in heuristic PA identification, specifically for work locations/activities as mentioned in Section 6.1.2. Next, the temporal variation of SAs is investigated based on stay-time duration (details are in Section 6.2.2). Lastly, it is used in the ActivityNET model as one of the input variables (details are in Section 7.2.1).

5.2.4 Direction (from/to) information

The locations from/to an activity are defined as the nearest spatial information relating to an activity. The from-activity location (FL) is defined as the last location before an activity and the to-activity location (TL) is defined as the next location after an activity. Hence, the activity is marked based on the position of PAs such as home and work. This is used for the identification of the SAs as described in Section 6.2.

5.2.5 Activity start and end stations and times

The start and end stations (locations) of activities are used as spatial attributes of the activities and the start and end hours of activities are used as their temporal attributes. These are important from a transport planning perspective to predict daily, weekly, monthly and yearly travel demand for travel networks within cities. They are also important indicators from a land use policy perspective, for optimising the use of city centres to prevent congestion or the desertion of areas at certain times. Activity starting and ending hours help to guide establishments' opening and closing times (Montgomery, 2017). Furthermore, activity start/end times and locations are used in identifying home and work locations/activities (Sections 6.1.2), enrichment of SAs (Section 6.2.2) and in the ActivityNET framework as input data (Section 7.2).

5.3 Activity–POIs consolidation algorithm[8]

Each trip is made for a purpose such as work, entertainment, eating, shopping, drop-offs/pick-ups, or part-time work. The purpose of the trip can be inferred from activities, but the location information from SCD is limited based on tap-in/out stations. To enrich the location information from SCD, land use information needs to be combined with the travel dataset, which provides more information about the location of the stations.

POIs from 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 (Rashidi et al., 2017). The dataset includes a broad classification of location and activity

[8] Part of this chapter has been presented in:
[1] N Sari Aslam, D Zhu, T Cheng, MR Ibrahim, Y Zhang 2020. Semantic enrichment of secondary activities using SCD and points of interests: a case study in London. Annals of GIS, 1-13.
[2] N Sari Aslam, MR Ibrahim, T Cheng, H Chen, Y Zhang 2021. "ActivityNET: Neural Networks to Predict Public Transport Trip Purposes from Individual Smart Card Data and POIs." Geo-spatial Information Science 24 (4): 711–721.

types in seven categories (home, work, entertainment, eating, shopping, outdoors and recreation and travel and transport), as well as the number of check-ins, the user count at each POI, the opening and closing hours of the POIs.

The proposed algorithm aims to determine the land use distribution of activity locations to link the activities from SCD to obtain the potential POIs using time variables. To infer activities from transit data, this section explains how both large datasets – SCD and POIs – are combined and used for the enrichment of SAs (in Section 6.2.3.3). In addition, data points through this algorithm (SSCD and POIs) are used as input for ANN in Section 7.2.1 and prediction purposes with (SCD and POIs) in Section 7.2.4.

Figure 5.1 explains the workflow in this section: First, the proposed activity–POIs consolidation algorithm filters relevant POIs for each activity (A). Second, the data characteristics are presented under the three subsections: spatial information match, temporal information match and attractiveness (B). Third, an example, with visualisations, is presented spatially (C), temporally (D) and for aggregated (sum) check-ins for the activity types (E). In addition, m refers to the number of POIs around the station.

Figure 5.1A illustrates the proposed activity–POIs consolidation algorithm to explain how relevant POIs are filtered for each activity. The algorithm starts by selecting a station and an activity in that station. If there are POIs within walking distance of the station, a POI is selected for that activity. Then, the activity–POI temporal information match is tested against two conditions: ‘the start time of the activity ≥ the opening time of POIs’ and ‘the end time of the activity ≤ the closing time of POIs’. If both conditions are met, the number of check-ins is added under the activity types of the POI. Then, the algorithm moves to the next POI for the same activity. Once all possible POIs have been checked, the activity has the total number of check-ins for each of the activity types: home (H), work (W), entertainment (ENT), eating (EAT), shopping (SHO), outdoor and recreational (REC) and travel and transport (TPO). This process is conducted for all activities in each station. Thus, the characteristics of land use information using the check-ins of POIs are assigned to each activity with different weights. In the second part, Figure 5.1B illustrates the same scenario using the data characteristics in three categories, including a spatial information match using the coordinates of both datasets, a temporal information match using the start/end time of activities from SCD and opening/closing hours of POIs and attractiveness of each activity using the total number of check-ins for the activity types from the POIs. Because the opening hours of the POIs may have some variation on different days, the earliest and latest working hours are used for each POI, e.g., if opening/closing hours of a place are 10:00/15:00 from Monday to Friday and 12:00/16:00 on Saturday and Sunday, the opening/closing hours are considered to be 10:00/16:00 for the place. On the other hand, if the opening and closing hours are not available, the solution is to interpolate the missing data points by applying the industry opening and closing hours to the missing POI. For example, if a restaurant is missing opening/closing hour data points, the most frequent values for the activity class are selected and applied to the missing POI for weekdays and weekends before using the earliest and latest working hours for each POI. The last/third column visualises the same scenario using an example. Figure 5.1C starts with the spatial information match for an activity (A1) at a station (Oxford Circus station) using ‘walking distance 800 m’, which captures 3023 POIs for A1. The same example, further investigated for A1 considering the temporal information match, is displayed in Figure 5.1D. The start/end times of A1 are 10:00/13:00 and the opening/closing hours of the first POI (POI1Sho) are 9:00/22:00. According to the temporal information match, the time variables overlapped; thus, POI1Sho is moved next step and the number of check-ins is saved for corresponding activity types (POISHOs) in Figure 5.1E. Then the next POIs (POI2Wor and POI3Eat) are similarly checked based on temporal information. The number of check-ins for POI2Wor is added in POIWORs, but the number of check-ins for POI3Eat is not counted in POIEATs due to non-overlapping temporal information. After running this process for each of the 3023 POIs, the aggregated check-ins are saved under the seven categories for A1 as the characteristics of land use information, as shown in Figure 5.1E.

5.4 Chapter summary

This chapter explained how activities and their characteristics were extracted from transit data, including combining activities with land use data (POIs). Hence, the extracted characteristics were used in methodological frameworks proposed in Chapters 6 and 7. First, Section 5.1 introduced the key assumptions of the data pre-processing steps with the relevant literature before extracting activities. Then, Section 5.2 explained the definition of activities and characteristics used in the proposed methodologies.

Finally, Section 5.3 detailed how activities from travel data (SCD and SSCD) were combined with POIs from Foursquare data using an activity–POIs consolidation algorithm in Section 5.3.

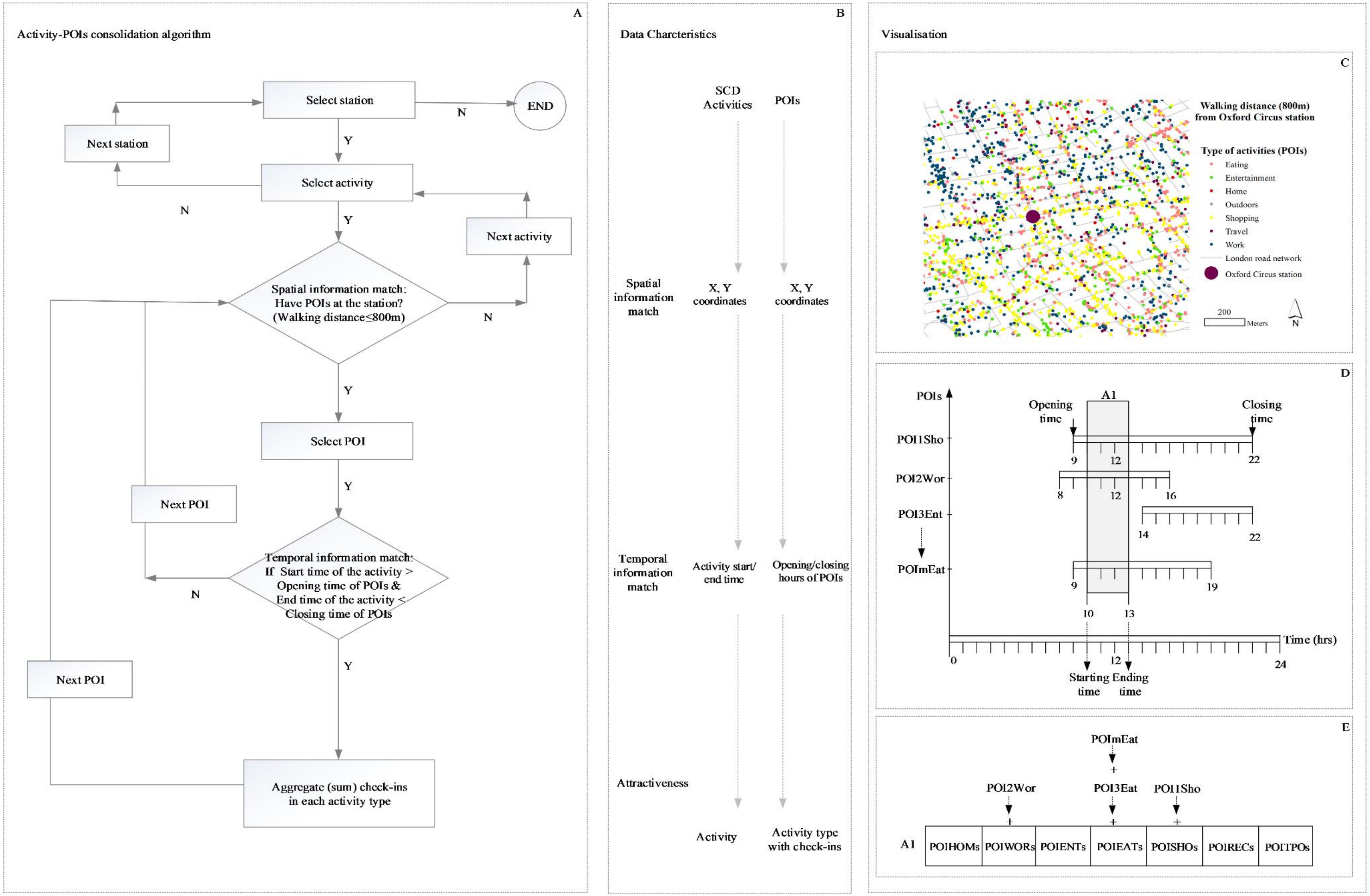


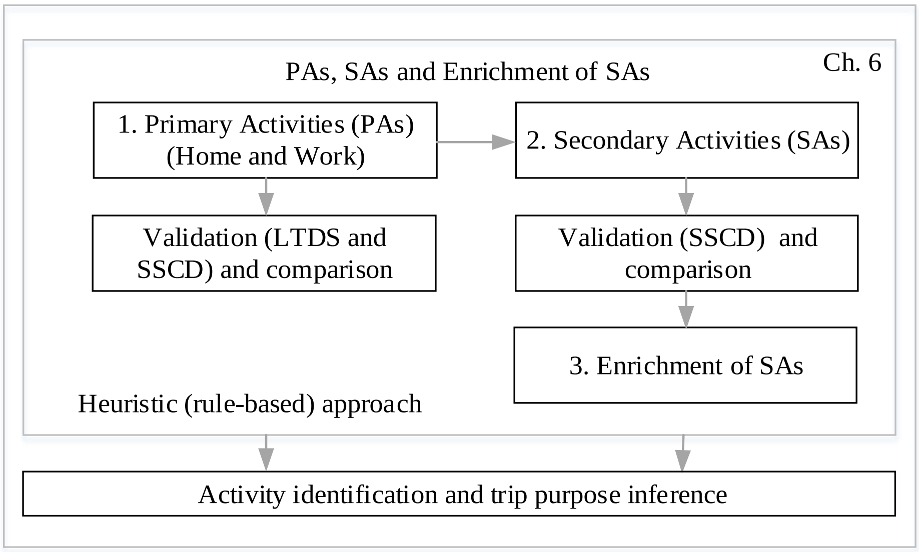
Figure 5.3 A workflow of the proposed Activity-POIs algorithms using SCD and POIs.

Chapter 6

Trip Purpose Inference with Heuristic Approach

6 TRIP PURPOSE INFERENCE USING A HEURISTIC METHOD[9]

This chapter presents activity identification models to investigate trip purposes using a heuristic approach in two parts. The first focuses on the primary location/activity identification model to define Primary Activities (PAs), with the methodology detailed in Section 6.1.2 and a case study with the results of the algorithms and validations using London Travel Demand Data (LTDS) and Survey Smart Card Data (SSCD) in Section 6.1.3. The PAs model is summarised in Section 6.1.4.



Ch. 6
PAs, SAs and Enrichment of SAs
1. Primary Activities (PAs) (Home and Work)
2. Secondary Activities (SAs)
Validation (LTDS and SSCD) and comparison
Validation (SSCD) and comparison
3. Enrichment of SAs
Heuristic (rule-based) approach
Activity identification and trip purpose inference

Figure 6.1 The workflow of PAs and SAs models using a heuristic approach.

The second part presents the Secondary Activities (SAs) identification model and enrichment of SAs. Section 6.2.2 details the methodology and Section 6.2.3 presents the results and validation of the model, including station-based trip purposes using spatiotemporal characteristics of Point-of-Interests (POIs). Lastly, the SAs models and the chapter are summarised in Sections 6.2.4 and 6.3, respectively.

[9] Part of this chapter has been presented in:
[1] N Sari Aslam, T Cheng, J Cheshire 2019. A high-precision heuristic model to detect home and work locations from SCD. *Geo-spatial Information Science* 22 (1), 1-11.
[2] N Sari Aslam, D Zhu, T Cheng, MR Ibrahim, Y Zhang 2020. Semantic enrichment of secondary activities using SCD and points of interests: a case study in London. *Annals of GIS*, 1-13.

6.1 Primary activities identification (PAs)

6.1.1 Introduction

The purpose of this section is to initially address the first research objective, which is to develop a model to identify primary locations and activities once the detailed surveys are not available. Thus, the heuristic PAs identification is proposed using only the start and end stations, visit-frequency and stay-time duration from SCD in a set of rules/algorithms.

This section proceeds as follows. First, primary and secondary location and activity identification models are proposed using what-if scenarios without using surveys. Second, the proposed models are evaluated using the LTDS and SSCD datasets and compared to benchmark models in a large urban environment. After validation, the algorithms are applied to SCD as a case study to represent home and work locations and activities in a large city, i.e., London.

6.1.2 Methodology

6.1.2.1 Identify regular commuters

Regular commuters use the transport network for their daily home- and work-related travel. To establish the regularity of usage, the journey count has been defined as the number of trips carried out by each user in the dataset after applying the data pre-processing steps outlined in Section 5.1. This separates regular users from sporadic users of the network. Too low a threshold will include a large number of irregular users in the dataset, whilst too high a threshold will be too restrictive for the analysis (Hasan et al., 2012). An appropriate journey count threshold will create a meaningful dataset for the study.

6.1.2.2 Home location/activity identifier

The algorithm selects the origin station of the first journey and the destination station of the last journey of the day for each user. If the start and end stations match or are within walking distance (≤ 800 m), the selected stations are analysed further. The next stage passes these selected stations through the criteria of the visit-frequency threshold. If a station is identified more than the defined threshold for visit-frequency, it is classified as a home location for that user.

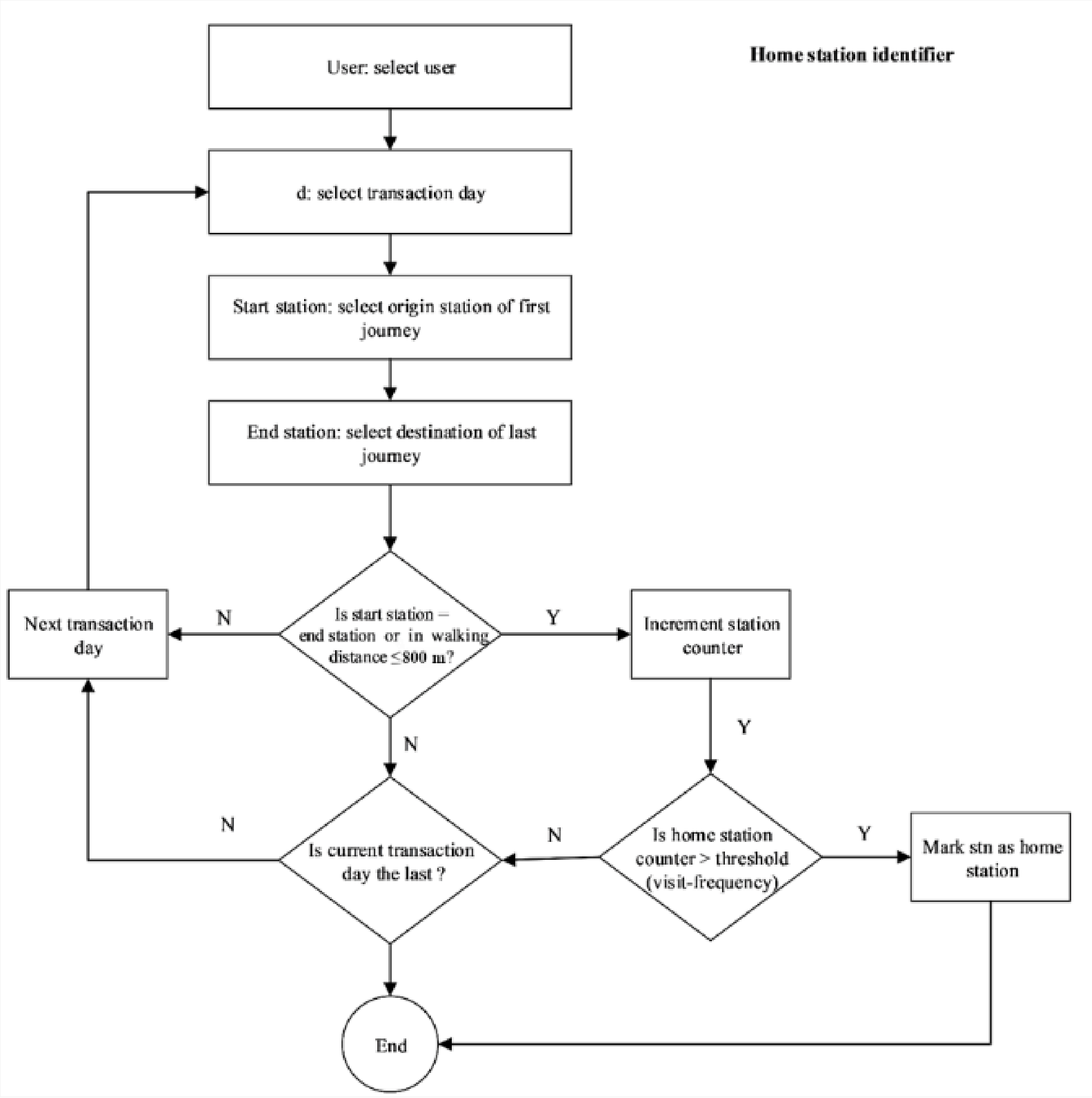

Home station identifier
User: select user
d: select transaction day
Start station: select origin station of first journey
End station: select destination of last journey
Is start station – end station or in walking distance ≤800 m?
N
Y
Next transaction day
Increment station counter
Y
N
Is current transaction day the last ?
N
N
Is home station counter > threshold (visit-frequency)
Y
Mark stn as home station
End

Figure 6.2 Flowchart of home location/activity identification.

The algorithm may fail to highlight any station as a home location if none meet the expected criteria (Figure 6.2). At the same time, it is also possible that the algorithm may find more than one station that fits the criteria for a user’s home location. In that case, Long, Zhang and Cui (2012) and Wang et al. (2017) assigned the home station based on residential areas by using land use information, e.g. POIs. However, this does not apply well to cities with extensive transportation networks, as multiple stations identified may fall within residential areas. Moreover, large cities are often not segregated by land use; residential areas frequently contain work locations and other land use types. Therefore, an approach based on distance and rank using frequency attributes provides a more meaningful outcome than dependence on only land use data.

The algorithm uses the assumption first made by Barry et al. (2002) that a high percentage of commuters end the last journey of their day at the same station where they started the first journey of their day. This location is significant as it represents the home location of the user. Although Zou et al. (2016) made a similar assumption, this paper combines the frequency threshold (Hasan et al., 2012) with a distance threshold (walking distance, ≤ 800 m) to increase the accuracy of the findings.

6.1.2.3 Work location/activity identifier

To identify work locations and activities, all consecutive journey/trip pairs (J1 and J2) for all working days are evaluated. In the model, the destination station of the first journey (J1.destination) and the origin station of the second journey (J2.origin) in the journey pairs are selected. If selected stations match or are within walking distance (≤ 800 m), the stay-time is extracted using the origin time of the second journey (J2.origin time) and destination time of the first journey (J1.destination time). The results selected are those which pass a predefined threshold for stay-time. The next stage is to pass these selected stations through the criteria of visit-frequency threshold. If a station is identified more than the defined threshold for visit-frequency, it is classified as a work location for that user. Based on this criterion, users can have one or more work locations. Similarly, it is also possible that the algorithm fails to highlight any station as a work location if no location meets the expected criteria (Figure 6.3).

In the situation where more than one station appears as a work station, Alexander et al. (2015) and Wang et al. (2017) applied criteria based on the visit-frequency and distance from the home location to identify the work location. The limitation of such an approach is that it fails to take into account the duration of the work activity, leading to the identification of multiple candidates for work locations. Additionally, limiting the work location to a single station can be inaccurate as an individual may have more than one station close to their work. Therefore, an approach based on distance (walking distance, ≤ 800 m), presented by rank, using frequency attributes and stay-time duration is used to provide a meaningful outcome.

The above work location identification algorithm is based on the stay-time and visit-frequency of consecutive journeys. A stay-time criterion is used based on the pragmatic assumptions made by Devillaine, Munizaga and Trépanier (2012) to identify work locations by using activity duration and visit-frequency is used as an indicator in a rule-based algorithm for the identification of work locations based on Hasan et al. (2012). The combination of both indicators along with the walking distance criterion forms the basis of work location identification used in this study.

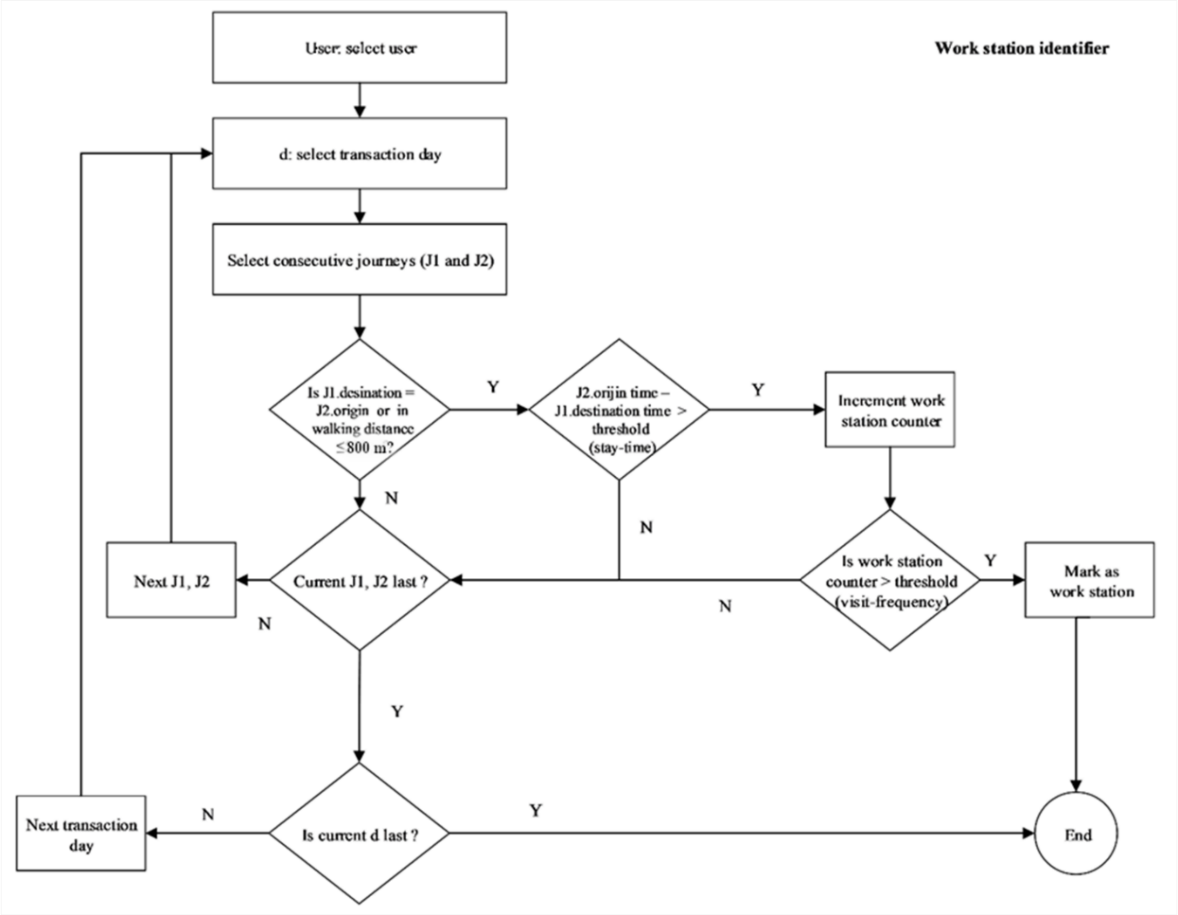

User: select user
Work station identifier
d: select transaction day
Select consecutive journeys (J1 and J2)
Is J1.desination = J2.origin or in walking distance ≤800 m?
Y
J2.orijin time – J1.destination time > threshold (stay-time)
Y
Increment work station counter
N
N
Next J1, J2
Current J1, J2 last ?
N
Is work station counter > threshold (visit-frequency)
Y
Mark as work station
N
Y
Next transaction day
N
Is current d last ?
Y
End

Figure 6.3 Flowchart of the work location identification.

6.1.2.4 Validation process

In this section, three validation methods have been applied in this work to gauge the accuracy of the results. The first method uses the LTDS data, which provides individuals’ home and work locations. After running home and work locations algorithms on LTDS data, correctly identified user locations are further analysed due to the imbalance number of stations in the postcode districts. The location-weighted average (LWA) is calculated using the correctly identified user locations as a percentage of total user locations in the postcode district and the total number of correct locations in the dataset as described in Equation 6.1:

$$LWA = [\left(\frac{VR}{TVR}\right) * \left(\frac{TU_{in\ postcode\ district}}{TU_{in\ London}}\right)] * 100 \tag{6.1}$$

where VR is the validated results, TVR is the total validated results and TU is the total number of users in the LTDS dataset. Hence, two results are presented in this section, i.e., the percentage of the correctly identified home and work locations from LTDS and the visualisation of the correctly identified home and work locations based on postcode district (Figure 6.9).

The second validation approach compares the results to another model, described by Hasan et al. (2012). The model uses the distribution of people's most visited places as the key driver to classify work and home locations in London. Then a subset of user data with locations identified is compared to the proposed models using LTDS in this study.

Finally, the LTDS dataset available for the evaluation is based on survey data collected in a single day in a typical year. In a dynamic city like London, where a large number of people move in and out of the city daily and where people are free to change their places of work and residence, this makes the LTDS data inaccurate for evaluation. Therefore, SSCD are also used for further validation as a recent survey, which is the last validation approach in this section.

6.1.3 Case study

The case study aims to apply the proposed models to SCD. The Oyster card is valid on all London public transport systems and the framework proposed in this study is generic to capture intermodal commuting patterns as long as the complete OD information is available in the journey record.

SCD are used in the PAs model to investigate home and work locations and activities. The models are validated using two survey datasets – SSCD and LTDS. The details of SCD, SSCD, LTDS and data pre-processing are outlined in Sections 4.2.1, 4.2.2, 4.2.3 and 5.1, respectively.

6.1.3.1 Results of the case study

Regular commuters use the network for their daily activities that involve travel to and from their primary locations. After the pre-processing of the dataset, the regularity of usage for individual users is examined through journey counts based on 60 days of SCD, including weekends. To create a meaningful dataset, it is necessary to make assumptions to define an acceptable journey count threshold. Since a minimum of 2 journeys is required to carry out an activity, a minimum of 2 and a maximum of 60 journey counts are considered to examine the regularity of usage.

Figure 6.4 highlights that there is a fall in the number of regular users above a journey count of ten, whereas, between the values of two and ten, the reduction in the number of regular commuters is approximately 52 per cent. The reason for the irregularity based on journey count might be due to the exclusion of bus journeys as described in Section 5.1. Therefore, as a threshold, journey counts of greater than ten are considered to define the regularity of usage. In other words, individuals who have greater than ten journeys in two months dataset are considered for further analysis.

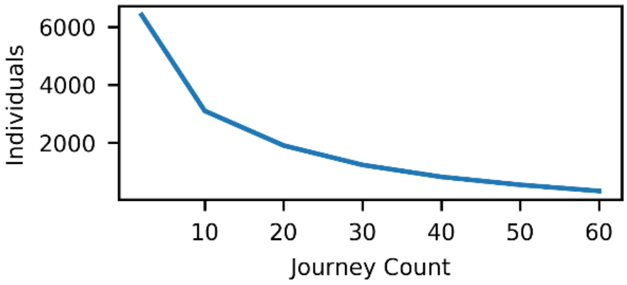


Figure 6.4 Histogram of journey counts during two months of SCD by number of unique users.

To reveal the temporal regularity of people’s mobility, two visit-frequency values, i.e., home (Figure 6.2) and work (Figure 6.3) locations per individual, are investigated to understand how many times defined activity happened in a week per individual. To attain a level of confidence in the results, the heuristic primary location model is presented against the number of different visit-frequency counts as a measure of mobility patterns.

Different indicator values provide different outcomes for the algorithms based on the available data. Low visit-frequency or high visit-frequency can be used depending on the aim of the work (Amaya, Cruzat and Munizaga, 2018). For instance, Wang et al.(2021) mentioned that no more than two days a week represents low visit-frequency. According to their study, three days and at least four days a week indicate mid and high visit-frequencies, respectively. Low visit-frequency is used for shopping or entertainment activities (Goulet-Langlois, Koutsopoulos and Zhao, 2016). High visit-frequency represents commuters (Huang et al., 2018) and is used for identifying primary location/activities (Wang et al., 2021).

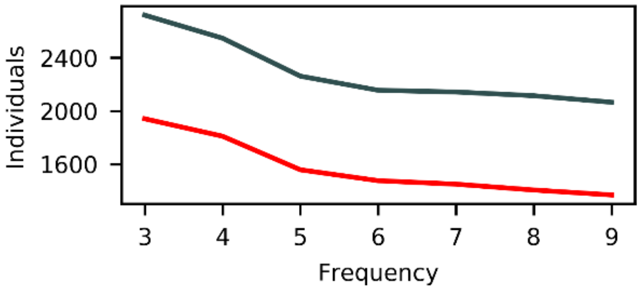


Figure 6.5 The impact of the visit-frequency threshold on the identification of individuals' home (blue line) and work (red line) locations.

After running different visit-frequency thresholds for home (blue line) and work (red line) locations, the number of individuals is presented in Figure 6.5. The expected visit-frequency threshold is considered between three and nine. Because the defined scenario for home locations, i.e., 'individuals' start and end stations are the same or close proximity' might happen one time in a day, which means seven times a week. In addition, a visit-frequency value of five provides a demarcation point for home and work locations. The number of individuals decreases at a slower rate above the value of five as compared to below the value of five. Therefore, a visit-frequency threshold of five is applied in the heuristic primary location model, meaning that the algorithm must correctly identify the home location minimum of five times a week classified as a home location. The same threshold value is applied to the work locations due to a similar reason.

Even though Wang et al. (2021) made an assumption that visit-frequency five captures commuters' behaviour to understand primary locations/activities, this study represents various visit-frequencies not only for home locations but also for work locations. Hence, users have the flexibility to pick different visit-frequency depending on the objective of their studies. Note that picking any frequency works with algorithms. The only difference appears in the accuracy.

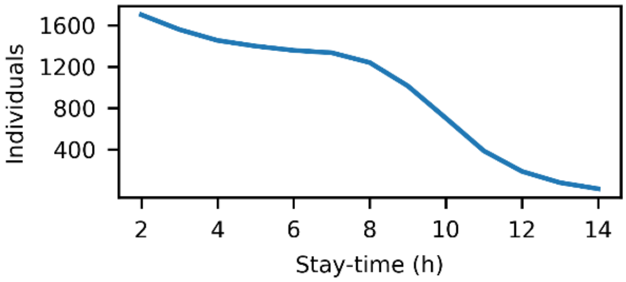


Figure 6.6 Stay-time threshold on the identification of individuals' work locations.

Another important indicator concerning temporal patterns is the stay-time duration, describing activities between two consecutive trips gathered from SCD, which enables the identification of work locations. Figure 6.6 highlights the impact of the stay-time threshold on the identification of work locations for the dataset within a range of two to fourteen hours. As the stay-time threshold increases, the number of users with identified work locations decreases. There is a decline of fourteen per cent for a change in stay-time duration from two to four hours. The drop in the number of the users identified is less significant between four and eight hours compared to the first four hours. It is expected that the activities with durations of eight hours represent the working hours of most regular commuters. Besides, this assumption has validated the information given by ONS (Office of national statistics), which is 'the average working hours a week are approximately 38 hours (eight hours per day)'(Leaker, 2020).

Figure 6.7 demonstrates the well-connected nature of the London transportation network, with home locations dispersed evenly around London. The data on the map are aggregated at the station level. The diameter of each data point represents the number of users that identified a given station as their home location. The top two home stations highlighted from the analysis are Brixton and Stratford. It also shows that the outer boroughs of London were represented by smaller data points in comparison to the inner London stations. This is representative of the high population density in central London.

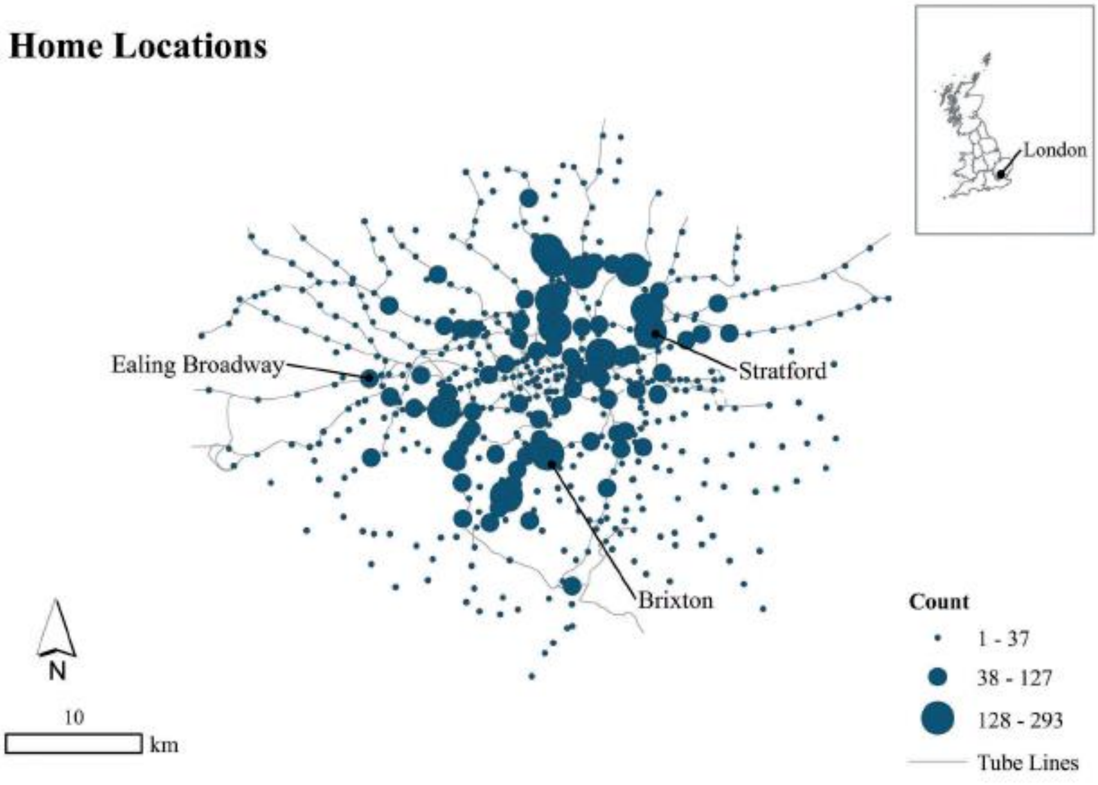


Figure 6.7 The geographic representation of home locations/activities in London.

Figure 6.8 displays the identified work locations, particularly the clustered in the centres of financial services around the City of London. The data on the map are aggregated at the station level. The diameter of each data point represents the number of users that identified a given station as their work location. Locations outside of central London were also identified as work locations, including Ealing Broadway, Stratford and Brixton. These locations are examples of commercial centres outside central London. The contribution of Figures 6.7 and 6.8 is that they highlight the locations – such as Stratford, Brixton and Ealing Broadway – that are significant as both home and work locations. In contrast to Long, Zhang and Cui (2012) and Wang et al. (2017), these figures illustrate that large cities are often not clearly segregated by land use; the residential areas frequently contain work locations and other land use types.

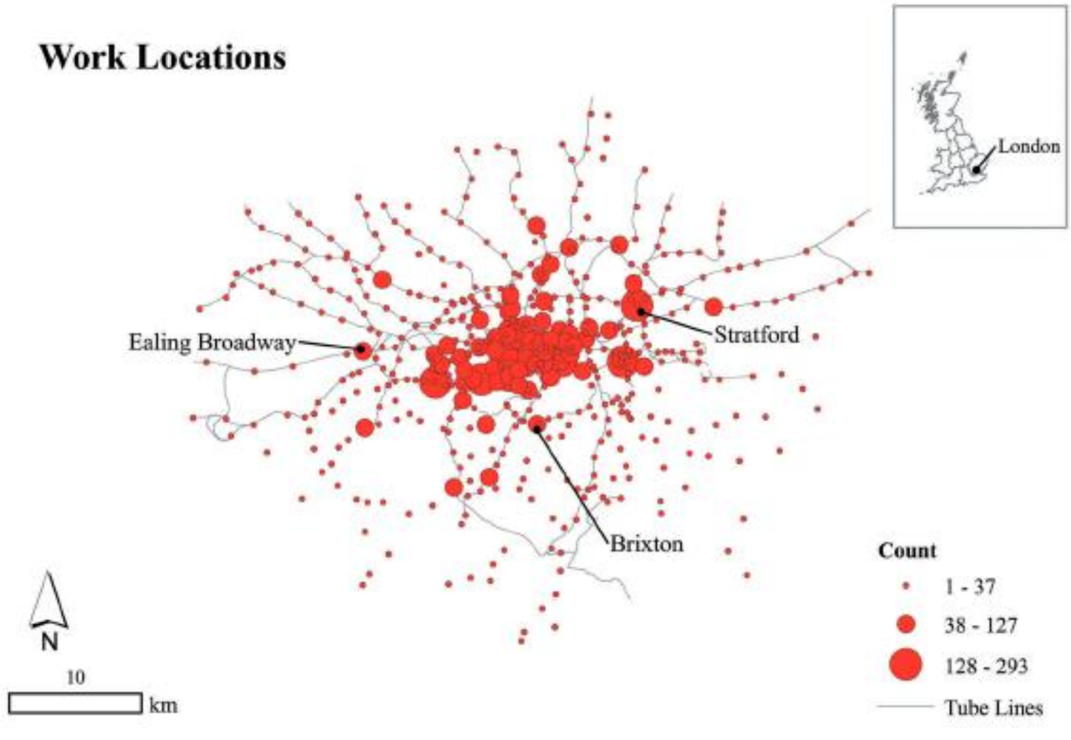


Figure 6.8 The geographic representation of work locations/activities in London.

6.1.3.2 Validating the results of the model

Validation of the analysis is carried out using the LTDS. There are two types of information from LTDS data, i.e., passengers' daily journeys/trips (station-based) and passengers' home and work locations at the postcode district level due to privacy concerns. Therefore, first, station-based journey data are used with proposed algorithms identifying home and work stations. Second, identified station-based home and work stations are matched to the postcode districts of the stations and then compared to the LTDS data. Hence, the results of algorithms are validated from LTDS data and 84 per cent of home and 61 per cent of work locations are correctly identified from LTDS data.

Figure 6.9 presents the validation result of the algorithm for the identification of home and work locations aggregated at the postcode district level. The weighted outcomes were calculated using the correctly identified user locations as a percentage of the total user locations in the postcode district and the total number of correct locations in the dataset. Thus, correctly identified home and work locations results are presented based on the available data as a percentage. For instance, the correctly identified results for

Camden station are 30% and 70 % for home and work locations, respectively. However, note that there are other activities (SAs) in the district area which have not been represented in this figure due to work focused on primary locations. Other activities (SAs) are investigated in Section 6.2.

The second part of the validation demonstrates a comparative analysis of another rule-based approach in London (Hasan et al. 2012) as the benchmark from LTDS data. The existing model results, when validated against the LTDS demographic dataset, corresponding to a success rate of 58 per cent for home locations and 37 per cent for work locations (excluding missing LTDS data). As a result, the heuristic primary location model in this study demonstrates a more accurate identification of home and work locations compared to more simplistic methods.

The proposed model also provides better accuracy than the existing model as a benchmark presented for London, but accuracy for home and work locations in London's highly variable context is difficult due to the city's high rate of residential and employment flux. This is especially the case for the LTDS dataset, which is therefore only an approximate gauge for comparisons. Because such survey data are limited in their ability to provide an accurate source for evaluation, the work has been supplemented with additional verification against a more recent dataset of journeys and locations collected from volunteer users. The third part of the validation, therefore, is to check the proposed algorithm from the new dataset (SSCD). The accuracies of identification of home and work locations from SSCD are 95 per cent and 91 per cent, respectively. The reason for the improvement in accuracy has related the nature of the data, such as longitudinal daily journeys, regular visit-frequency and stay-time duration.

As a result, validating results from LTDS data is a challenging task for this study due to station based journey data and postcode district level of LTDS data. LTDS data are used for both the proposed and existing model (Hasan et al. 2012).

6.1.4 Summary

This study aims to develop a framework to identify primary activities from SCD once the detailed surveys are not available. The model uses journey counts as an indicator of usage regularity and visit-frequency to identify activity locations for regular commuters and stay-time for the classification of work and home locations and activities. London is used as a case study and the model results are validated against data from the LTDS and SSCD. The results demonstrate that the proposed heuristic model can detect accurate home and work locations with high precision at 95 per cent and 91 per cent, respectively, from labelled smart card datasets (SSCD). Thus, a large amount of unlabelled data (SCD) from the Oyster network has the potential to vastly improve the way mobility analysis is carried out in large cities.

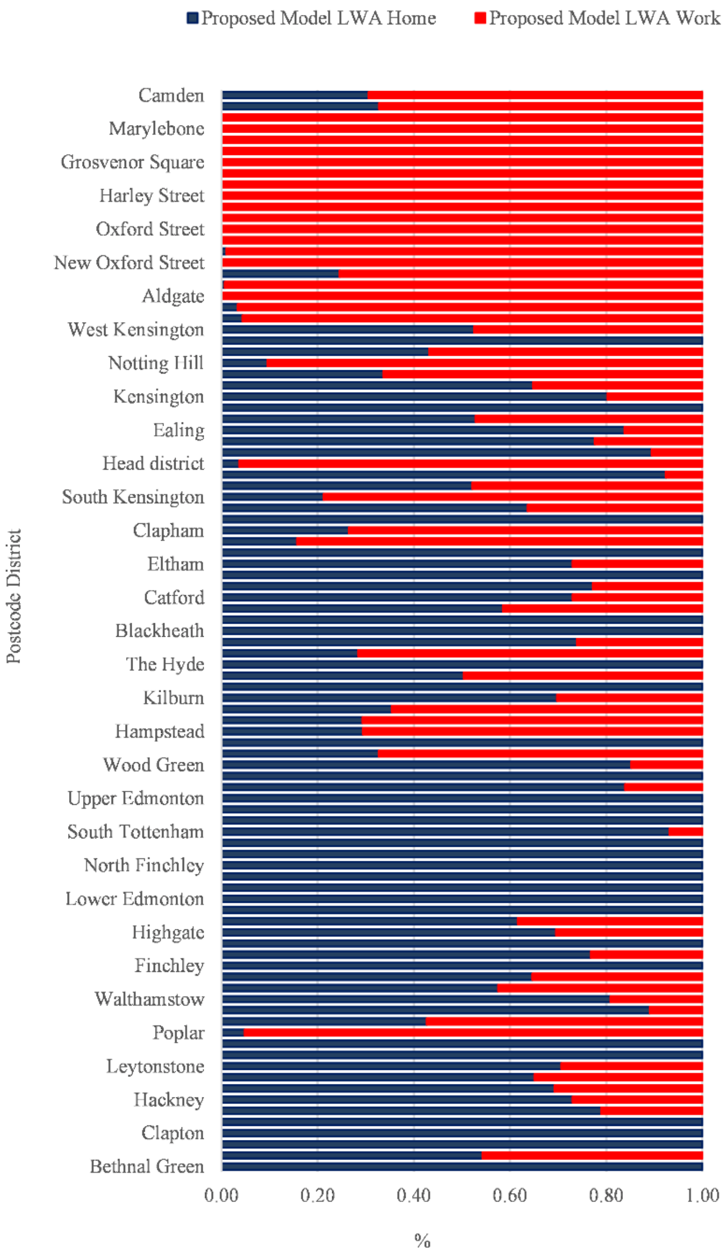


Figure 6.9 Proposed model – LWA home and work location validation results.

This study also shows that surveys from SCD present substantial prospects for the understanding of commuter behaviour and can provide an accurate and more reliable picture compared to user mobility profiles from sample surveys. This approach to human mobility can improve the understanding of wider mobility patterns at an aggregate level. Furthermore, it can help city officials recognise the complex underlying factors of transportation use and develop more efficient and sustainable urban transportation systems.

6.2 Secondary activity identification (SAs) and enrichment of SAs

6.2.1 Introduction

The purpose of this section is to address the second and third research objectives, which are to develop a model to identify SAs from SCD and to enrich SAs with land use attributes once the detailed surveys are not available. Thus, the heuristic approach is carried out to investigate the rest of the activities (SAs) called ‘others/other activities’ in the literature.

The section proceeds as follows. First, a heuristic SA identification algorithm is proposed in a methodology to extract SAs from SCD and SAs are presented as a holistic picture in a large city to reveal ‘where and when individuals move within the city’. Second, station-based SAs from SCD are enriched using POIs to address ‘why individuals move within the city’. Lastly, using London’s transit data as a case study, the model is compared with a volunteer survey to demonstrate its effectiveness and offer a cost-effective method to obtain travel demand research.

6.2.2 Methodology

In this section, the remainder of the activities are detected based on their direction (i.e., where an individual came from before an activity and where they went after an activity) relative to the primary locations and classified into four types – ‘before-work (BW)’, ‘midday (MD)’, ‘after-work (AW)’, or ‘undefined (UD)’– to represent individuals’ activity patterns (Pinjari et al., 2007; Rasouli and Timmermans, 2015; Wang et al., 2017). Such classification was used in early studies in activity-based modelling called CEMDEP (Bhat, Guo, Srinivasan and Sivakumar, 2004) for travel demand studies from survey data. The model mainly focussed on non-workers (home-based) and workers (work-based) using travel patterns, land use, socio-demographic and transportation level-of-service attributes under three layers such as pattern, tour and stops. (Rasouli and Timmermans, 2015). In this study, a similar activity classification was considered from SCD. This classification is considered to provide more information as compared to a single category of ‘others’ (Pinjari et al., 2007; Ma et al., 2017; Wang et al., 2017). Finally, in this section, the SAs are enriched into one of the semantic sub-categories using POIs – eating, entertainment, shopping, work, or other – to represent each trip’s purpose.

In addition, in the SA identification algorithm, work locations have been used as an anchor to identify SAs because the study mainly focuses on the centre of the city rather than residential areas. Besides, the additional data source, i.e., POIs is also more representative in terms of central locations rather than residential areas/locations in cities due to sampling bias. Alternatively, it would also be possible to describe SAs using home locations, where activities would be classified as before-home activities and after-home activities.

6.2.2.1 From and to activity locations

To understand SAs from the SCD perspective, activity from and to locations are defined as the nearest spatial information relating to an activity. After defining the primary locations of each individual in Section 6.1, the chain of activities is investigated based on the direction of travel (activity from or activity to) relative to those anchor locations. The from-activity location (FL) is defined as the last location before an activity. The to-activity location (TL) is defined as the next location after an activity. The FL and TL of each activity are translated into a binary vector based on the location types (home, work and other). Figure 6.10 illustrates the FL and TL of an SA for an individual. In this example, both locations match the work location (WL). The activity pattern suggests a midday activity where an individual travelled from work to carry out an activity and returned to work afterwards. FLs and TLs are extracted for each activity in a day for all individuals for further analysis.

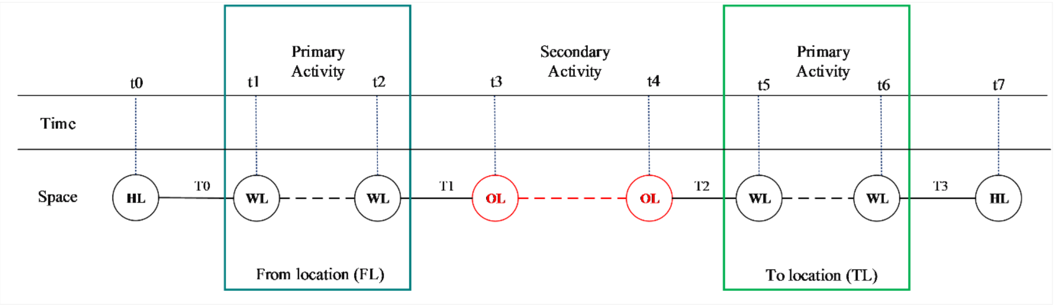


Figure 6.10 Schematic diagram illustrating activity chains in a day for an individual. In this case, the SA is marked in red as other location (OL) between the FL and TL. HL and WL refer to home and work locations as PAs, respectively.

6.2.2.2 Extracting secondary activities

To identify secondary activities in figure 6.11, first, an individual is selected for a particular day. Second, all of the individual's primary activities for the day are identified/considered. Third, ‘from and to locations’ are derived from the activities. As a result, SAs are categorized as ‘before-work (BW)’, ‘midday (MD)’, ‘after-work (AW)’, or ‘undefined (UD)’. After completing all days for the selected individual, the next individual’s data are considered using the same process. Before-work activities are defined as taking place between the home and work locations. If an individual came from home and spent time at a location before arriving at the work location, that activity is defined as a before-work activity. If an individual travelled from the work location for an activity and returned to the work location, or if they travelled from the home location and returned to the home location after the activity, the activity is labelled as a midday activity. If an individual came from the work location and spent some time at another location before going to the home location, that activity is labelled as an after-work activity. Finally, if the activity does not match any defined criteria strictly, which means the nearest spatial point relating to the activity (from/to) does not have complete anchor information (i.e., defined as home or work), then the activity is labelled as an undefined activity. An undefined activity can have either one anchor point (i.e., Home – Others, Work – Others, Others – Home, Others – Work) or none (Others – Others).

In the literature, only one type of SA (after-work activities) is investigated from public transport using time constraints (Wang et al., 2017). The first constraint is calculated as the threshold of the earliest time of departure from work and the second constraint is the finishing time of the tube lines, restricting individuals’ after-work activities to before midnight. However, in the present study, SAs are extracted using anchor points as well as the direction information of those locations. Therefore, the proposed algorithm can capture individual-level starting and ending working hours (flexible working hours) as a holistic picture with before-work and midday activities as captured in travel surveys (Rasouli and Timmermans, 2015)

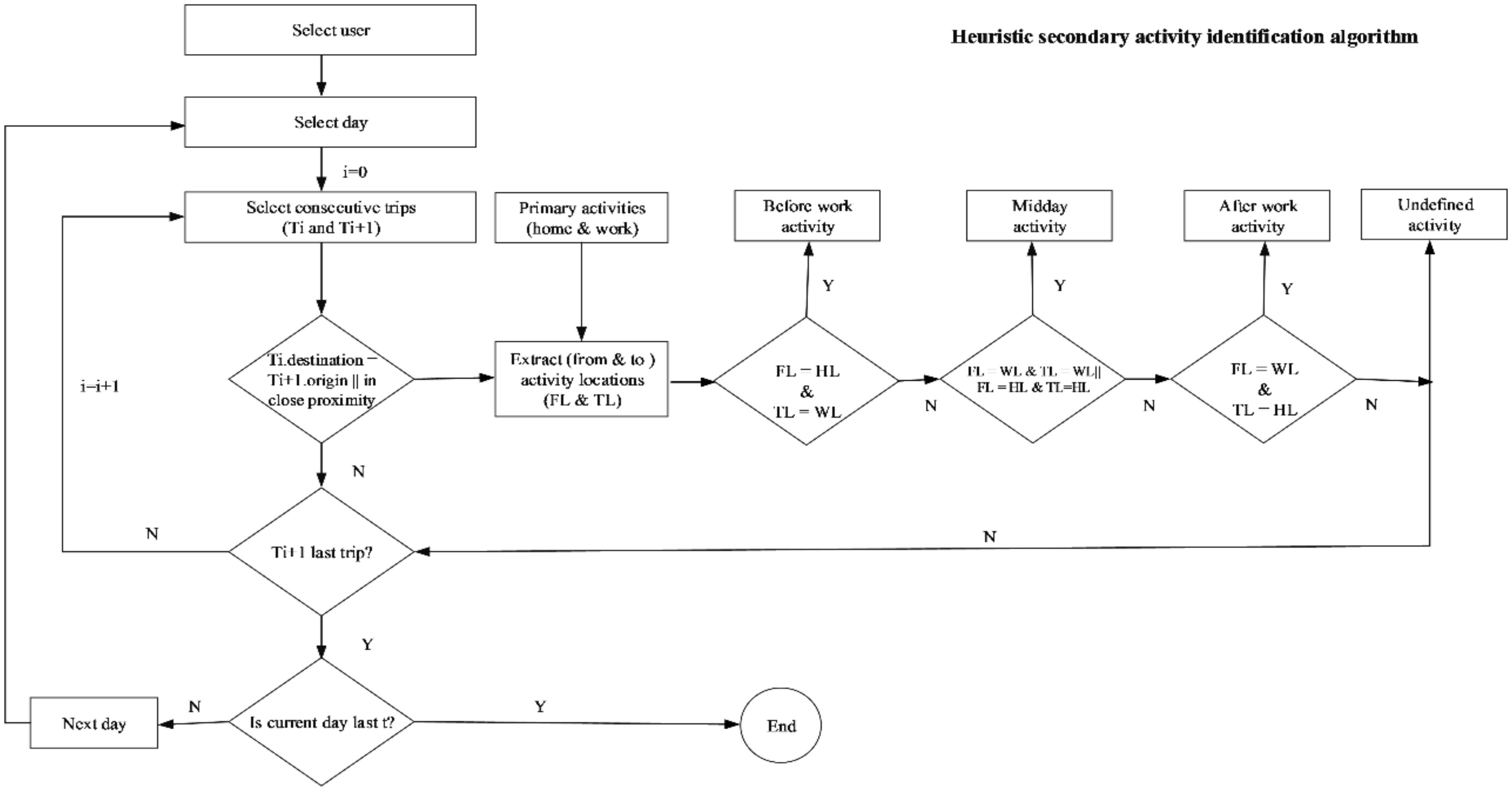


Figure 6.11 Flowchart of the SA identification algorithm for each user, where FL and TL are from and to activity locations, respectively and HL and WL are home and work locations for each user, respectively.

6.2.2.3 Enriched secondary activities

To assist in the inference process, large spatiotemporal transport data are commonly coupled with auxiliary information such as land use data and POIs, which can provide information on the type of performed activities (Noulas et al., 2015; Gong et al., 2016), thus facilitating inference tasks such as activity prediction and activity pattern classification (Hasan and Ukkusuri, 2014).

In activity-based modelling – designing a travel demand model based on the characteristics of activities gathered from surveys – PAs such as home, work and school are used for long-term forecasts and are usually considered 'mandatory activities', the least flexible in terms of scheduling, while SAs are mainly considered 'maintenance activities' (dropping off and picking up children and shopping) or 'discretionary activities' (eating out, entertainment, social visits, other recreational activities and doctor visits) (Castiglione, Bradley and Gliebe, 2015). Similar activities were found within the smart card dataset. However, from a land use policy perspective, the main objective is to optimise the use of city centres to prevent congestion or the desertion of areas at certain times. This is achieved by controlling for operating hours through planning permission (Montgomery, 2017). As trips to work, as well as eating, shopping and entertainment trips, are impacted by the opening and closing hours of establishments, these hours are the most appropriate measure for demand forecasting within cities (Alsger, 2017). In contrast, visiting a park, walking on a bridge and social visits are less time dependent and generate inconsistent trips on the transport network; this discretionary character is not enough to warrant policy changes within cities (Alsger et al., 2018). This study's POI dataset has similar activities, including opening and closing hours. Therefore, the POIs are sub-categorised as eating, entertainment, shopping, work (part-time) and others (travel and transport, outdoor and recreation and home) for the enrichment of the SAs at the tube/train station level.

POIs from Foursquare data are categorised based on industry classification of the visited place and easy-to-determine trip purposes (Rashidi et al., 2017), which offers some advantages compared to other data sources, such as land use data. For instance, the total number of check-ins can be used to assign different weights within a trip purpose inference model. Using opening and closing hours of POIs assists in presenting urban flow within cities (Rashidi et al., 2017). In this part of the study, we have matched the temporal attributes of SCD with POIs to refer to SAs. Reducing irrelevant POIs using a temporal variable produces a more meaningful inference of the SAs. The following steps have been taken:

i) Spatiotemporal match using the activity–POIs consolidation algorithm: a catchment area from each station (walking distance) and the starting and ending hours of each SA are used to filter POIs based on their opening and closing hours. This is applied for each activity from SCD to POIs to control the over-representation of activity types such as eating. The activity–POIs consolidation algorithm is detailed in Section 5.3.

ii) Station profile using the weighted average (WAi): activities are presented in five categories (eating, entertainment, shopping, work and others), denoted as:

$$\mathrm{Si} = [\mathrm{Ai}_{\mathrm{eat}}, \mathrm{Ai}_{\mathrm{ent}}, \mathrm{Ai}_{\mathrm{shop}}, \mathrm{Ai}_{\mathrm{work}}, \mathrm{Ai}_{\mathrm{others}}] \qquad (6.2)$$

Where Si is station i and Ai is activity at station i. The total count of activities at a specific station is defined as:

$$\mathrm{Ai}_{\mathrm{total}} = [\mathrm{Ai}_{\mathrm{eat}} + \mathrm{Ai}_{\mathrm{ent}} + \mathrm{Ai}_{\mathrm{shop}} + \mathrm{Ai}_{\mathrm{work}} + \mathrm{Ai}_{\mathrm{others}}] \qquad (6.3)$$

To obtain a better description of the station's characteristics, the weighted average (WAi) of each activity in station i is calculated. In Equation 6.4, only $\mathrm{WAi}_{\mathrm{eat}}$ is presented:

$$\mathrm{WAi}_{\mathrm{eat}} = (\mathrm{Ai}_{\mathrm{eat}} * 100 / \mathrm{Ai}_{\mathrm{total}}) * (\mathrm{Ai}_{\mathrm{total}} / \textstyle\sum_{\mathrm{i}=0}^{\mathrm{n}=0\ldots\mathrm{n}} \mathrm{Ai}_{\mathrm{total}}) \qquad (6.4)$$

Thus, the scaled activity values in Equation 6.2 can be replaced as:

$$\mathrm{Si} = [\mathrm{WAi}_{\mathrm{eat}}, \mathrm{WAi}_{\mathrm{ent}}, \mathrm{WAi}_{\mathrm{shop}}, \mathrm{WAi}_{\mathrm{work}}, \mathrm{WAi}_{\mathrm{others}}] \qquad (6.5)$$

Based on Equation 6.5, each station has five weighted values according to its nearest station's POI profile for geographic representation.

6.2.3 Case study

London is used as a case study to apply the proposed models using two types of transit data (SCD and SSCD) and land use data (POIs from Foursquare data). In this part of the analysis, even though the POIs from the Foursquare data have a broad classification (Section 4.2.4), only five POI activity types are considered for the enrichment part of this study: work (part-time), eating, entertainment, shopping and others. The details of SCD, SSCD, POIs and data pre-processing are outlined in Sections 4.2.1, 4.2.2, 4.2.4 and 5.1, respectively.

After extracting the SAs from SCD, the results of the analysis are presented as follows to explain temporal characteristics of SAs, validation of SAs from existing models and enrichment of SAs.

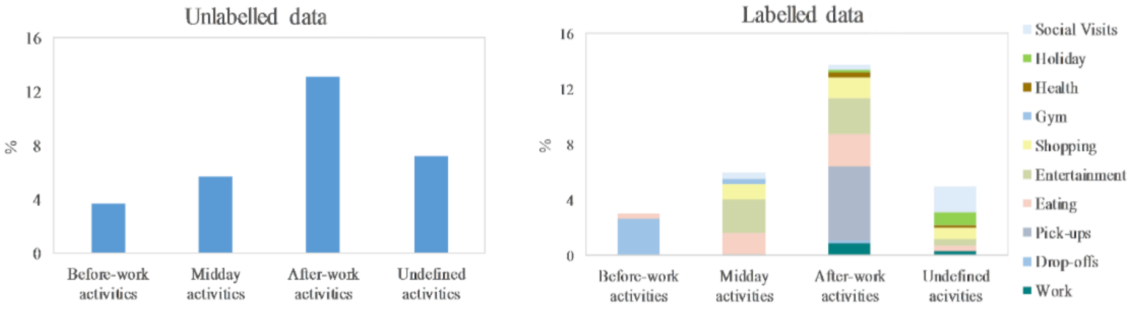


Figure 6.12 Representing SCD (unlabelled) and SSCD (labelled data) based on SAs.

Extracted SAs from SCD is compared to extracted and labelled activities by volunteers using Labelling-I (Section 3.2.2) in Figure 6.12. Almost 30 per cent of activities and 28.1 per cent of activities are SAs in SCD and SSCD, respectively. The highest and lowest activity counts in both datasets are after-work and before-work activities, respectively. Undefined activities account for 7.23 per cent of activities in the SCD and almost 5 per cent in the SSCD, where more information is observed about the nature of these activities. For instance, 1.8 per cent are labelled as social visits (the activity locations are not in the centre of the city [the City of London for this study] such as home locations) and 0.94 per cent are labelled as holidays (the activity locations are airports). The rest of the undefined activities are labelled as shopping (0.81 per cent), entertainment (0.44 per cent), eating (0.38 per cent), work (0.31 per cent) and other activities (0.25 per cent) such as walking in the city or park and doctor’s visits or other appointments.

6.2.3.1 The temporal characteristics of secondary activities

Although the algorithms are defined without considering the temporal variable, identified activities still have temporal characteristics such as boarding/alighting times and duration of the activity, as well as the day of the activity. Therefore, the temporal characteristics of SAs are investigated further. Figure 6.13 illustrates an aggregate analysis of SAs and their characteristics, such as activity duration on each day (heat maps) and activity start and end times (line charts). The first column presents before-work activities. The heat map of before-work activities highlights a consistent two- to three-hour window during the weekdays; the trend is less significant during the weekends. In addition, the start and end times peak from 08:00 to 09:00 (blue line) and 10:00 to 12:00 (red line), respectively. The second column represents midday activities. The heat map illustrates a consistent window of two to four hours during the weekdays and two to five hours during the weekends. There are two peaks in the total counts of start and end hours. The start hours peak at 12:00 and 16:00, while the end hours peak at 14:00 and 18:00. The smaller peaks appearing almost three hours later might be due to home-to-home midday activities, especially during the weekends.

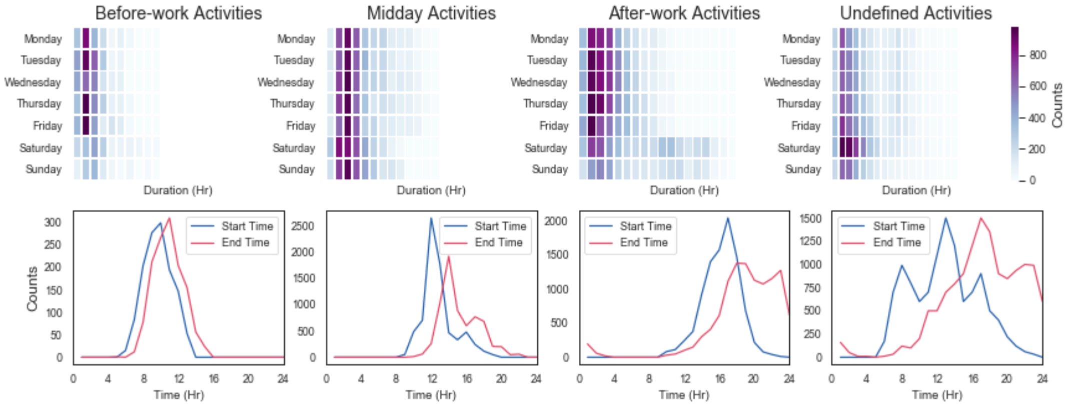


Figure 6.13 The temporal distribution of SAs (before-work, midday, after-work and undefined) by day of the week and time of day.

The third column presents after-work activities. The heat map of after-work activities shows that there is some difference in activity duration between weekdays and weekends. After-work activities are confined to a two- to four-hour window during the weekdays. However, after-work activities appear over a longer period during the weekend and do not present as consistently as on weekdays. This inconsistency is illustrated by the most intense colour appearing on Saturdays compared to the weaker colours on Sundays. Also, the line charts show a different pattern of start and end hours compared to before-work activities. The counts of start hours for after-work activities reach a peak from 15:00 to 17:00. The counts of end hours, on the other hand, show two peaks around 19:00 and 23:00. The starting of after-work activities may be regular due to fixed departure times from work, especially during the weekdays. The ending times of after-work activities are irregular due to the variable time needed to reach home locations. The last column shows the temporal variation of undefined activities. The duration of activities from the heat map is less than five hours during the weekdays compared to more than five hours for weekends, especially on Saturdays. The start hours of undefined activities present three peaks at 08:00, 12:00 and 16:00, while the end hours present three peaks at 11:00, 17:00 and 23:00. The reason for these three peaks is that they share an anchor point from one of the key locations – home or work – in the dataset.

As a result, the duration is an important characteristic in defining activities (Chakirov and Erath 2012; Zou et al., 2016). The duration of PAs is defined in the literature as ten to fifteen hours for home activities and six to nine hours for work activities (Chakirov and Erath, 2012; Devillaine, Munizaga and Trépanier, 2012; Zou et al., 2016). This study is the first to define the duration of the secondary activities – four hours or less, especially during the weekdays – while other studies have used time constraints directly to extract SAs (Wang et al., 2017). However, individuals' start and end hours of SAs present temporal variation. Thus, in this study, the sequence of activity chains for each individual is used to have an accurate estimate for travel purposes and the results of the analysis are presented at an aggregate level.

6.2.3.2 Validation of the identified Secondary Activities

Comprehensive validation of the activities identified from the SCD is difficult to achieve due to the limited availability of the survey data. Two validation approaches are used to test the accuracy of the proposed algorithm. The first approach is to compare the results of the proposed algorithm to SSCD. The proposed SA identification algorithm resulted in accuracies of 80 per cent for after-work, 76 per cent for before-work, almost 70 per cent for midday and 57 per cent for undefined activities.

The second method is to use another model as a baseline with which to compare the accuracy of after-work activities only (Wang et al., 2017). The estimation of after-work activities using the baseline approach is only 67.5 per cent accurate due to the earliest departure time from work being set as 16:00. However, almost 20 per cent of the after-work activities are labelled in the dataset as children's school pick-ups during 15:00–17:00. Besides, the baseline can be extended to include before-work and midday activities with time constraints of 07:00 to 09:00 and 12:00 to 14:00, respectively. This yields a success rate of only 62 per cent and 56 per cent, respectively, using the same validation dataset. Thus, the proposed algorithm provides better identification of SAs and demonstrates a complete picture compared to the existing baseline model.

6.2.3.3 The semantic meaning of secondary activities

The semantic meaning of SAs is illustrated using the number of check-ins for each of the five categories (eating, entertainment, shopping, part-time work and others) from SCD and POIs for London stations in Figure 6.14. The aggregated analysis of an individual trip's purpose according to where and why people spent their time within the city is presented in this figure. Each SA is explained using the three charts, reading anti-clockwise: first, the peak locations of SAs are presented in the London map using only SCD. Second, the percentages of activity types from the total counts of POI check-ins are illustrated for each SA in the bar chart (Eq 6.5). Finally, both the identified SAs and their enrichment from POIs are presented for the selected central London stations – Oxford Circus, Piccadilly Circus, Green Park, Leicester Square, Tottenham Court Road and Bond Street.

First, the count of before-work activity stations is illustrated. As well as central London stations, some residential and school stations are highlighted – Richmond, Clapham Common and Hampstead. From the count of check-ins from POIs, work activities (work and school) are found to be the main type of before-work activity, with the highest probability at about 42 per cent. The basis for this finding is twofold: first, work activity locations are identified using a duration threshold of more than 5 hours in Section 6.1. Therefore, some part-time workers' work activity which is less than 6 hours might be captured as before-work activity. Second, student activities (pick-ups and drop-offs) are highlighted as work activities in this study because TfL does not have student card information for children under 11 years old. (Chapter 4). Therefore, it is expected that most of the drop-off activities would appear here under work activities once parents consider different travel choices, e.g., bus or car. The main activity at the majority of the selected central London stations (except for Green Park and Bond Street) is also inferred to be work.

The count of midday activity in stations mainly in central London and Stratford are considered to represent office workers' lunch breaks. Due to home-to-home midday activities, especially during the weekends, the bar chart illustrates not only eating activities but also activities such as shopping and entertainment. In addition, upon inspection of the central London stations, the purpose of the majority of midday activities was determined to be 'eating' except for Oxford Circus and Bond Street, where it was determined to be 'shopping'.

The counts of after-work activity stations on the map suggest that there is an overlap with some of before-work and midday activities; this combination can also be seen from the SSCD (Figure 6.12). Although the total count of after-work activities in residential and school locations is similar to that of before-work activities, the total number of identified after-work activities overall (13.14 per cent) is more than the total numbers of both before-work (3.72 per cent) and midday (5.71 per cent) activities combined. This suggests that the biggest contribution comes from entertainment (40 per cent), eating (34 per cent) and shopping (almost 19 per cent) activities rather than activities at school locations, which can be seen from the bar chart as well as the central London stations. Furthermore, the selected London stations show that nearby stations have similar inferences under a certain category. For instance, the inferred activity at Covent Garden and Leicester Square is entertainment while the inferred activity at Oxford Circus and Bond Street is shopping.

Finally, the counts of undefined activity locations show that almost 2 per cent of the undefined activity appears either at interchange stations or London airports. A few examples are highlighted in the London map. The first reason is that the 20 min transfer time might be less for those transport hubs in metropolitan cities. Some studies have excluded those interchange stations as a step before defining primary locations (Li et al., 2015). However, most of the interchange stations in London have large spaces for passengers to spend their time while they are waiting to continue their journeys. Hence, the study provides station-based enrichment as well, even though the total count is presented simply as eating, entertainment, or shopping in the bar chart. Finally, the selected central London station activity is classified as eating and shopping. The reason for this could be that those undefined activities have a spatial point from one of the anchor points, such as work locations. For instance, an individual who comes from work may use a different mode of transport to go back home, such as a car or bike.

The SA identification algorithm is able to highlight before-work, midday, or after-work activities as locations strictly. In this study, the starting and ending hours of SAs are compared with the opening and closing hours of POIs for each location before assigning the highest probability of land use POIs. Hence, Figure 6.14 has provided meaningful enrichment.

Furthermore, the same London stations are enriched by different activity types during the day using the classification of SAs. For instance, Leicester Square is inferred as a work location under before-work activities, eating location under midday and entertainment location under after-work, while Marble Arch is inferred as a work location under before-work activities, eating location under midday activities and shopping location under after-work activities. That shows how incorporating SAs with dynamic POIs (opening and closing hours and user check-ins) may lead to meaningful activity inferences.

As a result, the framework uses big data sources to investigate individual SAs to infer travel purposes. The approach demonstrates how SAs can be derived from SCD using the proposed SA identification algorithm. The spatiotemporal characteristics of SAs for each individual have quantified in the aggregate analysis that the majority of SAs are four hours or less, especially during the weekdays.

The study presents how SAs are combined with POIs, which helps in the meaningful inference of SAs despite the limitations of POIs, such as contribution bias and demographic bias. As a result, the purpose of travel is different for the same stations and individuals during different times of day in dynamic cities, a finding which aids in representing urban flow as a more accurate and complete picture. The outcome is beneficial for urban and infrastructure/transport planners to develop more sustainable cities.

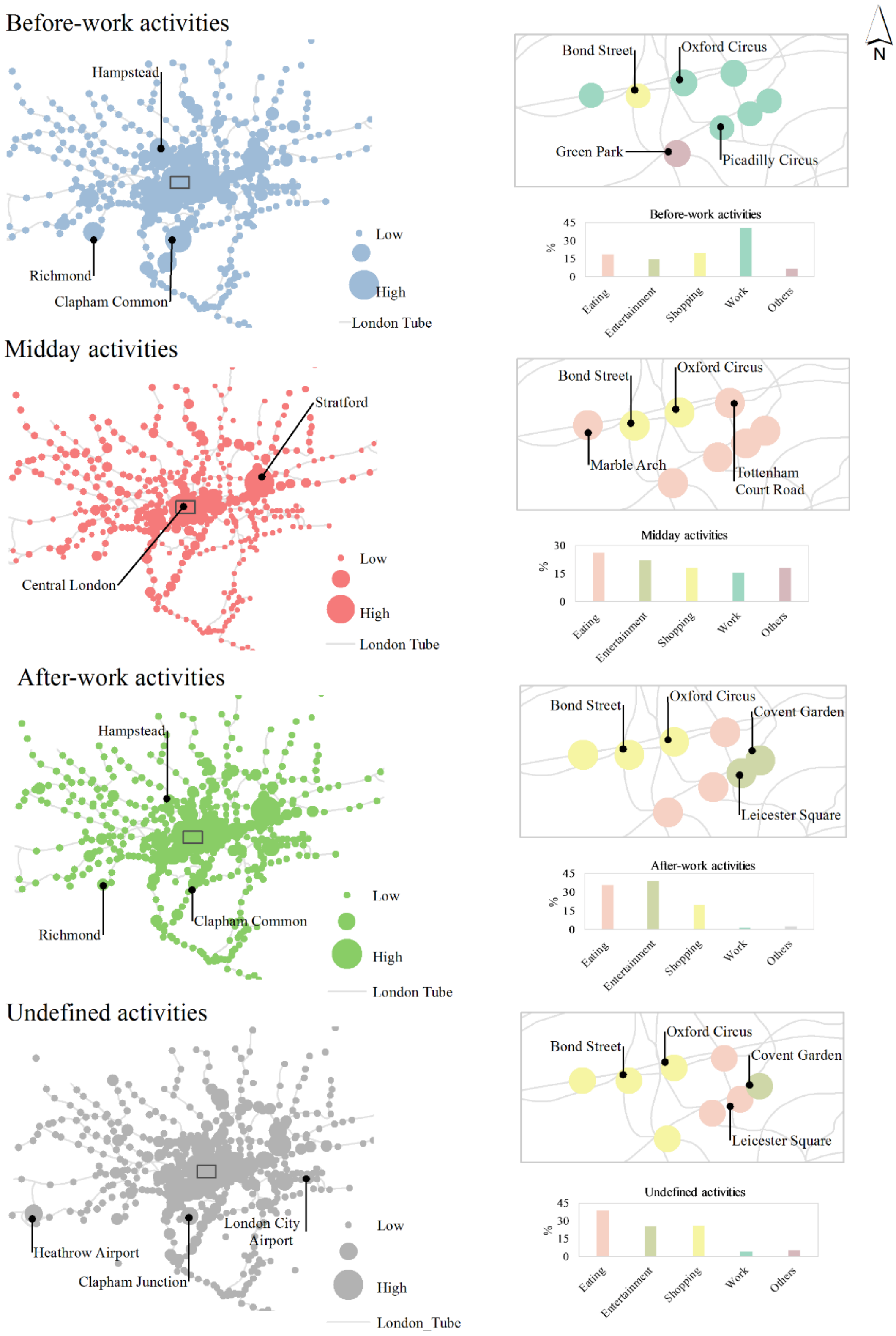


Figure 6.14 SAs mapped at the station level to infer the semantic meaning of SAs.

6.2.4 Summary

The large volume of individual-level SCD presents opportunities to generate new insights into travel behaviour research and urban modelling. This part of the study aimed to demonstrate a framework specifically for enriching the semantics of SAs by combining SCD with additional POIs in a complex urban environment. A heuristic model was proposed to identify SAs and enriched with ancillary POI data to estimate the likely nature of the SA once the detailed surveys are not available.

First, the proposed SA identification algorithm can detect meaningful locations for SA types. A heuristic SA identification algorithm was applied to tube/ train travellers in London and the algorithm reached accuracies of 80 per cent for after-work activities, 76 per cent for before-work activities and 70 per cent for midday activities based on volunteers' responses. Thus, the high-level classification of the activities fosters an understanding of the travel behaviour of users and facilitates more efficient and sustainable development of urban transport systems. Secondly, a framework integrating the SA identification algorithm with auxiliary information was introduced to investigate the reasons for travel in the case of regular users. In London, the identified SAs were enriched using Foursquare data under five sub-semantic categories – work, eating, entertainment, shopping and others. Hence, linking human travel behaviour with urban functions demonstrates how trip purposes are dynamic during the day at the same station/location, which is beneficial for transport and city planners. Lastly, the proposed method offers a cost-effective approach to human mobility as an alternative to the traditional travel demand survey.

6.3 Chapter summary

This chapter explained activity detection from SCD using a heuristic approach. The motivation for the model and contributions of the study were highlighted in Sections 6.1 and 6.2. The proposed methodology was explained in terms of PAs and SAs, including the validation of the models in Sections 6.1.2 and 6.2.2, respectively. The results of the proposed models (PAs and SAs) were presented in a case study in Sections 6.1.3 and 6.2.3 and summarised in Sections 6.1.4 and 6.2.4, respectively.

Chapter 7

Trip Purpose Inference with Machine Learning Approach

7 TRIP PURPOSE INFERENCE WITH A MACHINE LEARNING METHOD [10]

This chapter introduces a methodological framework called ActivityNET, which uses Machine Learning (ML) algorithms to investigate trip purposes. Section 7.1 provides a brief introduction to the study. Next, Section 7.2 explains the model details under four sections, i.e., input features, model structure, evaluation and validation and predicting trip purposes from Smart Card Data (SCD). Then the results of the model as a case study in London are presented in Section 7.3. Finally, the method and the chapter are summarised in Sections 7.4 and 7.5, respectively.

7.1 Introduction

The purpose of this section is to address the fourth research objective, which is to develop a model to predict Primary Activities (PAs) and Secondary Activities (SAs) under the same framework. The proposed framework is ActivityNET, which works with ML algorithm to identify trip purposes from Smart Card Data (SCD) with land use attributes such as Point-of-Interest (POIs). The framework uses Artificial Neural Network (ANN), which is a sequence of algorithms that identify underlying relationships in a set of data to classify multi-class trip purposes in a large dataset with high dimensionality (Xiao, Juan and Zhang, 2016).

ML models involved trip purpose inference using random forest (RF) (Breiman, 2001), support vector machine (SVM) (Cortes and Vapnik, 1995), logistic regression classifier (LR) and naïve Bayes (NB) from different data sources such as GPS, phone data, but rarely SCD. Compared with artificial neural networks (ANNs), these machine-learning approaches may have some drawbacks in the sense that multi-class trip purposes are inferred from noisy large datasets. RF is one of the best classification algorithms for large sizes of data. However, categorical variables with different levels built from an ensemble of trees create a bias with high dimensions that affects accuracy (Deng, Runger and Tuv, 2011). SVM belongs to the family of binary classifiers and uses kernel-trick to represent the data into a higher dimension (hyperplane), where the data can be linearly

[10] Part of this chapter has been presented in the following publications: N Sari Aslam, MR Ibrahim, T Cheng, H Chen, Y Zhang 2021. “ActivityNET: Neural Networks to Predict Public Transport Trip Purposes from Individual Smart Card Data and POIs.” *Geo-spatial Information Science* 24 (4): 711–721.

separable. However, once data size and dimensionality increase in multi-class classification, SVM cannot handle ambiguity, which trip purpose inference innately involves, due to class determination's non-probabilistic (binary) output (Jaakkola and Haussler, 1999) and provides lower accuracy. In addition, LR estimates the probability of an event occurring based on the data provided. However, the model is too simplistic to handle a complex relationship in large datasets. Lastly, NB used the Bayes theorem, suitable for multi-class classification problems. However, the model assumes that all features are independent, which means it cannot tolerate any relationship between features in large datasets with high dimensions. Within this mind, ANN can handle the complexity of land use attributes with noisy input data effectively. ANN does not depend on data belonging to any respective distribution. Neural Networks (NN) are capable of handling the dimensionality of the problem using spatial dependencies in a large dataset with high accuracy and low computing time (Lee and Buchroithner, 2010; Xiao, Juan and Zhang, 2016). Thus, ANN is appropriate to infer trip purposes in this study. The comparison between ANN and other baseline models is presented in Section 7.3.2.

The feasible framework includes the following: First, travel data (Survey Smart Card Data (SSCD) and SCD) are pre-processed using the steps mentioned in Chapter 5 and input features are extracted under three sub-groups: activity characteristics (activity start and end time, activity duration), day characteristics and land use characteristics. Second, input features from SSCD are fed to the ActivityNET model to predict trip purposes under PAs (home and work) and (enriched) SAs (entertainment, eating, shopping, child drop-offs/pick-ups and part-time work activities). Third, the proposed model is evaluated and validated to benchmark models to determine the effectiveness of the model for further developments in human mobility research and transport planning. Fourth, the trained model is used for trip purpose prediction from SCD in a large city, i.e., London.

7.2 Methodology

The proposed ActivityNET framework predicts trip purpose using ML methods to learn the relationship between input and output values. In addition, the model doesn’t need to define PAs first and then look at the rest of the data for SAs or enrich SAs. In contrast, all data points are considered at the same time as a single model for the predictive analysis, i.e., PAs and SAs. That makes the model flexible enough to learn from each data point rather than defined rules (e.g., SAs are not dependent on defining PAs).

ActivityNET framework is presented in two sections. Phase I focuses on extracting features from SSCD and SCD and combining them with POIs in three steps: spatial information match, temporal information match and attractiveness of the POIs.

The reason for these steps is to allow both datasets to link on common dimensions i.e. location and time. Then the attractiveness of the location has to be represented at the level of each activity, which requires the additional processing step. These steps allow data usability by converting it to make it compatible with each other, for instance, extracted activity is first considered based on the walking distance buffer and then checked with the temporal attributes, activity start and end time with POI's opening and closing hours before considering check-ins in each activity type.

The details of this section are explained in Section 5.3. Phase II presents the structure of the model, along with multiple scenarios and validation processes. The methodology is illustrated in Section 3.4.

7.2.1 Phase I: Data pre-processing to extract input features

SCD are combined with POIs using the activity–POIs consolidation algorithm (Section 5.3). Thus, a data point has a dimension that corresponds to input features. In other words, extracted activities from journey data have been considered using the rest of the features from the input feature table as a vector to represent a relationship between input and output. Temporal features (activity characteristics, day characteristics) and spatial features (land use characteristics) are presented in Table 7.1, accordingly.

7.2.2 Phase II: The structure of the artificial neural network

An artificial neural network (ANN) is applied for predictive analysis to classify multi-class trip purposes from SCD with land use attributes, i.e., POIs. The model can handle the dimensionality of the issues in a large dataset with high accuracy (Lee and Buchroithner, 2010; Xiao, Juan and Zhang, 2016).

Even though ANN can conduct complex problems with high dimensions, a class imbalance is still a common problem in classification techniques in ML. Once the class distributions are highly imbalanced, many classification learning algorithms have inadequate predictive accuracy for the infrequent class (minority class). Appearing rare data points in the dataset are considered either noise or outliers resulting in more misclassifications of the positive class (minority class) than the dominant class (majority class) even though the minority class is more important than majority class (Ali, Shamsuddin and Ralescu, 2015). To reduce the imbalance ratio in training data, one of the data level approaches, anointed sampling, is considered to rebalance the class distribution, which is a random over-sampling technique (randomly duplicates data points in the minority classes) and a random under-sampling technique (randomly removes data points from majority classes) (Ali, Shamsuddin and Ralescu, 2015; Brownlee, 2020c).

Table 7.1 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 ACT_ST_TIME ACT_EN_TIME	Duration of the activity (hours) Start time of the activity in 24 hours End time of the activity in 24 hours
Day characteristics	Weekdays/ends	1: If the activity has happened on weekdays/0: Otherwise
Land use characteristics	HOM WOR ENT EAT SHO REC TPORT	Aggregated check-ins for home locations Aggregated check-ins for work locations Aggregated check-ins for entertainment locations Aggregated check-ins for eating locations Aggregated check-ins for shop locations Aggregated check-ins for outdoors/recreation Aggregated check-ins for transport stations

In this area, there is another argument about spatial and temporal attributes. For instance, early researchers sometimes used only temporal (Faroqi and Mesbah, 2021) or the combination of spatial and temporal features (Alsger et al., 2018; E. Kim, Y. Kim and D. Kim, 2020). The reason is that the complexity of the land use attributes decreases the accuracy, especially in traditional ML algorithms such as logistic regression classifier, naïve Bayes, etc. or in what-if scenarios. Therefore, this study also focuses on investigating how the model behaves once temporal attributes (input feature without POIs) or spatial and temporal attributes (input feature with POIs) are used in the model.

The details of the model structure, illustrated in Figure 7.1, are as follows:

1. Input layer: The first layer of neural networks transfers the information from input features using the same dimensionality. To investigate class imbalance issues, a random over-sampling technique (ROS) and a random under-sampling technique (RUS) (Brownlee, 2020c) are applied and compared to unchanged values (UD) to understand the level of this problem. In addition, the dimensionality of the layer is increased and decreased, including (input dimension = 11, with POIs) and excluding (input dimension = 4, without POIs) spatial features to evaluate overall accuracy with different scenarios in the model (Section 7.3.1).

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

The dropout regularisation technique (Hinton et al., 2012) is considered after the hidden layers step with a dropout rate of 0.5 to reduce overfitting. Cross-entropy loss is 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.

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

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

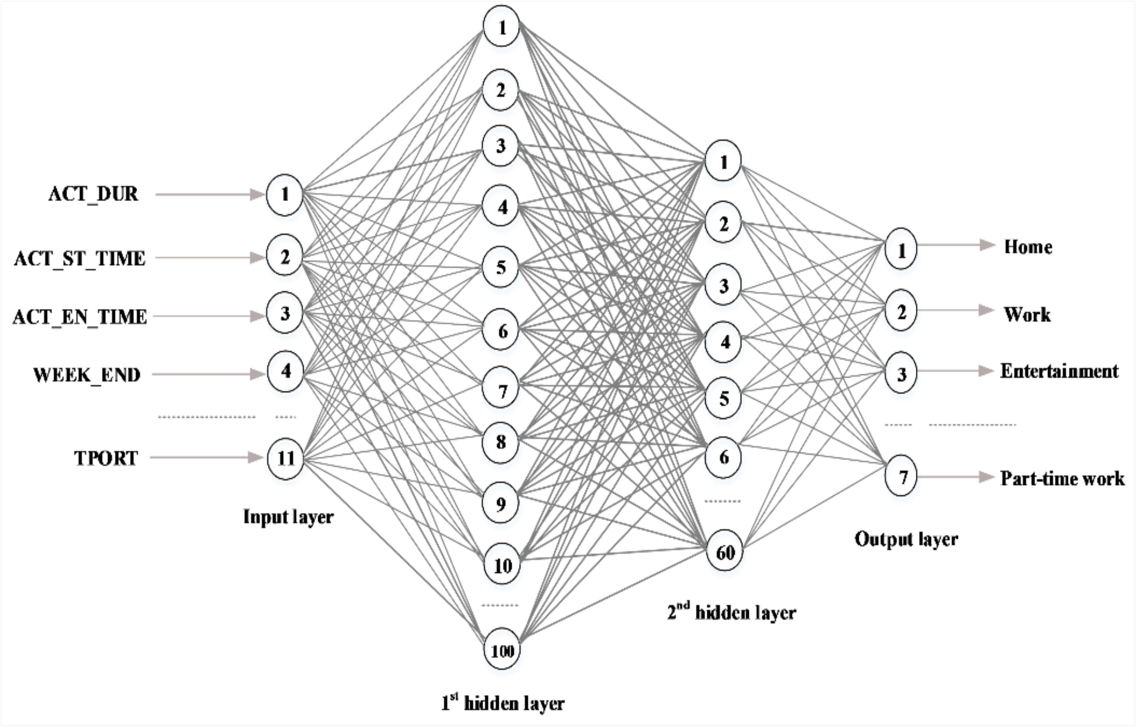


Figure 7.1 The structure of the ANN model

As a result, the proposed model is trained with 70 per cent of the data (train data) and can be used as a predictive model to identify trip purposes with the rest of the dataset (30 per cent, test data) or other datasets.

7.2.3 Evaluating and validating the model performance

Validation of the model is crucial for the study, which provides better insights into the prediction as compared to accuracy that focuses on only correct prediction (True Positive (TP) and True Negative (TN)). There is a need to understand false prediction from precision and recall metrics, which are dependent on False Positive (FP) and False Negative (FN). The definitions of TP, TN, FP and FN are given in Table 7.2. In addition, F1-score summarises both measurements for the model performance. However, there is still a limitation in terms of the distribution of FP and FN in each class. Therefore, the confusion matrix is required to illustrate where the model is misclassifying during the prediction. Apart from the first two evaluation methods, two more arguments were combined for the evaluation of the model: comparing the model's effectiveness to other baseline models and presenting the variance in each model using cross-validation, which means dividing the training dataset into k fold, use one fold for validation and the rest for training. Thus, it is possible to see the highest and the lowest accuracy in the dataset as compared to other based line models.

The model's valuation under two sections is presented and the first part is as follow: (1) evaluating the model performance with measures of precision, recall and F1-score as shown in Table 2 (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.

Table 7.2 Performance evaluation metrics and the definition of acronyms

Acronym	Definition	Metrics	Formula
TP	The model correctly predicts the positive class. In other words, actual values are positive and predicted as positive	Precision (prec)	$\frac{\text{TP}}{\text{TP} + \text{FP}}$
TN	The model correctly predicts the negative class. In other words, actual values are negative and predicted as negative	Recall (rec)	$\frac{\text{TP}}{\text{TP} + \text{FN}}$
FP	The model incorrectly predicts the positive class. In other words, actual values are negative and predicted as positive	F1-Score (F1)	$2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$
FN	The model incorrectly predicts the negative class. In other words, actual values are positive and predicted as negative	Accuracy	$\frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$

The second part of the evaluation determines the accuracy obtained from the highest probability of land use information (Alsger et al., 2018). Thus, after phase I, activities have been inferred from SCD using the highest probability of POIs as a benchmark model and the results are compared with SSCD. The activity type validation is calculated as follows:

$$V_{A_T} = \frac{CA_T}{TA_{T_n}} \times 100 \tag{7.1}$$

where A_T is the activity type (home, work, etc.), V_{A_T} is the percentage of the validated activity type, CA_T is the correctly identified activity points from SSCD using the highest probability of land use (POI) values and TA_{T_n} is the total number of n check-ins for the activity type. Hence, CA_T is normalised based on the total number of check-ins. As a result, the accuracy for each activity type is presented in Section 7.3.3.

7.2.4 Predicting trip purposes from SCD

In this section, the ActivityNET framework, a trained model from SSCD, is used to predict trip purposes from SCD instead of test data from SSCD (Phase 2B in Section 3.4). The results are visualised based on the temporal characteristics of the predicted trip purposes to demonstrate the temporal signature of each activity type. Then the consistency of findings is checked with common sense. Besides, similar locations/stations are compared between inferred activities from semantic enrichment of SAs in Section 6.2 and predicted activities from ActivityNET.

7.3 Case study

London is considered as a case study using two types of available datasets for the analysis: SSCD and Foursquare data (POIs). These datasets are explained in Sections 4.2.1, 4.2.2 and 4.2.4 and their pre-processing steps are detailed in Chapter 5, respectively.

The reasons for inferring trip purposes in only seven categories of activities are summarised as follows: first, the collected survey data (SSCD) is under the same seven categories, which helps comparison analysis during validation of the proposed models in the heuristic approaches in Sections 6.1 and 6.2. Second, the National Travel Survey (NTS) also classifies and recognises the activities under the sub-categories home, work, children drop-offs/pick-ups, shopping, eating, entertainment (museum, theatre and nightclubs) and other work-related activities (part-time work or personal business) (DfT, 2019). In the Foursquare data, only outdoors and recreation activities are excluded from the dataset. This is due to policy measures, including opening and closing hours, resulting in a lack of consistency when investigating trips outdoors and recreation activities such as visiting a park or walking on a bridge (Alsger et al., 2018).

7.3.1 The result of the model using multiple scenarios

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 a different number of data points may create class imbalance issues in the ML approach (Brownlee, 2020c).

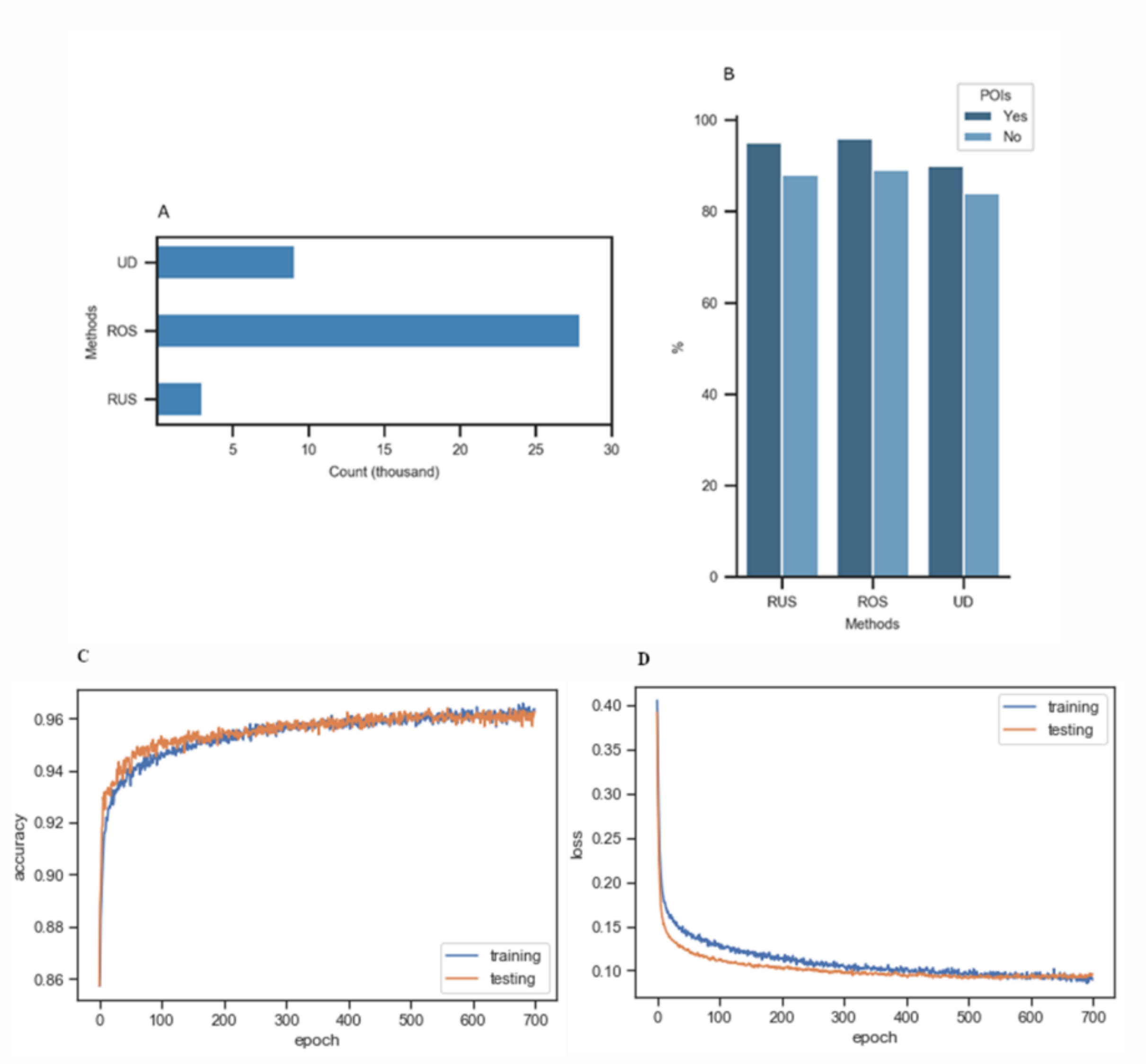


Figure 7.2 The number of data points in each method (A); the results of overall prediction with/without POIs using unchanged data (UD) and random under- and over-sampling (RUS and ROS, respectively) techniques (B); model accuracy (C); and loss (D) using random under-sampling with POIs.

In survey data (SSCD), almost 60 per cent of the activities are PAs and 40 per cent are SAs, which reveals that the count of each SA is much lower than the count of each PA. To reduce the imbalance ratio in training data, random over-sampling techniques (ROS) and under-sampling techniques (RUS) are compared to unchanged values (UD) to evaluate overall accuracy in each class. In addition, the classification accuracy using different scenarios, such as including and excluding land use attributes (with and without POIs, respectively), is also evaluated in this stage to obtain the best possible model performance.

According to the results in Figure 7.2, using RUS with POIs achieved an overall accuracy of almost 95 per cent. Conversely, without POIs, this number decreases almost seven per cent for an overall accuracy of 88 per cent. ROS with POIs increased the accuracy of the model to 96 per cent and without POIs, the accuracy was 89 per cent. Finally, without balancing any classes, the overall accuracies were 89 per cent and 83 per cent with and without POIs, respectively. In addition, Figures 7.2C and 7.2D illustrate the convergence of the model accuracy (the number of correctly predicted values is based on the total number of the predicted values) and loss (the difference between the predicted value and the actual value) using under-sampling with POIs.

Training speed using the RUS has a lower impact compared to the ROS. In other words, training ANN takes longer in a large dataset, without any substantial improvement in the performance measures. Therefore, the rest of the analysis is presented using random under-sampling with POIs.

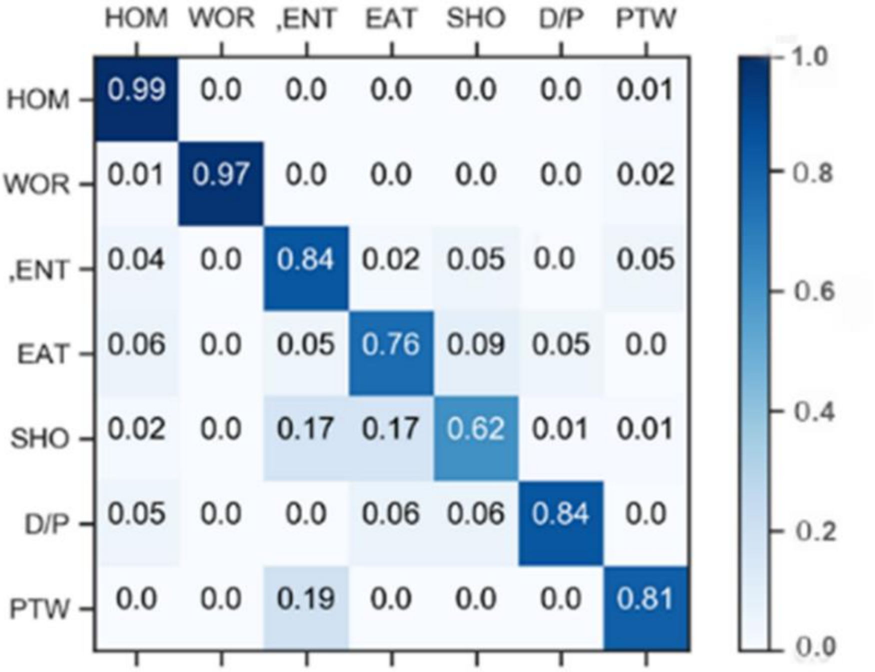
7.3.2 Evaluating the model performance

The first part of the model's valuation has presented in three sections. The first one evaluates the model using three performance metrics in each class: precision, recall and F1-score (Brownlee, 2020a). The high precision and recall values represent the low FP and FN from the model, respectively. That is attained with the best results in precision, recall and F1 for work activities (PAs) and child drop-offs/pick-ups and part-time work activities (SAs), as illustrated in Table 7.3.

Table 7.3 Prediction performance using precision, recall and F1-score on test data.

Trip purpose		Precision	Recall	F1-score
Home	Primary activities	0.84	0.98	0.90
Work		0.99	0.97	0.98
Entertainment	Secondary activities	0.73	0.84	0.78
Eating		0.74	0.76	0.75
Shopping		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

The second approach presents the confusion matrix to clarify the prediction performance for each class independently. The confusion matrix, using test data as shown in Figure 7.3, illustrates that the probability of a correct prediction is larger than that of misclassification. The lowest prediction score is for shopping activities with 17 per cent misclassified as entertainment or eating activities. The misclassification may suggest overlapping temporal variation in the three activities. 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 the PAs is fairly close, with 99 per cent of home and 97 per cent of work activities correctly predicted. The best prediction of inference among SAs is obtained for drop-offs/pick-ups (84 per cent) and PT work activities (81 per cent) as a result of regular activity patterns. The rest of the SAs present similar outcomes with high temporal stability and regularity, such as 84 per cent of entertainment activities and 76 per cent of eating activities correctly predicted.



	HOM	WOR	,ENT	EAT	SHO	D/P	PTW
HOM	0.99	0.0	0.0	0.0	0.0	0.0	0.01
WOR	0.01	0.97	0.0	0.0	0.0	0.0	0.02
,ENT	0.04	0.0	0.84	0.02	0.05	0.0	0.05
EAT	0.06	0.0	0.05	0.76	0.09	0.05	0.0
SHO	0.02	0.0	0.17	0.17	0.62	0.01	0.01
D/P	0.05	0.0	0.0	0.06	0.06	0.84	0.0
PTW	0.0	0.0	0.19	0.0	0.0	0.0	0.81

Figure 7.3 Inferring trip purposes using the confusion matrix with POIs.

The third approach is the comparison of the model with other baseline models using 10-fold cross-validation, in which trip purpose prediction accuracy is compared with several baseline models: 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 SCD. Therefore, they are considered as baseline models to compare to the proposed model in this study.

As shown in Figure 7.4, the original data is randomly partitioned into ten subsamples and the highest accuracies of between 86 per cent and 99 per cent are achieved using neural networks with almost 13 per cent variance. The second highest accuracy of 84 per cent to 89 per cent is achieved using RF with almost 5 per cent variance. The third highest accuracy of 78 per cent to 81 per cent is captured using SVM, with the lowest variance. Last, LR and NB are presented 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 SCD with the help of land use POI data.

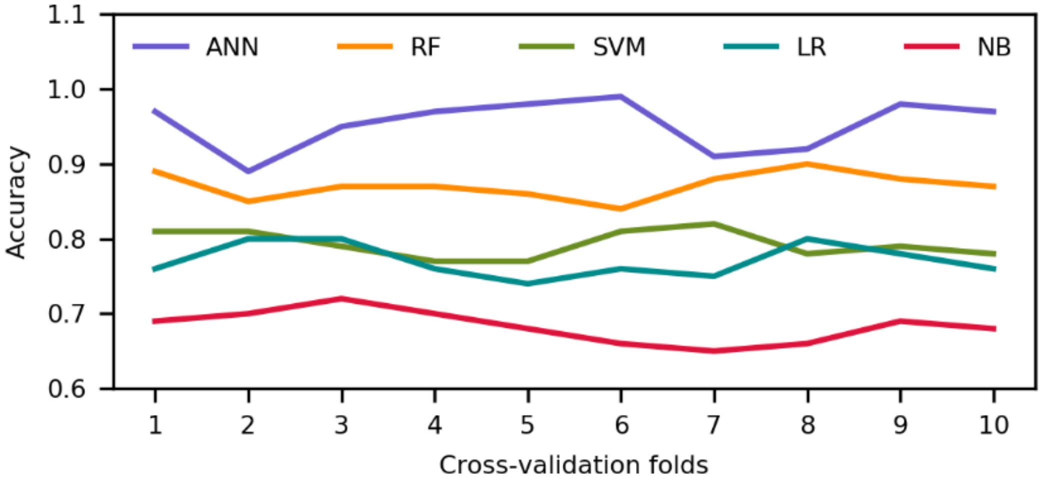


Figure 7.4 The accuracy of 10-fold cross-validation using baseline methods.

7.3.3 Validating the model from the literature

This section aims to compare the accuracy of the proposed framework to existing models using the highest probability of land use information from POIs (Alsger et al., 2018). Note that this part of the enrichment is obtained after applying equation 7.1. As a result, 51 per cent of work and 49 per cent of home activities, 44 per cent of entertainment, 33 per cent of eating, 35 per cent of shopping, 34 per cent of drop-off/pick-up and 39 per cent of part-time work activities are identified as correct. As a result, the proposed ActivityNET framework demonstrates a higher success rate than 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. Furthermore, more sophisticated techniques provide higher accuracy to predict trip purposes (Anda, Erath and Fourie, 2017).

7.3.4 The representation of trip purposes from SCD

This section illustrates the predicted values from SCD to understand findings based upon temporal and spatial representations, which assists in validating the results of the model. For instance, Sparks et al. (2017) characterise temporal signatures for eating and shopping locations, i.e., restaurants and other retail, from Foursquare data for different cities. Opening/start time of the locations or closing/end time of the locations are important measurements for policy implications in urban environments. Thus representing temporal signatures of the activities provide valuable details to understand the behaviour of the trip purposes. Note that temporal signatures may vary from various cities. Besides, the model from ActivityNET may also involve biases due to mismatched data sources and data collection methods. Second, the results in this section from ML approach can also be compared to Sections 6.1 and 6.2 from the heuristic approach.

Figure 7.5 illustrates temporal characteristics of predicted trip purposes from SCD using the activity duration for each day of the week (heat maps) and the start (blue line) and end (red line) times of the trips (line charts/temporal signature) under seven sub-categories: home (H), work (W), drop-offs/pick-ups (D/P), part-time work activities (PTW), entertainment (ENT), eating (EAT) and shopping (SHO).

The first subsection of the figure presents home activities. The heat map of home activities highlights a consistent 10- to 16-hour window, especially Monday through Thursday. Fridays and weekends are less significant compared to the first four days of the week. In addition, the line chart shows a sharp peak for the counts of start and end times at 17:00 and 8:00, respectively. However, some commuters start and end home activities later than those peak hours.

The heat map for the second subsection, work activities, illustrates a consistent window between 7 and 11 hours during the weekdays, with a similar but less significant pattern during the weekends. The counts of start and end hours of work activities peak from 07:00 to 9:00 and 16:00 to 18:00, respectively. Some commuters end work activities later than peak times, i.e., around 22:00.

Overall, PAs show regular temporal behaviour from heat maps, especially during the first four days of the week. Work activities on Fridays are regular, but home activities. The temporal signature of the PAs based on the counts of start and end hours is mainly represented by two peaks in the morning and evening.

The third subsection, showing drop-off/pick-up activities, highlights that the activity durations during weekdays (three hours) are less than during the weekends (four hours). The reason for this might be that most people have less spare time during the weekdays due to tight working schedules, as compared to weekends. There are two sets of peaks in the counts of start and end hours for drop-off/pick-up activities. Drop-offs start and end around 8:00 (first peak, blue and red lines). Pick-ups (second peaks) start between 15:00 and 17:00 (blue line) and end between 17:00 and 20:00 (red line).

The heat map of the fourth subsection, showing part-time work activities, highlights a three- to seven-hour window during the weekdays. The duration of part-time work on Saturday are similar to weekdays, while Sunday part-time work activities tend to have durations of less than six hours. The start/end times of the activities (the first peak) are more regular in the morning compared to the rest of the day. There are peaks at approximately 13:00 and 17:00 for the part-time work start times (blue line) and 17:00 and midnight for the end times (red line).

The heat map of the entertainment activities in the fifth subsection highlights a three-hour window on weekdays and a four-hour window on weekends, especially Saturdays. The counts of entertainment activities' start and end hours present one peak significantly, which starts around 16:00 to 17:00 and ends between 18:00 and midnight.

The heat map in the sixth subsection for eating activities presents a two-hour window for weekdays and a three- to four-hour window for weekends. The counts of the start hours for these activities (blue line) show three peaks at midday, 14:00 and 17:00, after which there is a sharp decrease from 18:00 to 22:00. Similarly, the counts of the end hours of eating activities (red line) also show three peaks at midday, from 16:00 to 17:00 and from 18:00 to 22:00.

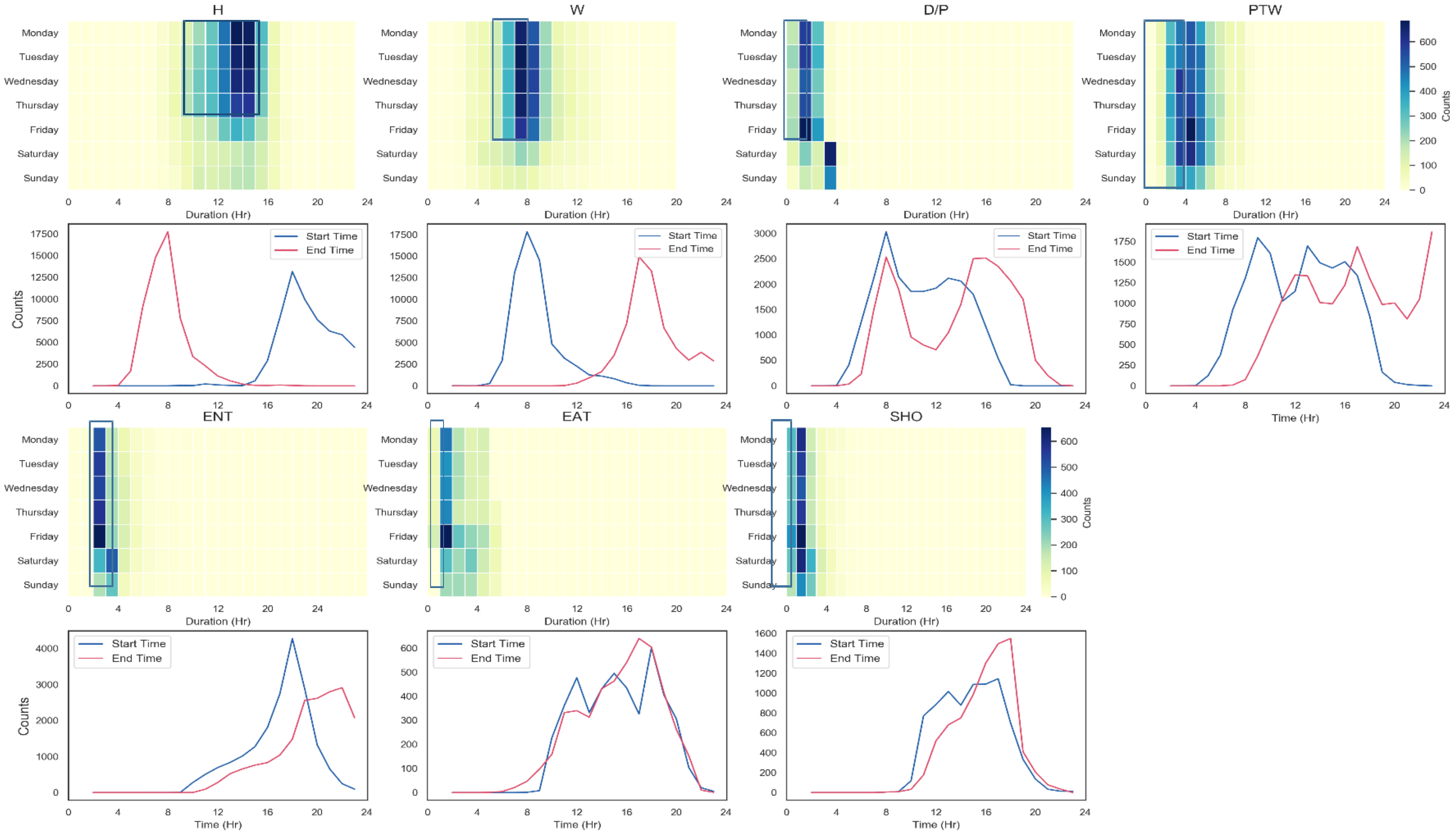


Figure 7.5 The temporal characteristics of trip purposes from SCD.

The heat map of the seventh subsection presents windows of less than three hours and four hours for shopping activities during the weekdays and weekends, respectively. The counts of the start hours of shopping activities show two peaks, from 10:00 to14:00 and from 15:00 and 17:00. The counts of the end hours of shopping activities show a weak/ insignificant peak from 10:00 to14:00 and a sharp peak around 16:00 to 18:00.

Each SAs represents different temporal behaviour observed in Figure 7.5. For instance, of the SAs, only drop-off/pick-up activities show a significantly different pattern between weekdays and weekends. However, one common theme is that the counts of the start and end hours of SAs are regular during the morning peaks compared to the rest of the day, except for entertainment activities.

The spatial distribution of the predicted values is based on some London stations – Oxford Circus, Leicester Square and Covent Garden for central London, Stratford for inner London and East Croydon and Ealing Broadway for outer London – are illustrated to show the types of activities. The results are presented under six temporal windows, such as morning (before 12:00), midday (12:00 to 17:00) and afternoon (after 17:00) for weekdays and weekends. Thus, findings can be compared to the enrichment of SAs in Section 6.2.

The first chart in the series represents the predicted count of activities at Oxford Circus, which has high counts of work activities during the morning hours on weekdays. Predicted values represent that shopping activities at Oxford Circus have appeared midday during the weekdays/ends more than other activities. According to the enrichment of SAs in Section 6.2, Oxford Circus has a high count of work activities for before-work, shopping for midday and after-work activities, similar to predicted results.

Furthermore, the selected central London stations, such as Covent Garden (Figure 7.7) and Leicester Square (Figure 7.8), which are nearby locations (can be seen in Figure 6.15), show similar inferences under a particular category. The high count of work activities is predicted during the weekdays/ends' morning hours. The rest of the days do not represent home or work but secondary activities, except eating. On the other hand, the enrichment of SAs in Section 6.2 also represents work activities for before-work and entertainment for after-work activities, except eating at midday (Figure 6.15).

The fourth chart (Figure 7.9) represents the predicted counts of activities at Stratford (inner London) station. The high count of work activities is predicted during the weekdays/ends' morning hours. In contrast, predicted high count home activities are weekdays/ends' afternoon hours. Besides, SAs have predicted midday during the weekdays/ends.

East Croydon and Ealing Broadway (outer London) stations are presented in Figures 7.10 and 7.11. These locations are examples of commercial centres outside central London so that they can serve as home and work locations for locals, captured from Figures 7.10 and 7.11, including Figures 6.7 and 6.8.

As a summary, the duration is an important characteristic in defining activities (Chakirov and Erath 2012; Zou et al., 2016) and is defined in the literature as ten to fifteen hours for home activities and six to nine hours for work activities (Chakirov and Erath, 2012; Devillaine, Munizaga and Trépanier, 2012; Zou et al., 2016). Besides, work activities were determined with eight hours for most regular commuters in Section 6.1. In this part of the study, the predicted values highlight a consistent eleven to sixteen hours window for home and a seven to eleven hours window for work activities from Figure 7.5, consistent with common sense.

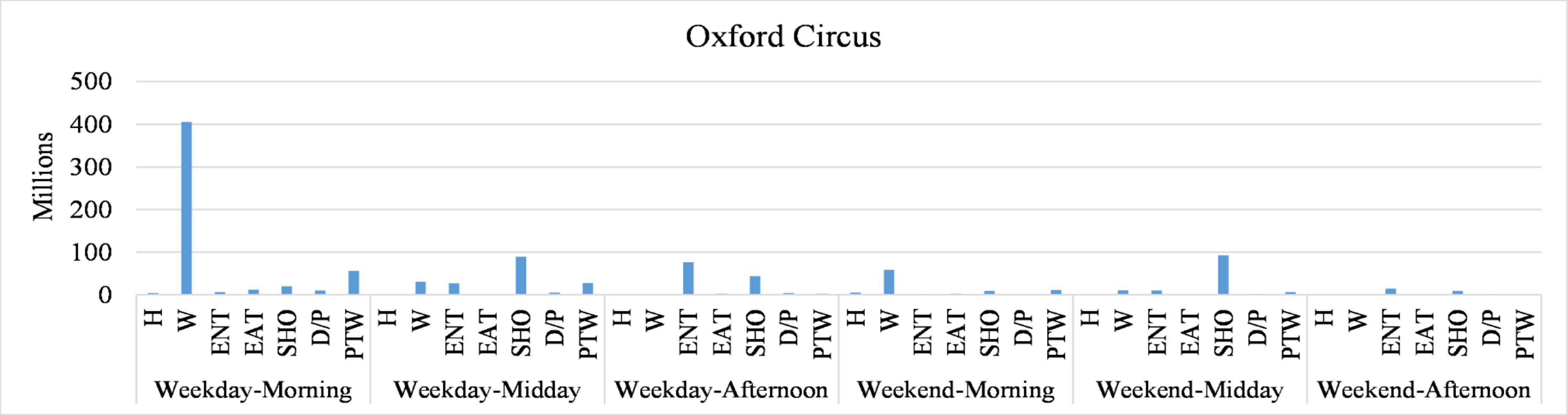


Figure 7.6 Trip purposes prediction at Oxford Circus (central London) station using the ActivityNET framework.

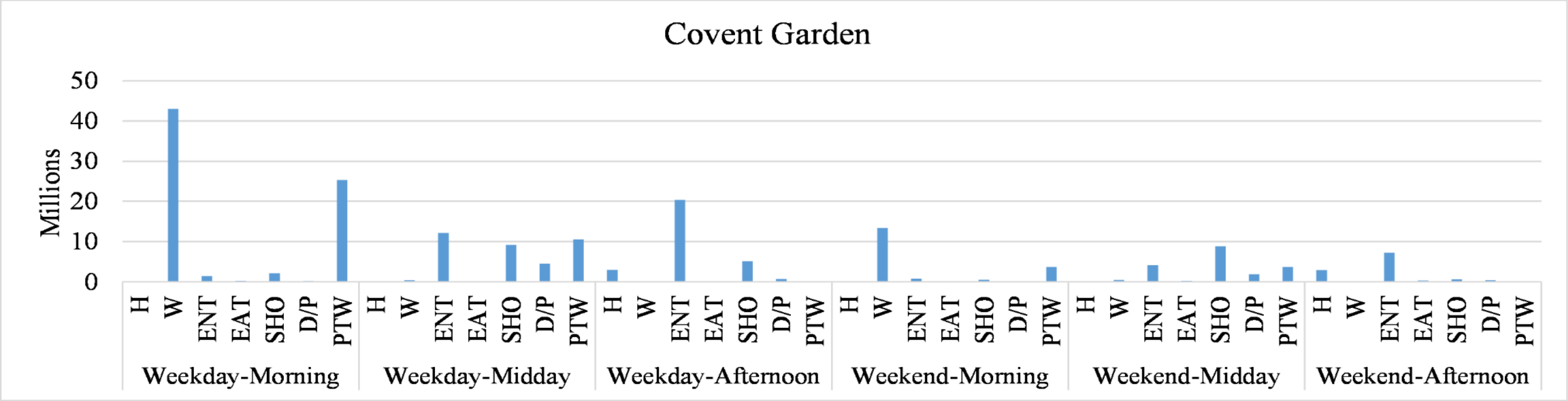


Figure 7.7 Trip purposes prediction at Covent Garden (central London) station using the ActivityNET framework.

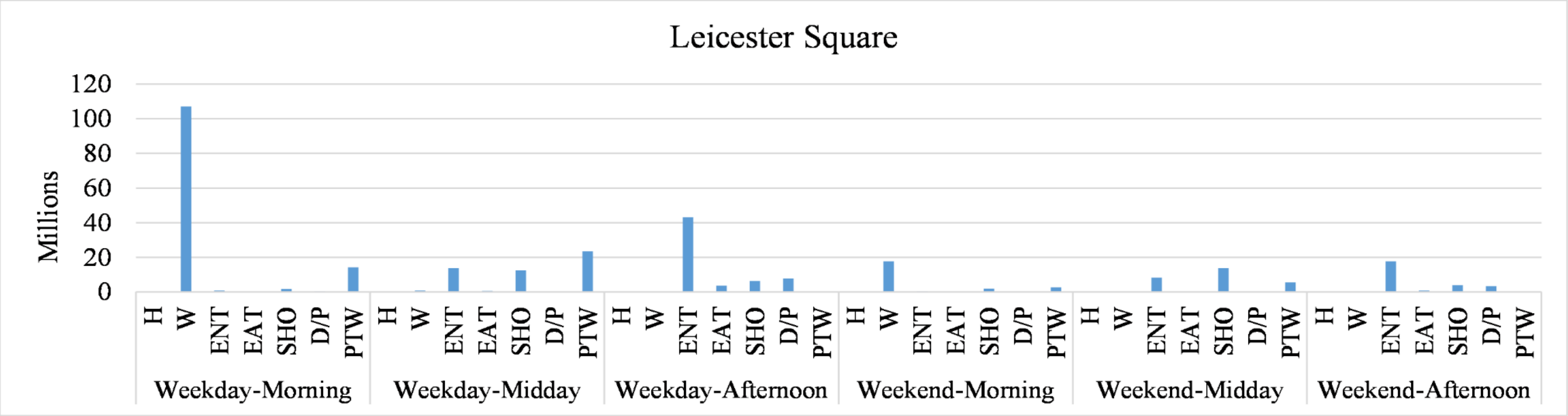


Figure 7.8 Trip purposes prediction at Leicester Square (central London) station using the ActivityNET framework.

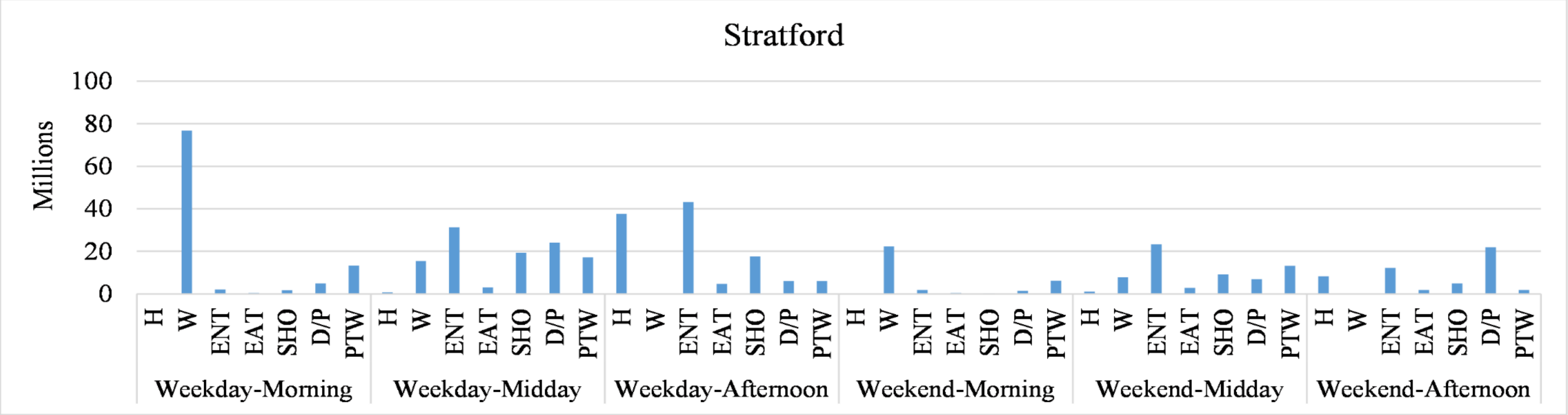


Figure 7.9 Trip purposes prediction at Stratford (inner London) station using the ActivityNET framework.

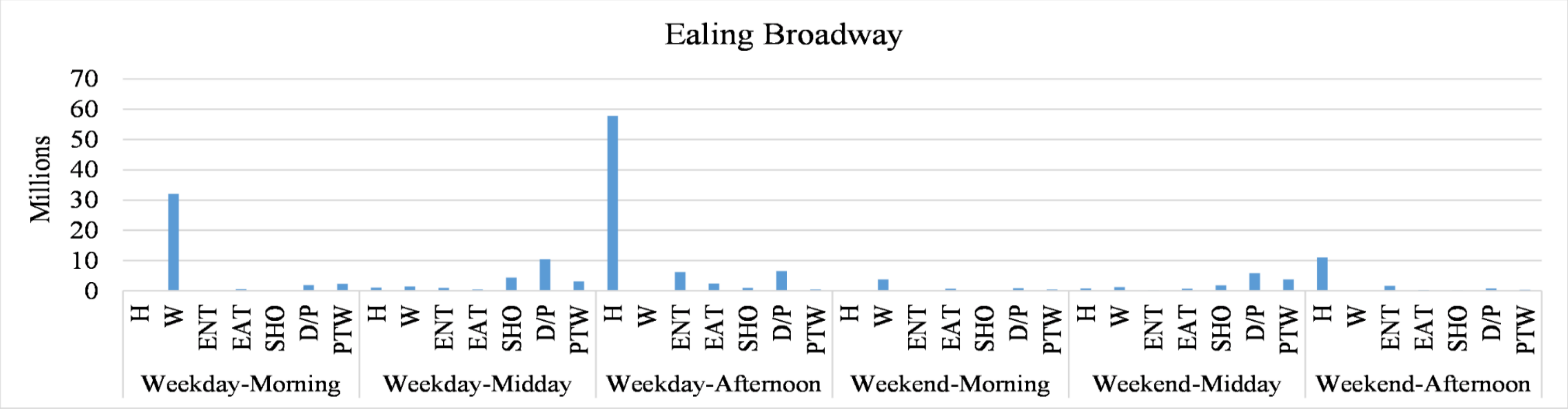


Figure 7.10 Trip purposes prediction at Ealing Broadway (outer London) station using the ActivityNET framework.

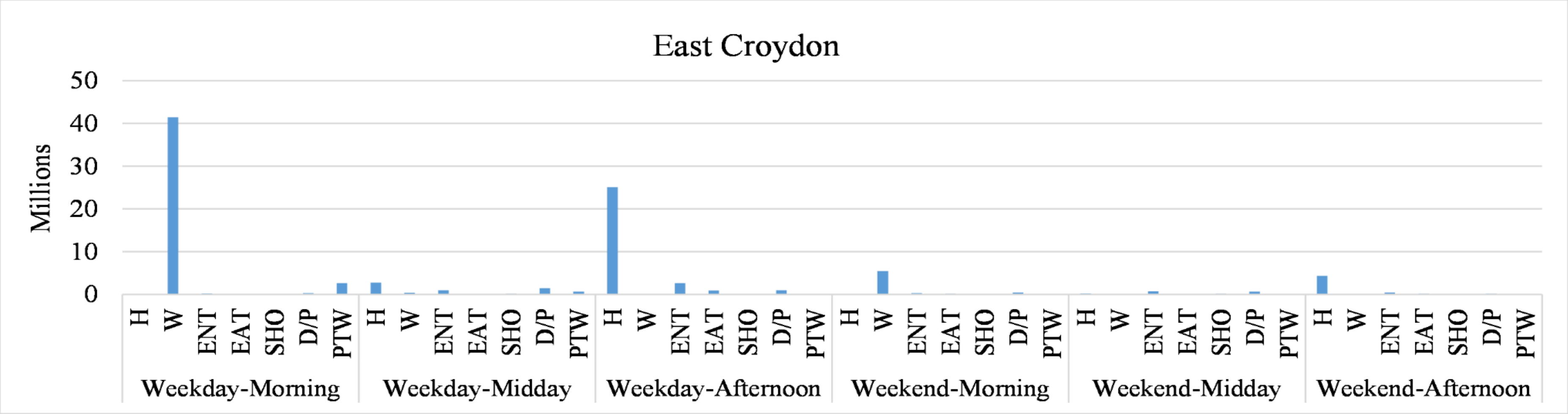


Figure 7.11 Trip purposes prediction at East Croydon (outer London) station using the ActivityNET framework.

Finally, Figure 7.12 presents the results of the identification of PAs and SAs from SCD aggregated at the level of each London station, illustrating that some of the stations in London are used for PAs as well as for SAs. The figure also highlights that most drop-off/pick-up and part-time work activities correspond to the stations used for home and work activities, respectively.

7.4 Summary of the ActivityNET framework

This study aims to predict trip purposes in a feasible framework using the spatiotemporal 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, demonstrates that a neural network provides improved accuracy in trip purpose prediction. First, the framework focuses on the proposed activity–POIs consolidation algorithm, which combines travel behaviour with land use information from POIs under three sub-groups: activity characteristics (activity start and end time, activity duration), day characteristics and land use characteristics. Second, the framework illustrates that the ANN model predicts trip purposes for PAs (home and work) and SAs (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 achieves 95per cent overall accuracy using random under-sampling techniques with POIs. In addition, high accuracy for PAs – 99 per cent for home and 97 per cent for work – are obtained from SCD. Furthermore, SAs provide improved accuracies of 84 per cent for entertainment, 84 per cent for drop-offs/pick-ups, 81 per cent for part-time work, 76 per cent for eating and 62 per cent for shopping activities. The validation results using the benchmark method show that the framework provides improved accuracy.

Moreover, ActivityNET framework promises accurate outcomes as compared to the proposed heuristic PAs and SAs. The reason is that the heuristic approach has been applied to identify PAs first and then SAs from SCD. Therefore, using PAs and SAs in the same framework makes the identification of SAs dependent on the correctness of the PAs. This separation is not the case for the ML approach, which investigates PAs and SAs simultaneously under the same framework with high accuracy. However, the model achievement depends on the novel SSCD with their associated SCD, which may require transport planners to revisit data collection methods for future developments.

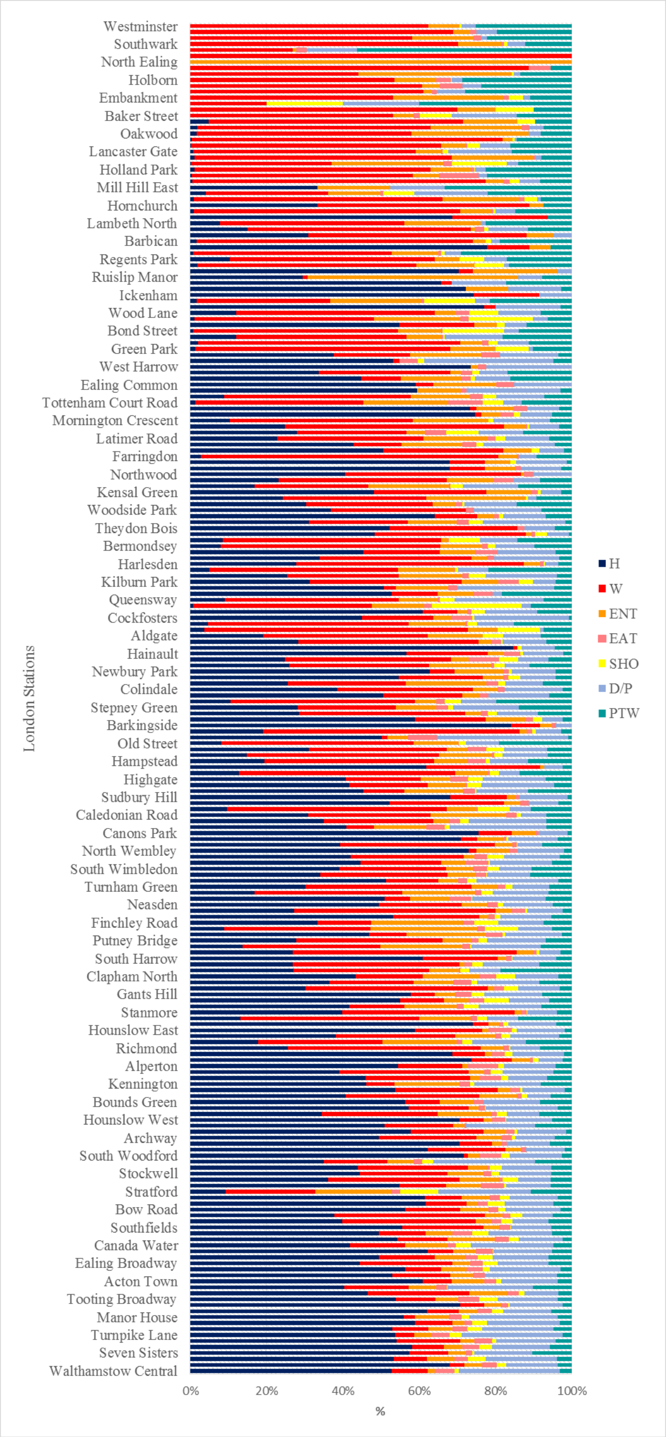


Figure 7.12 Predicted activity types by stations in London.

7.5 Chapter summary

This chapter explained trip purpose prediction from SCD using an ML approach. The contribution of the model with a brief introduction was presented in Section 7.1. The proposed methodology was explained in two parts – i.e., data pre-processing and the structure of the model, including evaluation and validation of the model performance and predicting trip purposes from SCD – in Section 7.2. Finally, the proposed framework results were presented in Section 7.3, followed by a summary of the ActivityNET framework in Section 7.4 and the chapter summary in Section 7.5.

Chapter 8

Conclusions and Future Work

8 CONCLUSION REMARKS AND FUTURE WORK

This chapter presents the conclusions of this thesis. Section 8.1 summarises the study's findings, while Section 8.2 details its contribution to the literature based on Chapters and Sections presented in this study. The limitations of the study and directions for future research are stated in Section 8.3. Finally, Section 8.4 presents the concluding remarks.

8.1 The summary of the study

Large volumes of individual SCD offer new opportunities for travel behaviour research and urban modelling. This work aimed to develop a framework that can utilise large and longitudinal individual SCD to investigate trip purposes. The cost-effective study also alleviates/eliminates the need for conventional travel surveys while enriching SCD, collecting SSCD and exploring trip purposes. To achieve this aim, five objectives, including their scope and limitations, were explained in Chapter 1.

Chapter 2 started with the usefulness of trip purposes in relevant research domains, such as transport, human mobility, retail analysis, environment and public health. Next, the conventional and new data sources are described for gathering trip purposes for transport planning. Transport SCD and other data sources were compared to highlight their scopes and limitations while mentioning the challenges in gathering trip purposes for validation requirements. Furthermore, the state-of-the-art models for trip purposes – the heuristic approach, the ML approach and the statistical approach from SCD and other data sources – were explained along with their definitions and limitations. Although the study primarily focused on the heuristic approach using only the characteristics of SCD without surveys, enrichment of SAs struggled to represent spatial dependencies. Moreover, the statistical approach could not deal with large and complex spatiotemporal data. On the other hand, ML techniques were appropriate for high-dimensional and large datasets to capture underlying patterns in the complex spatiotemporal dataset

Chapter 3 described the methodological framework for inferring trip purposes from SCD. Due to the limitations mentioned in Chapter 2, the data collection methodology for SSCD was explained in detail, followed by the databases and their challenges. Then, the proposed methodologies and the issues resolved were explained with/without detailed surveys.

The details of the study area, including London's public transport network and the data sources, were introduced in Chapter 4 and brief descriptions of the datasets with their scopes and limitations.

Chapter 5 highlighted the data preprocessing steps, i.e., data cleaning process, activity extraction, transfer activity definition and walking distance thresholds. Then, the relevant characteristics of the data sources used in the proposed methodologies, including journey count, start and end stations, visit frequency, activity duration, direction (from/to), opening and closing hours, activity type and the number of check-ins. Lastly, the 'activity–POIs consolidation algorithm' was explained using travel data with auxiliary information to infer trip purposes.

Chapter 6 illustrated a methodological framework to infer trip purposes using a heuristic approach. In this section, the PAs and SAs and the enrichment of SAs were presented using the characteristics of SCD. This part of the research is invaluable for representing trip purpose inference from SCD when surveys are unavailable.

Chapter 7 presented the ActivityNET framework to infer trip purposes using an ML approach. In this section, the model took advantage of innovative SSCD with land use attributes to provide high accuracy predictions of PAs and SAs under the same framework. In addition, the trained model was used to predict values from SCD to represent spatial and temporal characteristics of trip purposes. This part of the study is invaluable because it presents accurate outcomes for both PAs and SAs simultaneously. Besides, the study also encourages transport planners to improve the current data collection methods (i.e., SSCD).

Finally, Chapter 8 includes concluding remarks and highlights the major contributions of the thesis. The limitations of the methods are explained in detail. The chapter ends by highlighting several possible directions forward from the progression of this study.

8.2 Contributions to the existing literature

Innovative contributions are summarised based on the gaps mentioned in the literature accordingly.

- Data-driven heuristic activity identification models for PAs (Sari Aslam, Cheng and Cheshire, 2019) and SAs (Aslam et al., 2021) are introduced using SCD characteristics such as individuals' start and end stations, frequency, duration and direction (from/to) of the activities defined in heuristic algorithms using big data sources (without relying on surveys). Besides, the introduced models for PAs and SAs are incorporated under the same framework in Section 3.4.1. Hence, the first two gaps in the literature mentioned in Section 2.5 are addressed in Sections 6.1 and 6.2, respectively.

- ActivityNET framework is introduced to simultaneously identify PAs and SAs with higher accuracy using the deep learning method (Sari Aslam et al., 2021). Thus, the third gap in the literature mentioned in Section 2.5 is accomplished and presented in Chapter 7.
- To overcome the limitation of the ML approach, a new data source, i.e., survey SCD (SSCD), is collected for this research to investigate trip purposes from SCD, which was the last gap in the literature mentioned in Section 2.5. SSCD, with the innovative extension discussed in Section 8.3.4, offer a new data source and data collection methods for further developments in the big data era.

8.3 Critique of limitations and future work

Several innovations have been mentioned that contribute to the current literature. However, the outcome of this thesis can be a starting point for new research. The new data source (i.e., SSCD) with the proposed models can be combined in an application or software to enrich current SCD with extended trip purpose classifications. That way, the models and novel data sources can contribute to the re-thinking of existing sources of trip purpose information – i.e., surveys, SCD and other data collection practices. The limitations of the study are explained in the following sections

8.3.1 Improving trip chaining assumptions

The current model can be improved as follows: First, in the data processing section (Chapter 5), single trips are excluded from the smart card dataset before extracting activities, which is a limitation in the literature. Single trips (a day) could be included considering other weeks of the days with regularity parameters, i.e., frequency using appropriate thresholds. Second, bus journeys were excluded in the first phase of the study due to a lack of destination information. However, bus journeys could be included using appropriate time thresholds for train-to-bus trips, bus-to-train trips and bus-to-bus trips. Finally, multi-modal trips and activities could be analysed with appropriate rules and methods for a more holistic picture in large cities, including different transfer thresholds and walking distances in a large city

8.3.2 Investigating multi-purpose activities

Trip purpose detection inherently involves uncertainty (Xiao, Juan and Zhang, 2016; Faroqi, Mesbah and Kim, 2018). For instance, an extended shopping activity may be interrupted by an eating activity (e.g., drinking coffee/tea) at a location in which both shopping and eating places are available. Similarly, long hours of entertainment activities, such as at theatres or cinemas, may involve eating activities at a location in which both entertainment and eating places are available. Although it is difficult to separate those activities in individuals' daily lives, SSCD do not have any multiple/multi-purpose activities as labelled data for the analysis. Therefore, we assume that this is not an issue for this research. However, multi-purpose activities, e.g., a location for eating and shopping activities or a location for eating and entertainment activities, should be investigated in future trip purpose classification models once collected survey data have such multi-purpose activities as ground truth.

8.3.3 Methodological improvements in ML approach

Some researchers have mentioned that the current ML approach for inferring trip purposes needs to focus on unsupervised learning rather than confined classes, i.e., home, work and others (Faroqi, Mesbah and Kim, 2018). Such studies have used k-means clustering with a defined number of clusters. However, the results of the clusters are discussed based on the available surveys or prior knowledge of the study area (Medina and Erath, 2013). On the other hand, some other studies have preferred not to define the number of clusters using a hierarchical clustering algorithm but instead enriched the data from surveys and used only temporal features to capture high accuracy (Faroqi and Mesbah, 2021). The benefit of the ML approach using a large, high dimensional dataset and including the complexity of spatial dependencies is ignored in the current literature on SCD. Therefore, future studies should focus on unsupervised learning with high dimensional datasets, including land use's spatial dependencies.

Methodological improvements in ML approaches are suggested for inferring trip purposes from SCD by combining the current models (Anda, Erath and Fourie, 2017). For instance, the semi-supervised learning (SSL) technique may provide a better solution due to available limited survey data compared to large SCD (Sari Aslam et al., 2019). Besides, deep learning models can be combined with computer vision techniques using pictures or video records to enrich extracted activities from SCD for further development. Thus, not only the model but also enriched data represents inferring trip purposes under a different methodological framework.

8.3.4 Uniting data collection methods with a mobile phone application

Inferring trip purposes from SCD provides great opportunities for transport planners. Large, detailed SCD may empower current surveys at an unprecedented level. The advantages of SSCD are the longitudinal similarity to SCD and the size and scale, which provide broader coverage in urban environments, making it easier to represent the whole population compared to traditional travel surveys. On the other hand, the disadvantage of SSCD appears during the data collection process and labelling process, i.e., time-consuming. However, volunteer labelling of SCD can be leveraged using mobile phones or web-based applications to collect trip purposes from volunteers in a more streamlined way. For instance, travel data is currently web-based, downloadable information, which can be offered through an application designed by transport planners. Once volunteers register an application with the details of their SC and contactless card (if applicable), they can be guided by the applications to extract activities, which will automate data processing. The ActivityNET framework can be applied in this stage to label the trip purposes for volunteers, then ask volunteers appropriate questions about the labelled activities. The reason for this can be summarised as follows:

- SCD may present similar activities and there is no need to take volunteers' time labelling similar/the same activities. For instance, 60%–70% of each individual's data involves home and work activities. The right question may save the volunteers' time in this process.
- Asking individuals' activities, e.g., home and work, as a first question through the application (similar to surveys) may restrict different daily scenarios. For instance, in real scenarios, there will be more home and work scenarios than can be captured in a large dataset. If the question is related to that scenario, more information can be captured from volunteers.

- Furthermore, the extra required labelling process can also be tied to a loyalty program (Sharp and Sharp, 1997) to keep people interested in gaining points, money, or free travel such as a journey pass or daily or weekly passes.

Thus, the proposed trip purpose inference models with SSCD can represent correct and continuous outcomes than the current travel surveys, improving transport planning applications. Such data can be systematically collected using the same infrastructure without any additional cost. Although SSCD are not comparable to national travel surveys (NTS) as they belong to private establishments in different cities, this study may improve current systems, e.g., TfL, which may open up new ideas/opportunities for future development.

8.4 Final remarks

This thesis proposed a framework to infer trip purposes from SCD. Trip purpose inference methods using PAs and SAs models and ML models were compared based on accuracy, methods, advantages and disadvantages. Comparable to the literature, the highest accuracy for predicting PAs and SAs are achieved using ML approaches once the labelled (ground truth) data are available. The study also offered data collection techniques for a novel SSCD conducted from SCD using the nature of the travel card data volume and size.

The proposed models, either separate or together, have wide-ranging applications for trip purposes, mobility research and behavioural analysis from transport or other data sources. Although there is room for improvement, the current framework provides a good understanding of how SCD is useful and how it will be more useful in transport- and urban-related fields once surveys are automated as a function of the smart cards.

9 APPENDICES

Appendix A: Data collection templates

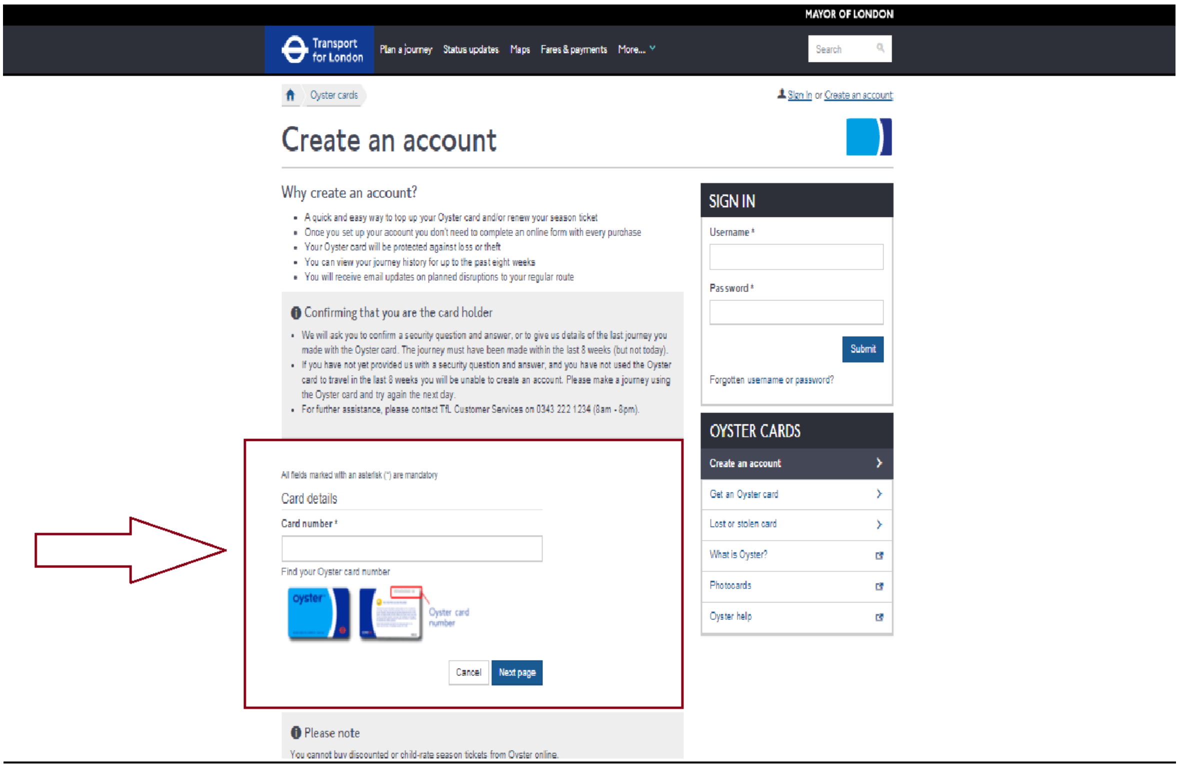
The following three templates for volunteers explain how travel data needs to be downloaded for this study.

Template 1: How can we register and download our smart card data?

1. Go to the website below and please create an account.

https://oyster.tfl.gov.uk/oyster/link/0004.do

2. Please fill the sections pointed out by arrow in the following figures.



Create an account

Why create an account?

- A quick and easy way to top up your Oyster card and/or renew your season ticket
- Once you set up your account you don't need to complete an online form with every purchase
- Your Oyster card will be protected against loss or theft
- You can view your journey history for up to the past eight weeks
- You will receive email updates on planned disruptions to your regular route

Confirming that you are the card holder

- We will ask you to confirm a security question and answer, or to give us details of the last journey you made with the Oyster card. The journey must have been made within the last 8 weeks (but not today).
- If you have not yet provided us with a security question and answer, and you have not used the Oyster card to travel in the last 8 weeks you will be unable to create an account. Please make a journey using the Oyster card and try again the next day.
- For further assistance, please contact TfL Customer Services on 0343 222 1234 (8am - 8pm).

All fields marked with an asterisk (*) are mandatory

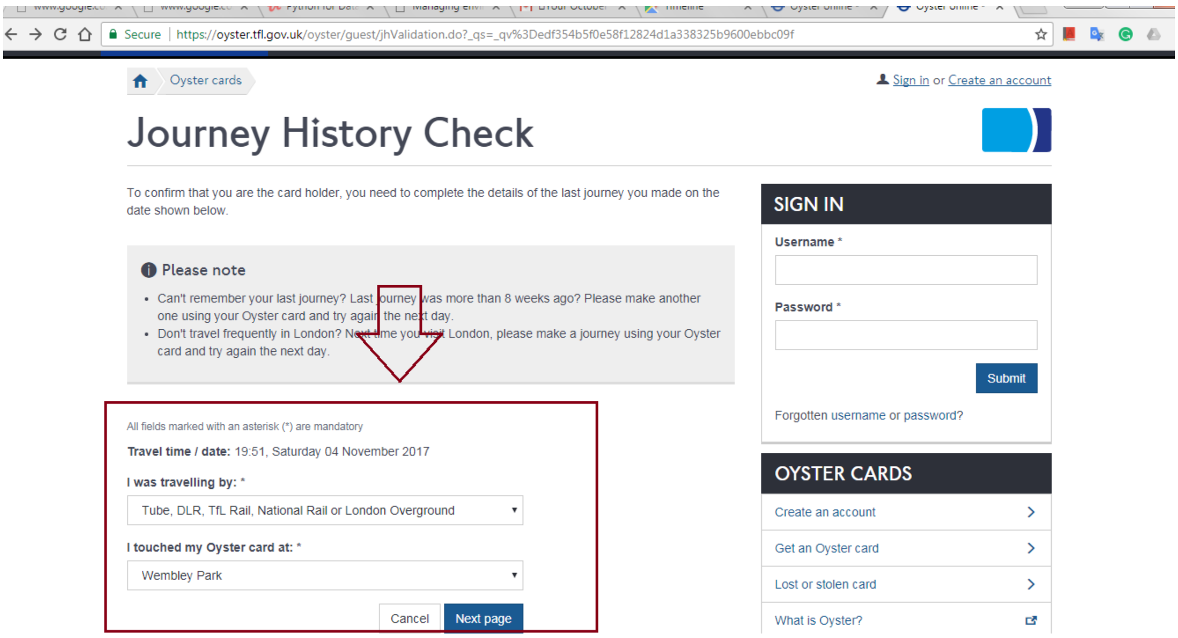
Card details

Card number *

Find your Oyster card number

Please note

You cannot buy discounted or child-rate season tickets from Oyster online.



Secure | https://oyster.tfl.gov.uk/oyster/guest/jhValidation.do?_qs=_qv%3Dedf354b5f0e58f12824d1a338325b9600ebbc09f

Oyster cards

Sign in or Create an account

Journey History Check

To confirm that you are the card holder, you need to complete the details of the last journey you made on the date shown below.

Please note

- Can't remember your last journey? Last journey was more than 8 weeks ago? Please make another one using your Oyster card and try again the next day.
- Don't travel frequently in London? Next time you visit London, please make a journey using your Oyster card and try again the next day.

All fields marked with an asterisk (*) are mandatory

Travel time / date: 19:51, Saturday 04 November 2017

I was travelling by: *

Tube, DLR, TfL Rail, National Rail or London Overground ▾

I touched my Oyster card at: *

Wembley Park ▾

Cancel | Next page

SIGN IN

Username *

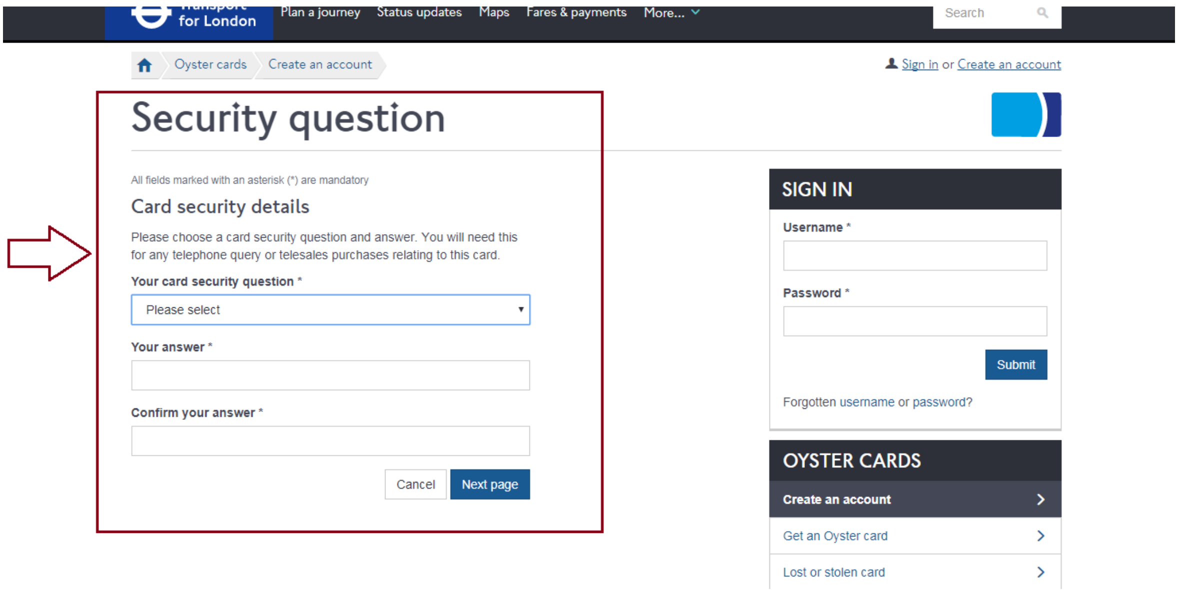
Password *

Submit

Forgotten username or password?

OYSTER CARDS

- Create an account
- Get an Oyster card
- Lost or stolen card
- What is Oyster?



Oyster cards | Create an account

Sign in or Create an account

Security question

All fields marked with an asterisk (*) are mandatory

Card security details

Please choose a card security question and answer. You will need this for any telephone query or telesales purchases relating to this card.

Your card security question *

Please select ▾

Your answer *

Confirm your answer *

Cancel | Next page

SIGN IN

Username *

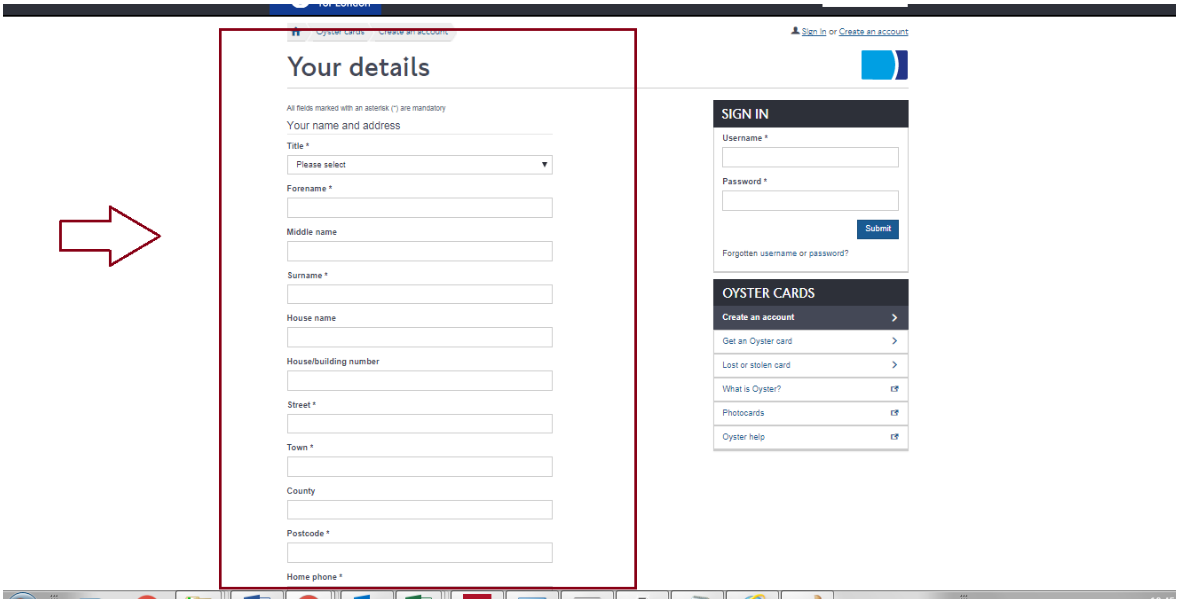
Password *

Submit

Forgotten username or password?

OYSTER CARDS

- **Create an account**
- Get an Oyster card
- Lost or stolen card



Oyster cards › Create an account

Sign in or Create an account

Your details

All fields marked with an asterisk (*) are mandatory

Your name and address

Title *

Please select

Forename *

Middle name

Surname *

House name

House/building number

Street *

Town *

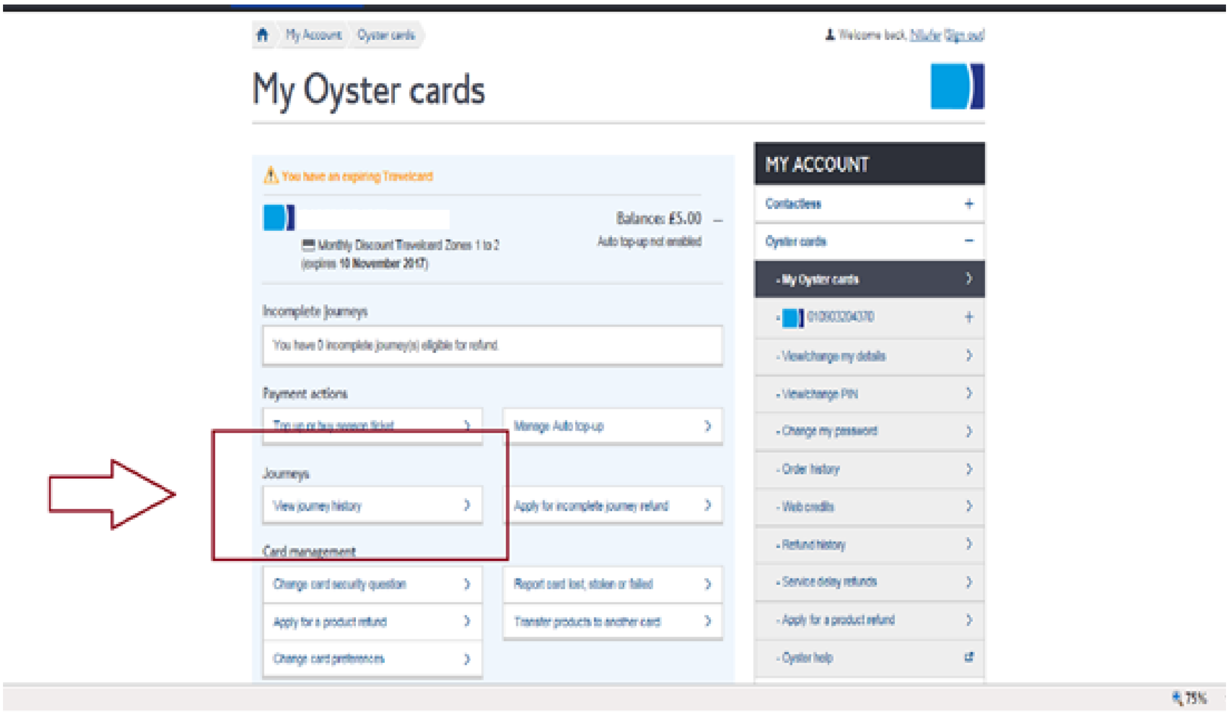
County

Postcode *

Home phone *

Over here, the column called <u>House name</u> and <u>House number</u> may create some problems. Please follow the format below.

House name:	Avenue Mansions
House number:	4A



My Account › Oyster cards

Welcome back, Nilufar Sign out

My Oyster cards

You have an expiring Travelcard

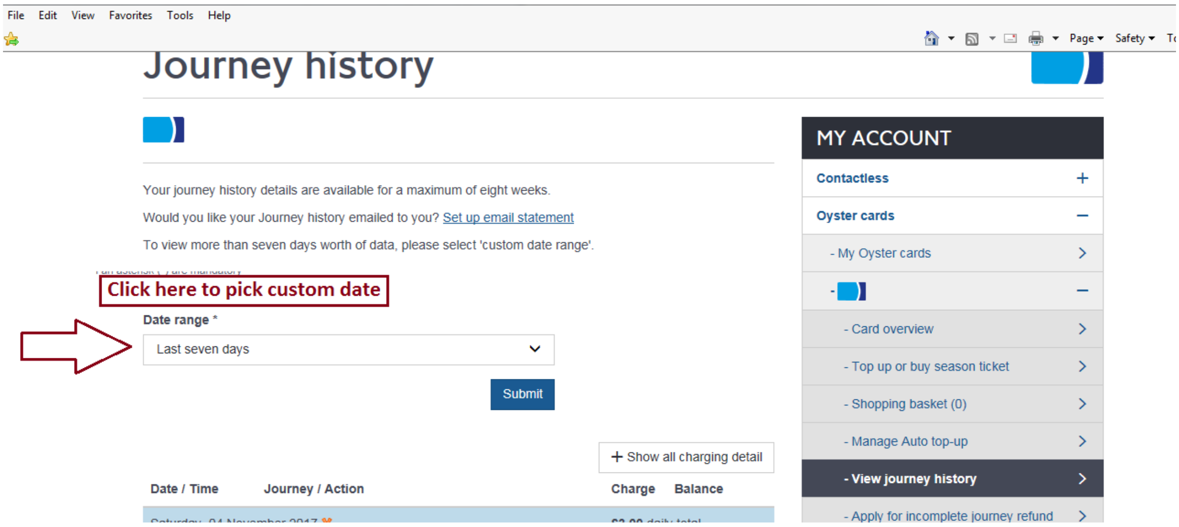
Balance: £5.00

Auto top-up not enabled

Monthly Discount Travelcard Zones 1 to 2 (expires 10 November 2017)

Incomplete Journeys

You have 0 incomplete journey(s) eligible for refund.



Journey history

Your journey history details are available for a maximum of eight weeks.

Would you like your Journey history emailed to you? Set up email statement

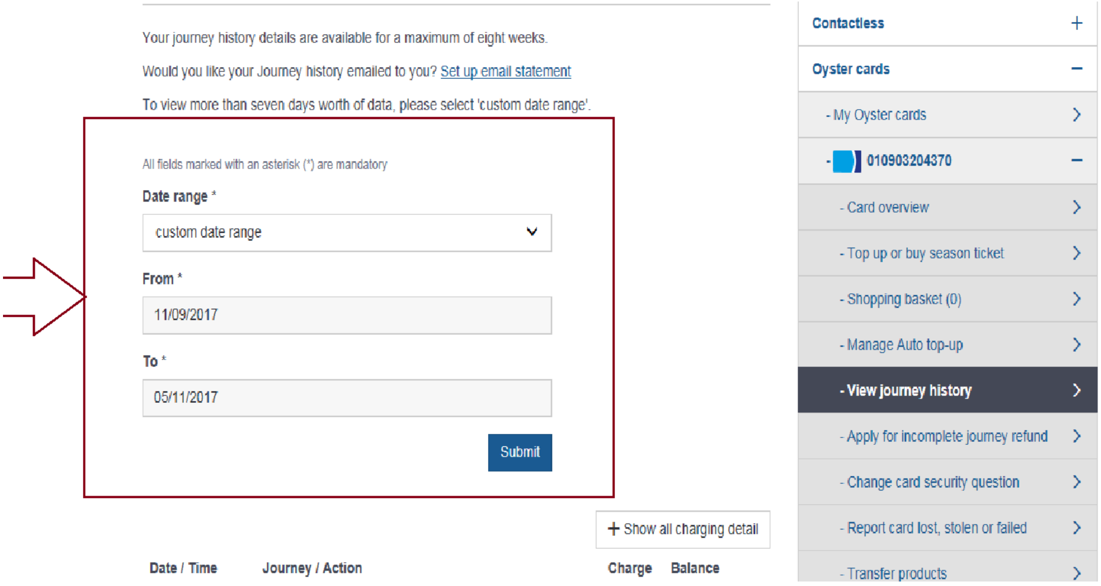
To view more than seven days worth of data, please select 'custom date range'.

Click here to pick custom date

Date range *

Last seven days

Date / Time	Journey / Action	Charge	Balance
Saturday, 04 November 2017		£3.00 daily total	


Contactless
Oyster cards
- My Oyster cards
- 010903204370
- Card overview
- Top up or buy season ticket
- Shopping basket (0)
- Manage Auto top-up
- View journey history
- Apply for incomplete journey refund
- Change card security question
- Report card lost, stolen or failed
- Transfer products
Submit
+ Show all charging detail

Your journey history details are available for a maximum of eight weeks.

Would you like your Journey history emailed to you? Set up email statement

To view more than seven days worth of data, please select 'custom date range'.

All fields marked with an asterisk (*) are mandatory

Date range *

custom date range

From *

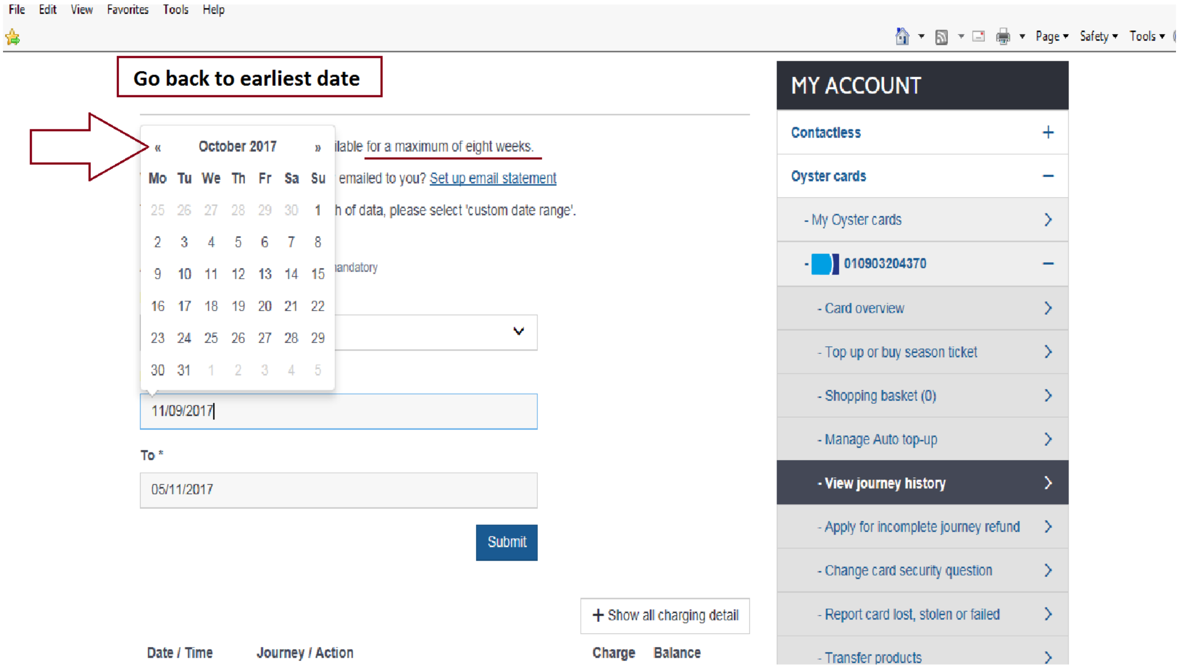
11/09/2017

To *

05/11/2017

Date / Time	Journey / Action	Charge	Balance

Go back to earliest date



«		October 2017				»
Mo	Tu	We	Th	Fr	Sa	Su
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

ilable for a maximum of eight weeks.

emailed to you? Set up email statement

h of data, please select 'custom date range'.

andatory

11/09/2017

To *

05/11/2017

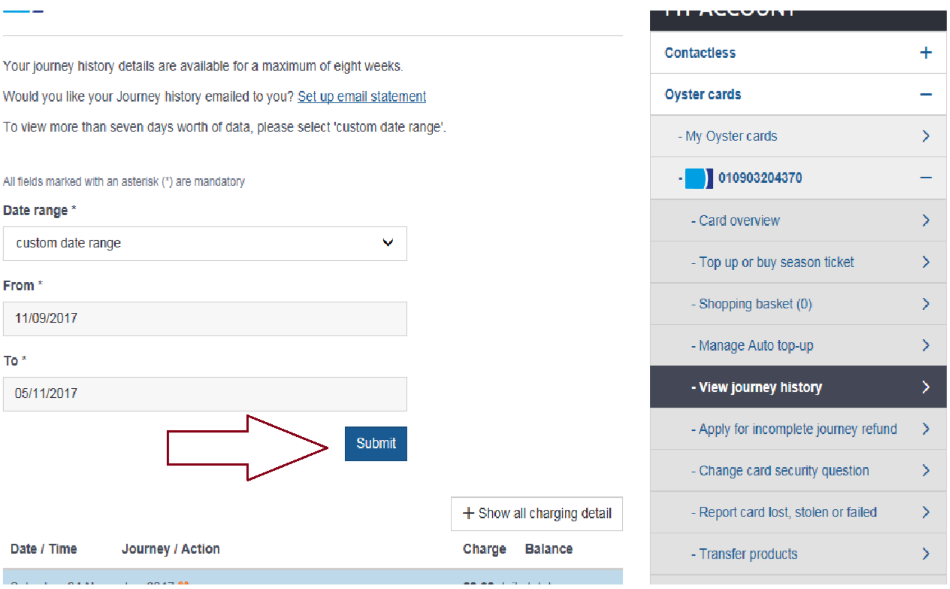
Submit

+ Show all charging detail

Date / Time Journey / Action Charge Balance

MY ACCOUNT

- Contactless +
- Oyster cards –
- - My Oyster cards
- - 010903204370 –
- - Card overview
- - Top up or buy season ticket
- - Shopping basket (0)
- - Manage Auto top-up
- - View journey history
- - Apply for incomplete journey refund
- - Change card security question
- - Report card lost, stolen or failed
- - Transfer products



Your journey history details are available for a maximum of eight weeks.

Would you like your Journey history emailed to you? Set up email statement

To view more than seven days worth of data, please select 'custom date range'.

All fields marked with an asterisk (*) are mandatory

Date range *

custom date range

From *

11/09/2017

To *

05/11/2017

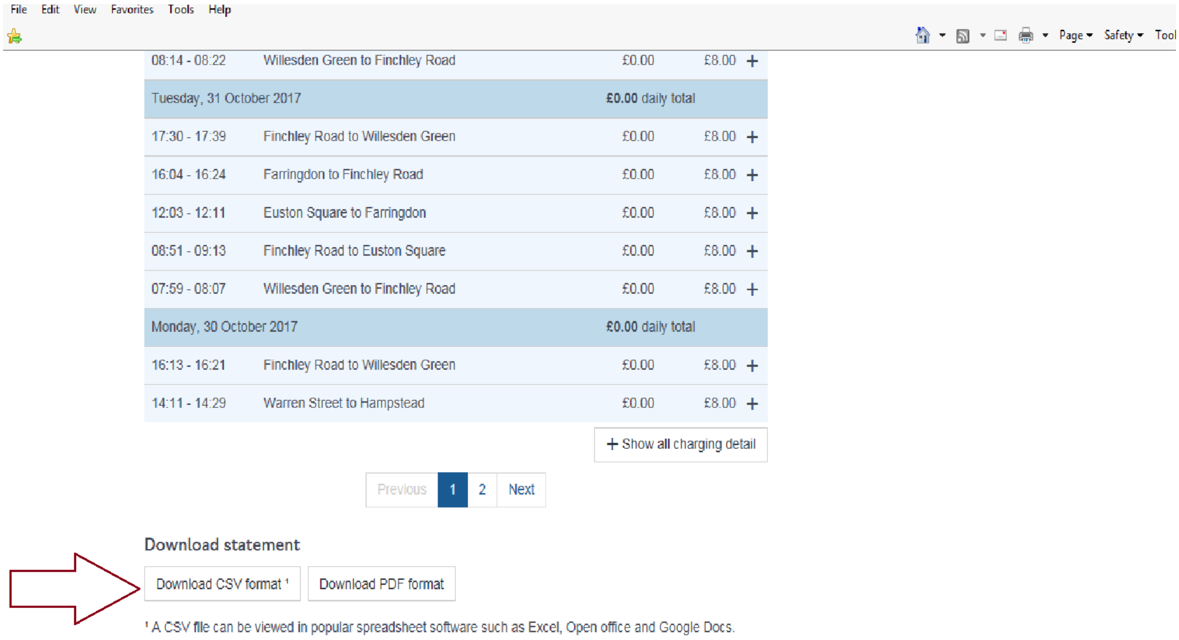
Submit

+ Show all charging detail

Date / Time Journey / Action Charge Balance

MY ACCOUNT

- Contactless +
- Oyster cards –
- - My Oyster cards
- - 010903204370 –
- - Card overview
- - Top up or buy season ticket
- - Shopping basket (0)
- - Manage Auto top-up
- - View journey history
- - Apply for incomplete journey refund
- - Change card security question
- - Report card lost, stolen or failed
- - Transfer products



08:14 - 08:22	Willesden Green to Finchley Road	£0.00	£8.00	+
Tuesday, 31 October 2017		**£0.00** daily total		
17:30 - 17:39	Finchley Road to Willesden Green	£0.00	£8.00	+
16:04 - 16:24	Farringdon to Finchley Road	£0.00	£8.00	+
12:03 - 12:11	Euston Square to Farringdon	£0.00	£8.00	+
08:51 - 09:13	Finchley Road to Euston Square	£0.00	£8.00	+
07:59 - 08:07	Willesden Green to Finchley Road	£0.00	£8.00	+
Monday, 30 October 2017		**£0.00** daily total		
16:13 - 16:21	Finchley Road to Willesden Green	£0.00	£8.00	+
14:11 - 14:29	Warren Street to Hampstead	£0.00	£8.00	+

+ Show all charging detail

Previous 1 2 Next

Download statement

Download CSV format [1] Download PDF format

[1] A CSV file can be viewed in popular spreadsheet software such as Excel, Open office and Google Docs.

Please save the .csv file in the same format as below.

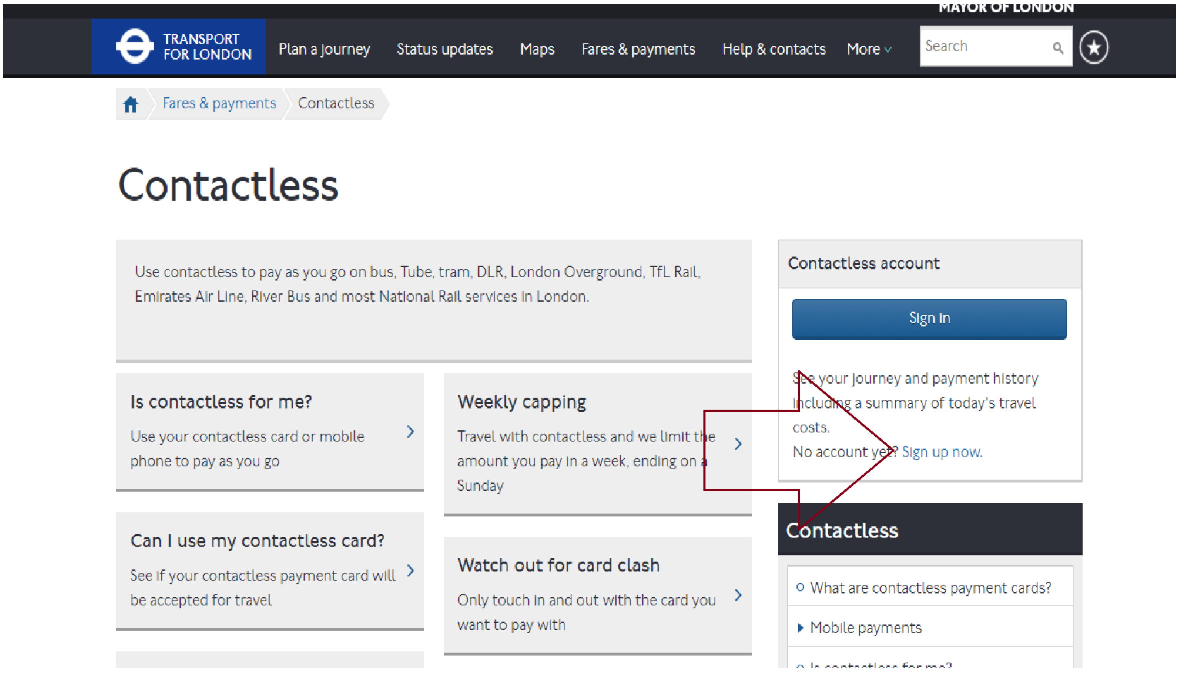
Nilufer_Aslam_11Sep_04Nov.csv

Template 2: How can we register and download our contactless card data?

1. Go to the website below and please create an account.

https://tfl.gov.uk/fares-and-payments/contactless

2. Please fill the sections pointed out by arrow in the following figures
3. If you have registered your card before, please see 4.



Contactless

Use contactless to pay as you go on bus, Tube, tram, DLR, London Overground, TfL Rail, Emirates Air Line, River Bus and most National Rail services in London.

Is contactless for me?

Use your contactless card or mobile phone to pay as you go

Weekly capping

Travel with contactless and we limit the amount you pay in a week, ending on a Sunday

Can I use my contactless card?

See if your contactless payment card will be accepted for travel

Watch out for card clash

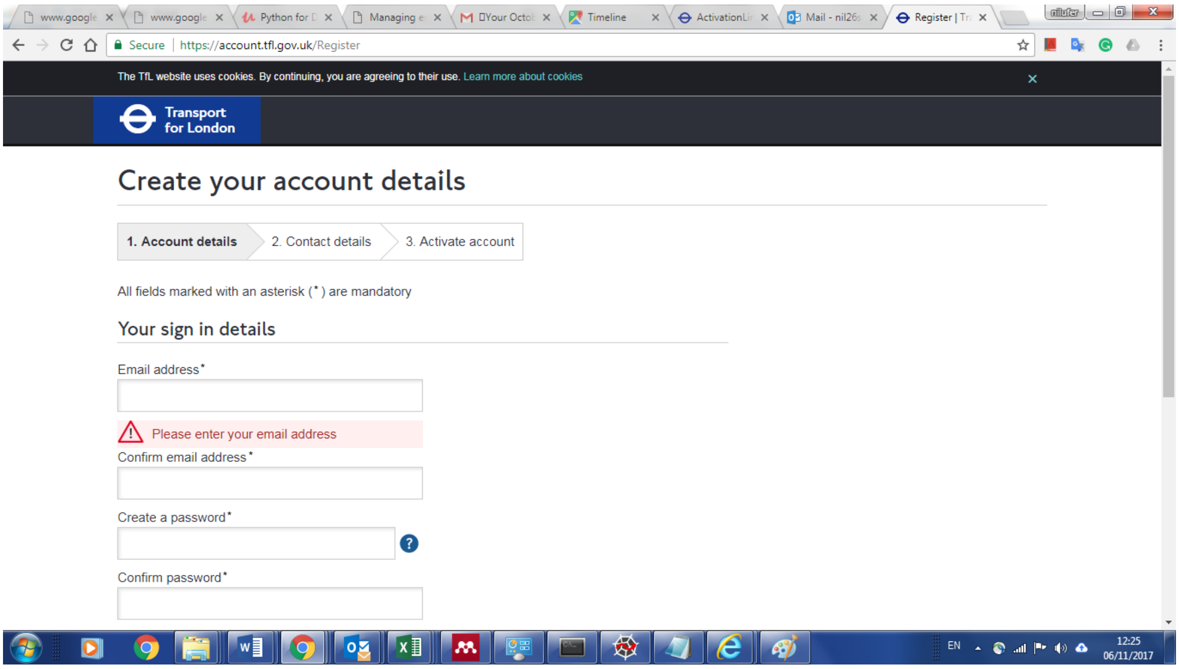
Only touch in and out with the card you want to pay with

Contactless account

See your journey and payment history including a summary of today's travel costs.

No account yet? Sign up now.

Contactless



The TfL website uses cookies. By continuing, you are agreeing to their use. Learn more about cookies

Create your account details

1. Account details 2. Contact details 3. Activate account

All fields marked with an asterisk (*) are mandatory

Your sign in details

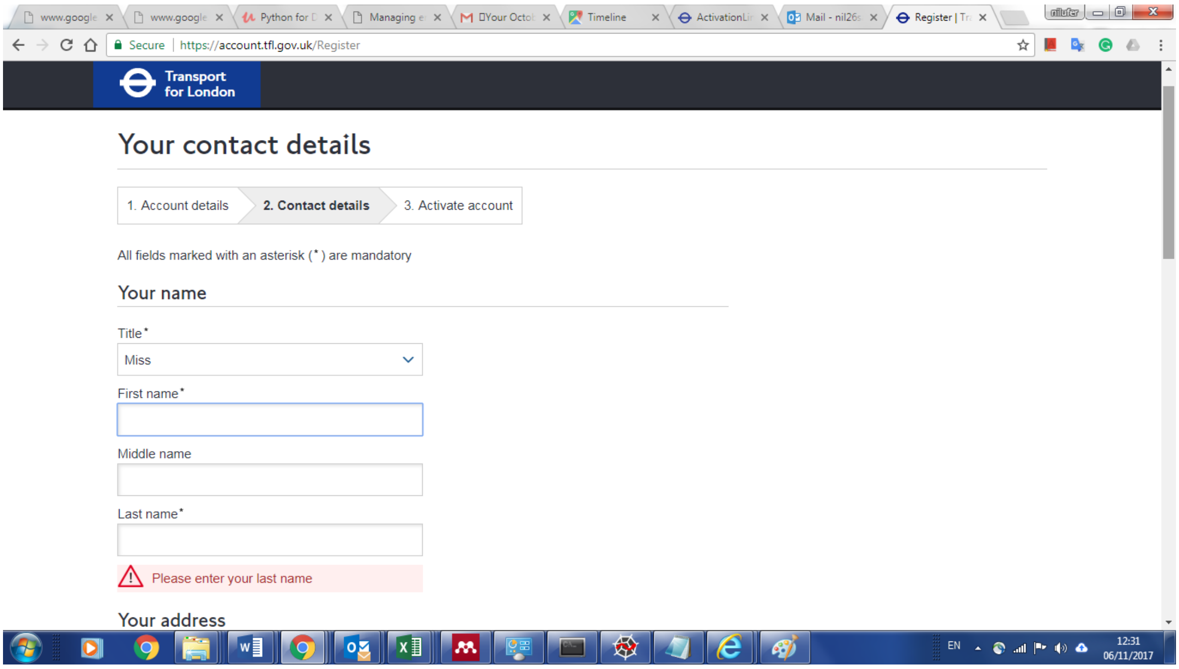
Email address *

Please enter your email address

Confirm email address *

Create a password *

Confirm password *



Your contact details

1. Account details 2. Contact details 3. Activate account

All fields marked with an asterisk (*) are mandatory

Your name

Title *

Miss

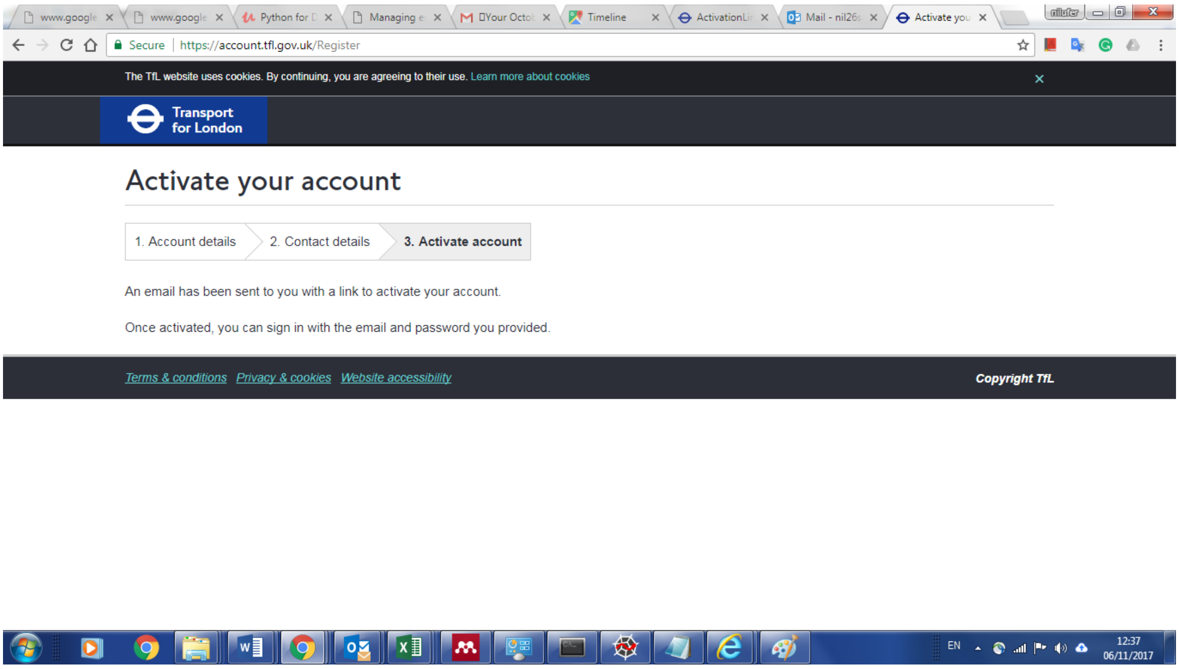
First name *

Middle name

Last name *

Please enter your last name

Your address



The TfL website uses cookies. By continuing, you are agreeing to their use. Learn more about cookies

Transport for London

Activate your account

1. Account details 2. Contact details **3. Activate account**

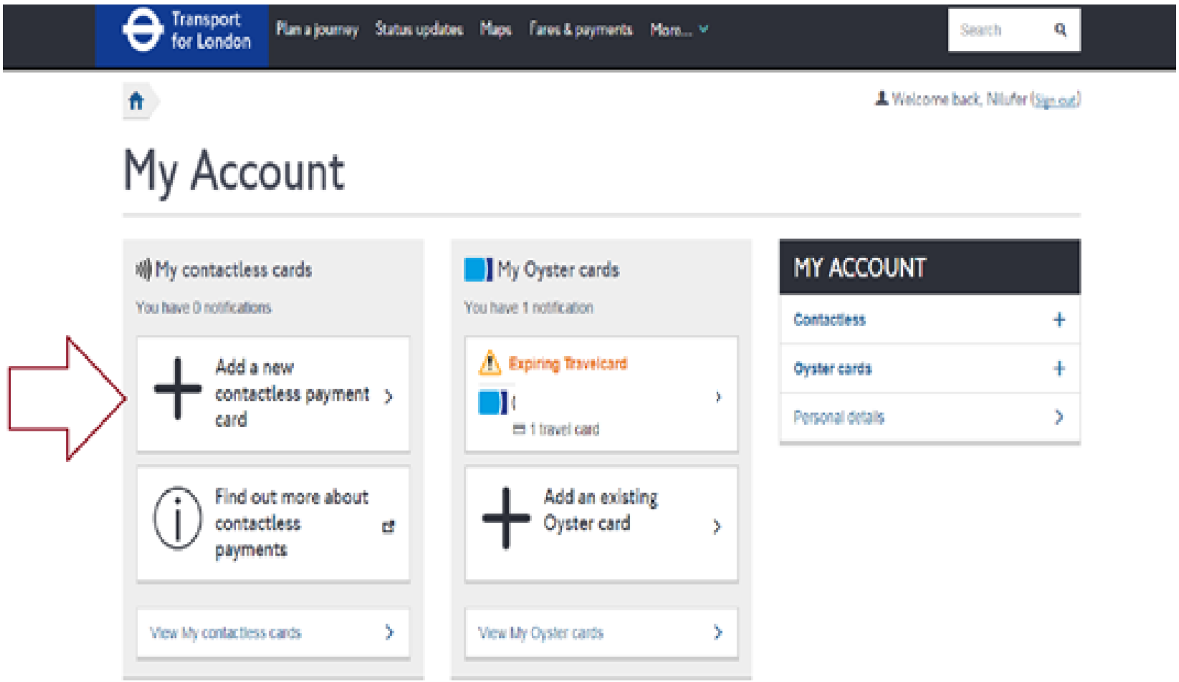
An email has been sent to you with a link to activate your account.

Once activated, you can sign in with the email and password you provided.

Terms & conditions Privacy & cookies Website accessibility

Copyright TfL

4. If you have registered before, you only need to sign in as below.



Welcome back, Nilufer (Sign out)

My Account

My contactless cards

You have 0 notifications

Add a new contactless payment card

Find out more about contactless payments

View My contactless cards

My Oyster cards

You have 1 notification

Expiring Travelcard

1 travel card

Add an existing Oyster card

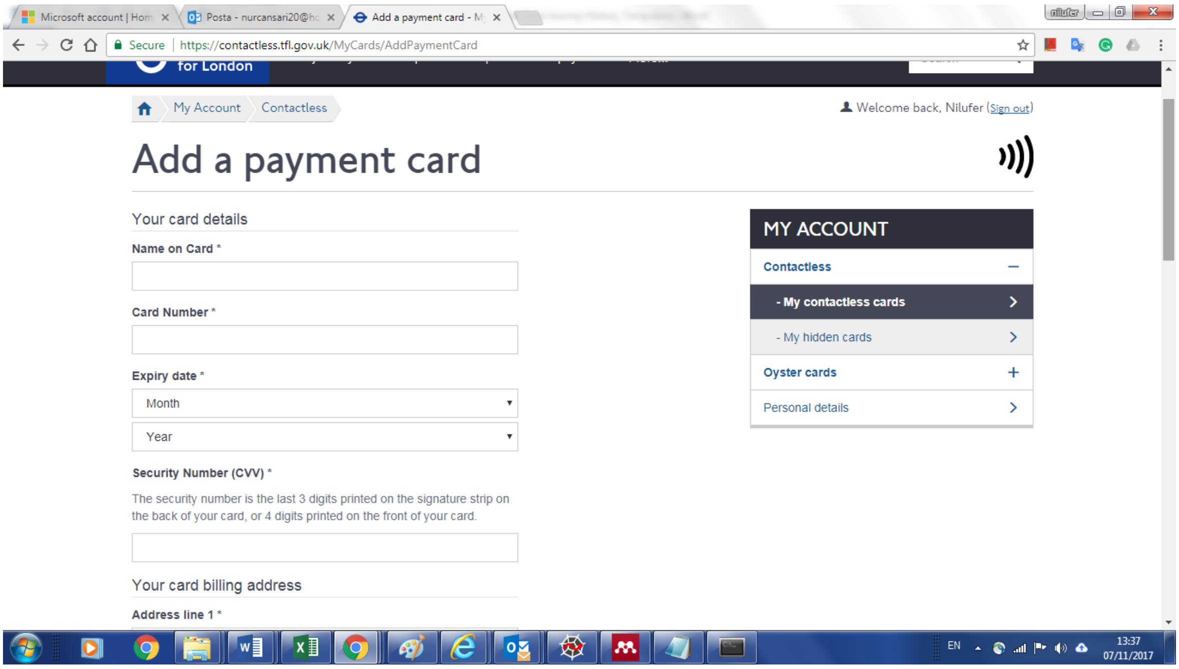
View My Oyster cards

MY ACCOUNT

Contactless +

Oyster cards +

Personal details



My Account › Contactless

Welcome back, Nilufer (Sign out)

Add a payment card

Your card details

Name on Card *

Card Number *

Expiry date *

Month ▾

Year ▾

Security Number (CVV) *

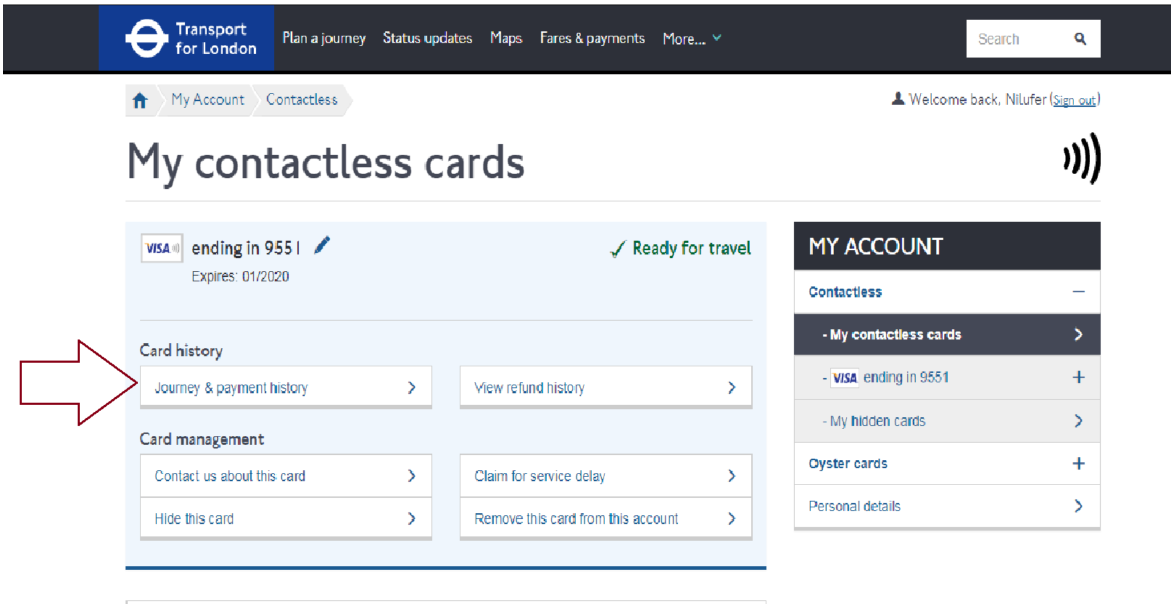
The security number is the last 3 digits printed on the signature strip on the back of your card, or 4 digits printed on the front of your card.

Your card billing address

Address line 1 *

MY ACCOUNT

- Contactless –
- - My contactless cards >
- - My hidden cards >
- Oyster cards +
- Personal details >



My Account › Contactless

Welcome back, Nilufer (Sign out)

My contactless cards

VISA ending in 9551

Expires: 01/2020

✓ Ready for travel

Card history

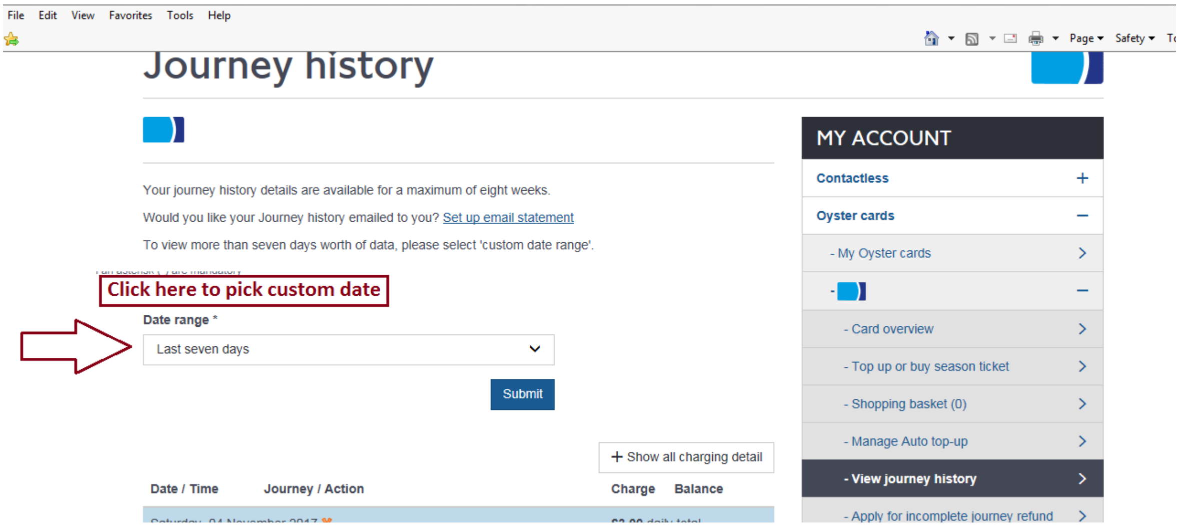
- Journey & payment history >
- View refund history >

Card management

- Contact us about this card >
- Claim for service delay >
- Hide this card >
- Remove this card from this account >

MY ACCOUNT

- Contactless –
- - My contactless cards >
- - VISA ending in 9551 +
- - My hidden cards >
- Oyster cards +
- Personal details >



Journey history

Your journey history details are available for a maximum of eight weeks.

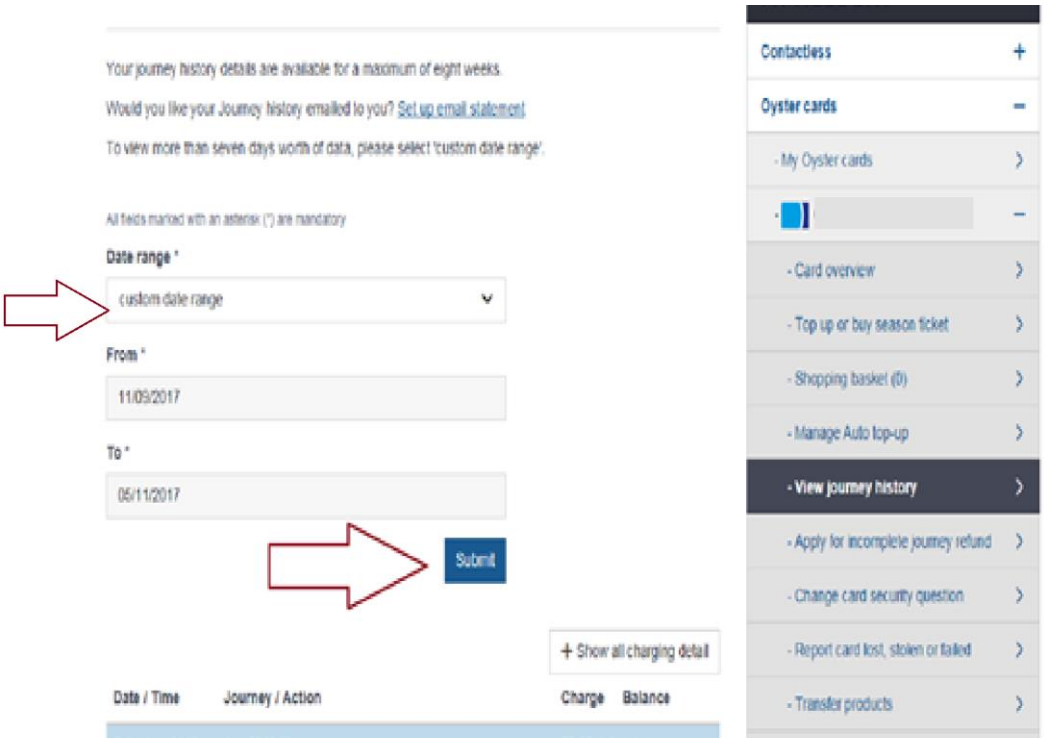
Would you like your Journey history emailed to you? Set up email statement

To view more than seven days worth of data, please select 'custom date range'.

Click here to pick custom date

Date range *

Last seven days



Your journey history details are available for a maximum of eight weeks.

Would you like your Journey history emailed to you? Set up email statement

To view more than seven days worth of data, please select 'custom date range'.

All fields marked with an asterisk (*) are mandatory

Date range *

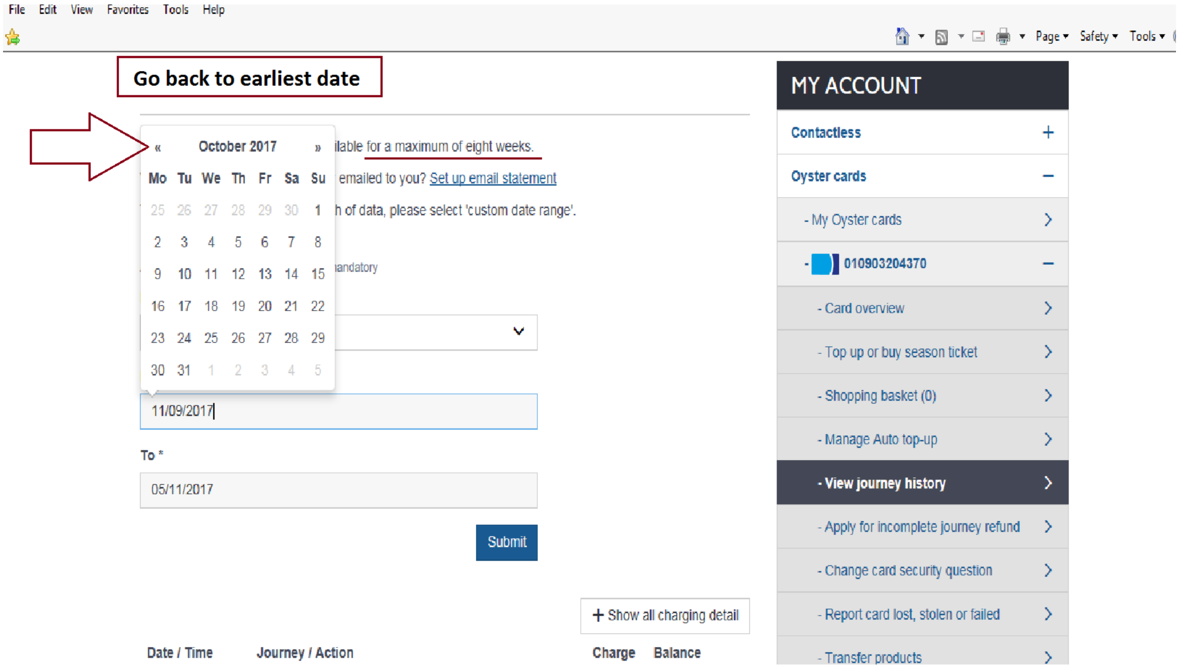
custom date range

From *

11/09/2017

To *

05/11/2017



Go back to earliest date

« October 2017 »

Mo	Tu	We	Th	Fr	Sa	Su
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

ilable for a maximum of eight weeks.

emailed to you? Set up email statement

h of data, please select 'custom date range'.

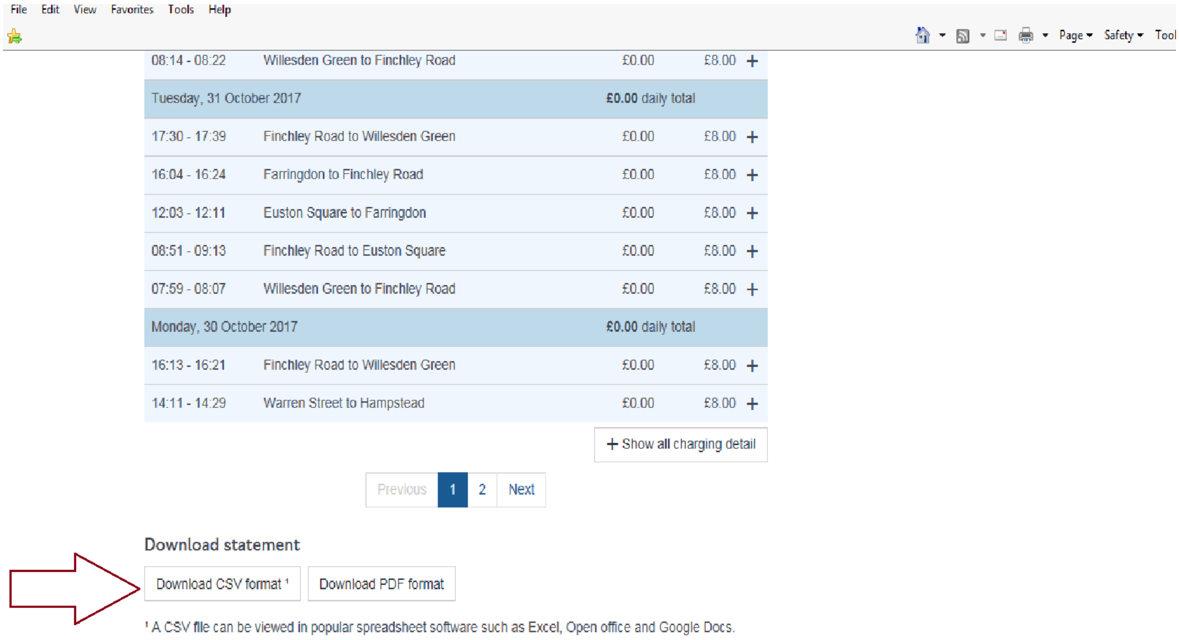
andatory

11/09/2017

To *

05/11/2017

Date / Time	Journey / Action	Charge	Balance


File Edit View Favorites Tools Help
Page ▾ Safety ▾ Tool
+ Show all charging detail
Previous 1 2 Next

08:14 - 08:22	Willesden Green to Finchley Road	£0.00	£8.00 +
Tuesday, 31 October 2017		**£0.00** daily total	
17:30 - 17:39	Finchley Road to Willesden Green	£0.00	£8.00 +
16:04 - 16:24	Farringdon to Finchley Road	£0.00	£8.00 +
12:03 - 12:11	Euston Square to Farringdon	£0.00	£8.00 +
08:51 - 09:13	Finchley Road to Euston Square	£0.00	£8.00 +
07:59 - 08:07	Willesden Green to Finchley Road	£0.00	£8.00 +
Monday, 30 October 2017		**£0.00** daily total	
16:13 - 16:21	Finchley Road to Willesden Green	£0.00	£8.00 +
14:11 - 14:29	Warren Street to Hampstead	£0.00	£8.00 +

Download statement

Download CSV format ¹ Download PDF format

¹ A CSV file can be viewed in popular spreadsheet software such as Excel, Open office and Google Docs.

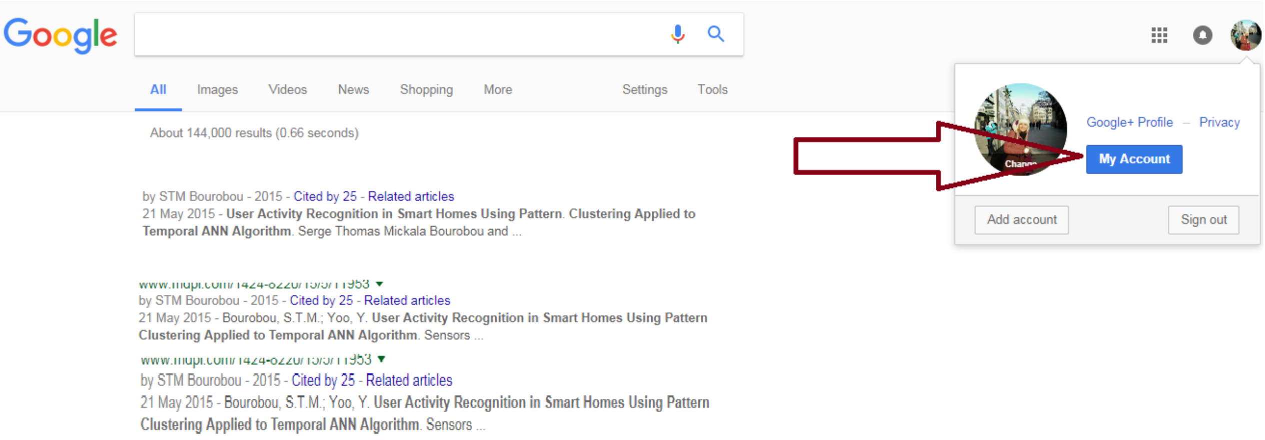
Please save the .csv file in the same format as below.

Nilufer_Aslam_11Sep_04Nov.csv

Template 3: How can we download our Google activity data?

To download your activity data, you need to have a Google account. If you have one and are willing to share, please follow the steps below.

1. Please click your Google account.



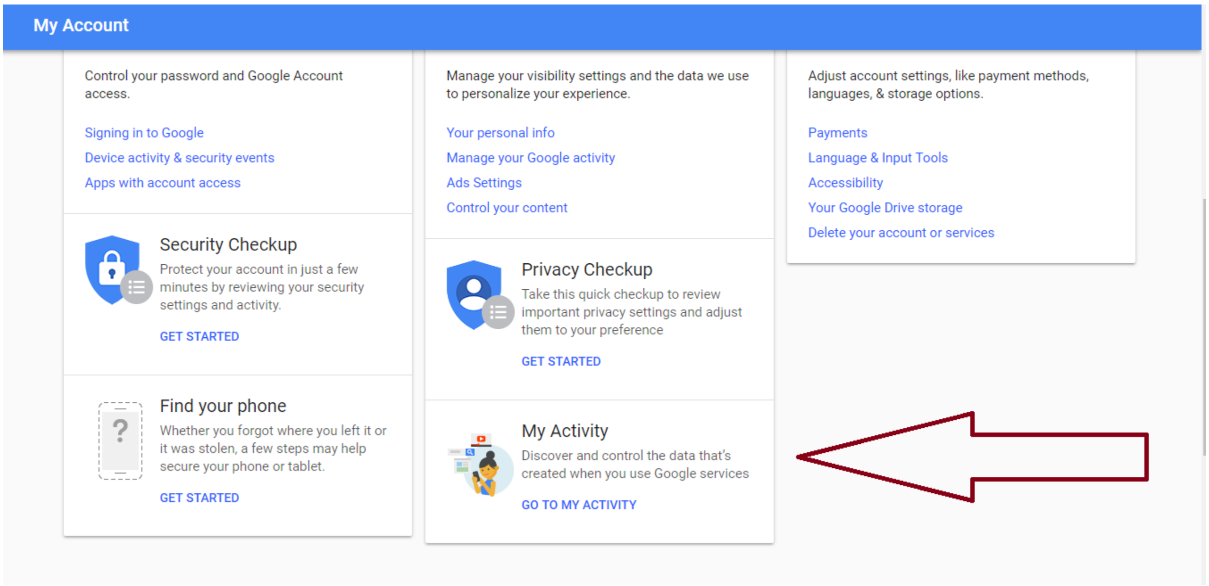
About 144,000 results (0.66 seconds)

by STM Bourobou - 2015 - Cited by 25 - Related articles
21 May 2015 - **User Activity Recognition in Smart Homes Using Pattern. Clustering Applied to Temporal ANN Algorithm**. Serge Thomas Mickala Bourobou and ...

www.mdpi.com/1424-8220/15/5/11953 ▾
by STM Bourobou - 2015 - Cited by 25 - Related articles
21 May 2015 - Bourobou, S.T.M.; Yoo, Y. **User Activity Recognition in Smart Homes Using Pattern Clustering Applied to Temporal ANN Algorithm**. Sensors ...

www.mdpi.com/1424-8220/15/5/11953 ▾
by STM Bourobou - 2015 - Cited by 25 - Related articles
21 May 2015 - Bourobou, S.T.M.; Yoo, Y. **User Activity Recognition in Smart Homes Using Pattern Clustering Applied to Temporal ANN Algorithm**. Sensors ...

2. GO TO MY ACTIVITY



My Account

Control your password and Google Account access.

Signing in to Google
Device activity & security events
Apps with account access

Security Checkup
Protect your account in just a few minutes by reviewing your security settings and activity.
GET STARTED

Find your phone
Whether you forgot where you left it or it was stolen, a few steps may help secure your phone or tablet.
GET STARTED

Manage your visibility settings and the data we use to personalize your experience.

Your personal info
Manage your Google activity
Ads Settings
Control your content

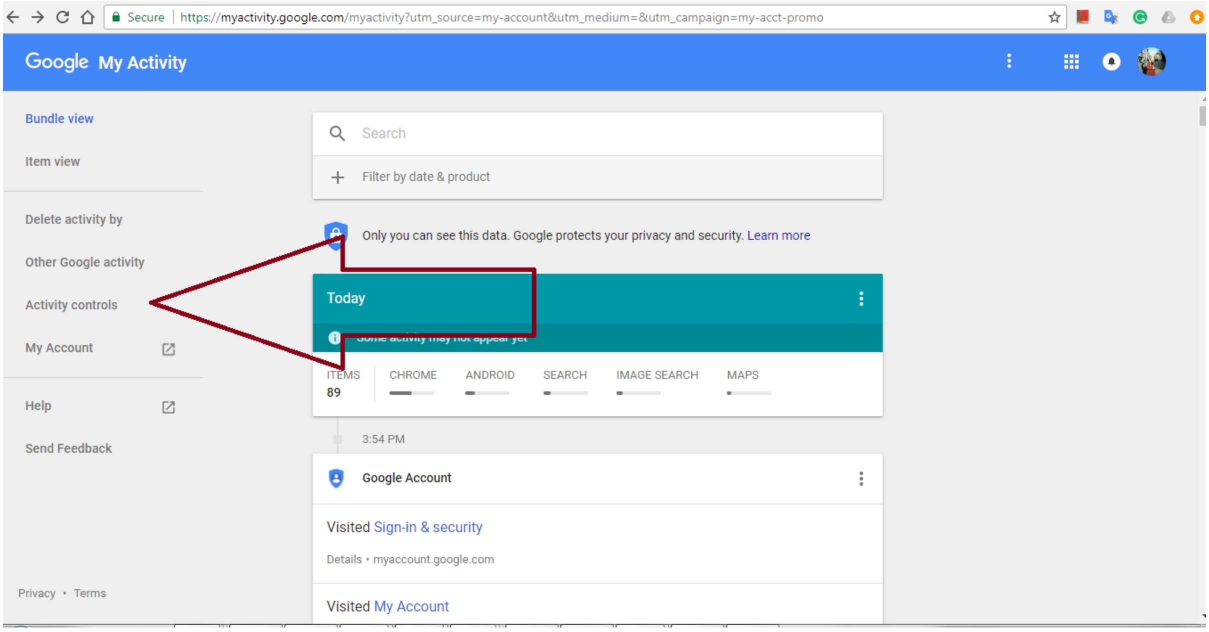
Privacy Checkup
Take this quick checkup to review important privacy settings and adjust them to your preference
GET STARTED

My Activity
Discover and control the data that's created when you use Google services
GO TO MY ACTIVITY

Adjust account settings, like payment methods, languages, & storage options.

Payments
Language & Input Tools
Accessibility
Your Google Drive storage
Delete your account or services

3. Click activity controls.



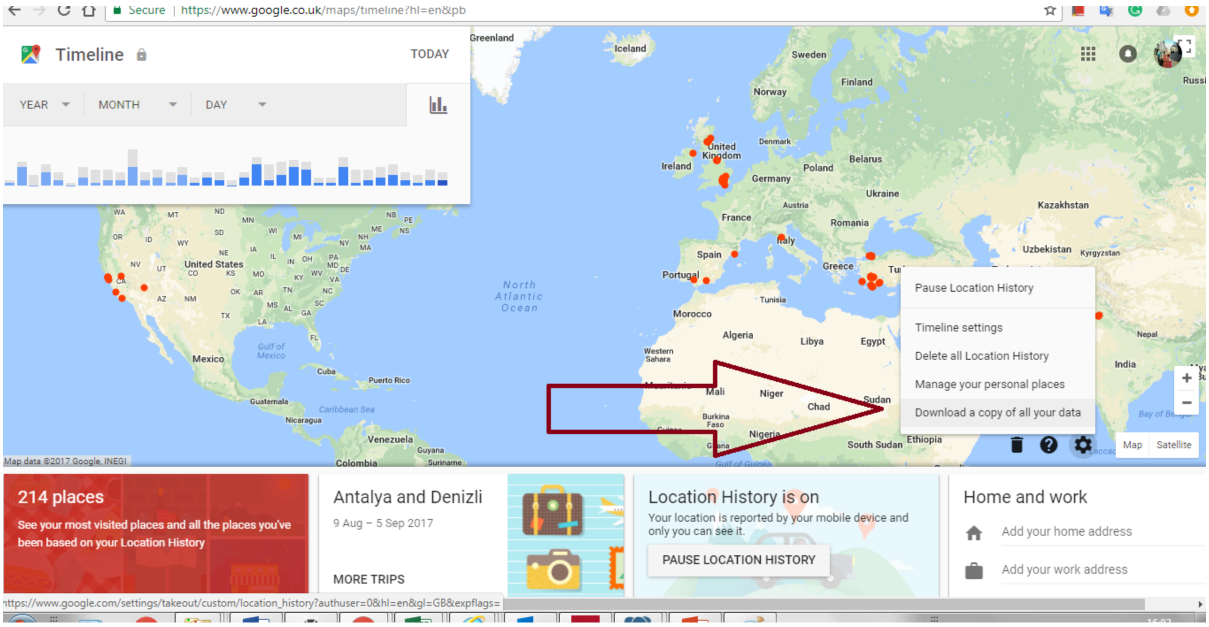
Only you can see this data. Google protects your privacy and security. Learn more

Visited Sign-in & security

Details • myaccount.google.com

Visited My Account

4. Download a copy of all your data.


Secure | https://www.google.co.uk/maps/timeline?hl=en&pb
Timeline
TODAY
YEAR
MONTH
DAY
Pause Location History
Timeline settings
Delete all Location History
Manage your personal places
Download a copy of all your data
Map
Satellite
Map data ©2017 Google, INEGI
214 places

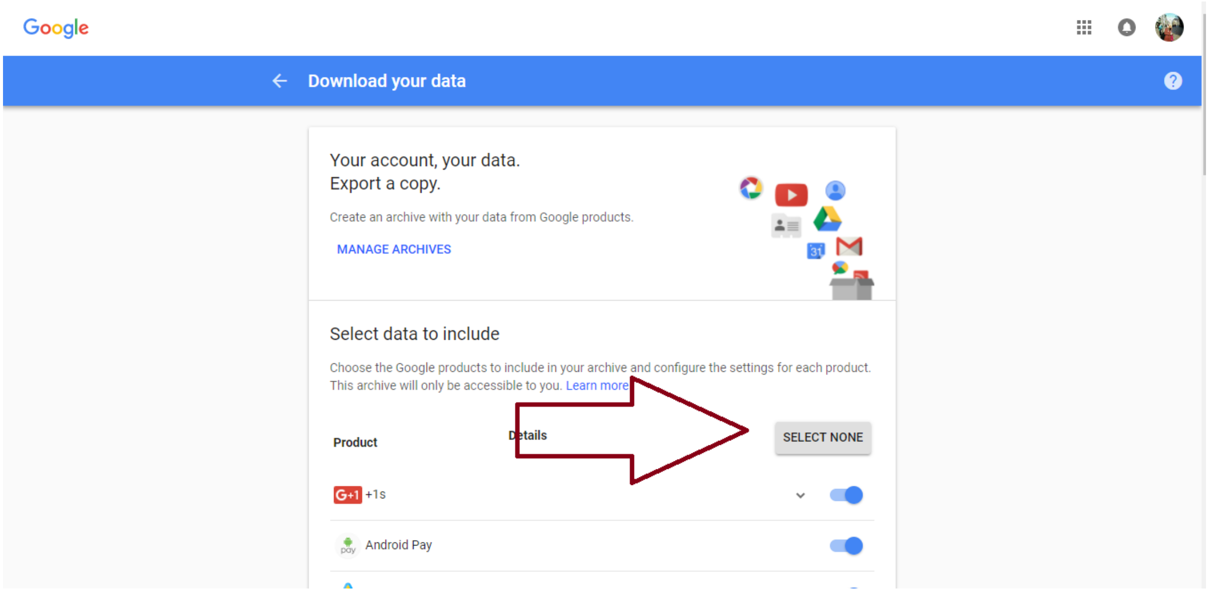
See your most visited places and all the places you've been based on your Location History

Antalya and Denizli
9 Aug – 5 Sep 2017
MORE TRIPS
Location History is on

Your location is reported by your mobile device and only you can see it.

PAUSE LOCATION HISTORY
Home and work
Add your home address
Add your work address
https://www.google.com/settings/takeout/custom/location_history?authuser=0&hl=en&gl=GB&expflags=

5. Scroll down the page called ‘Download your data’. Please select none first.



Your account, your data.
Export a copy.

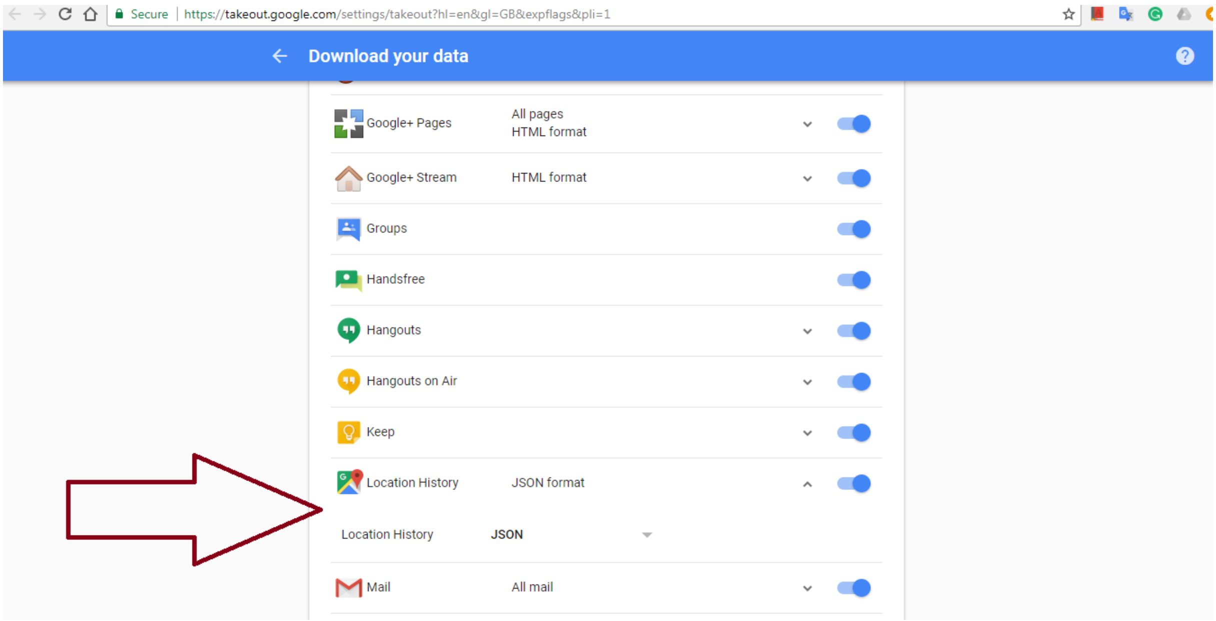
Create an archive with your data from Google products.

Select data to include

Choose the Google products to include in your archive and configure the settings for each product. This archive will only be accessible to you. Learn more

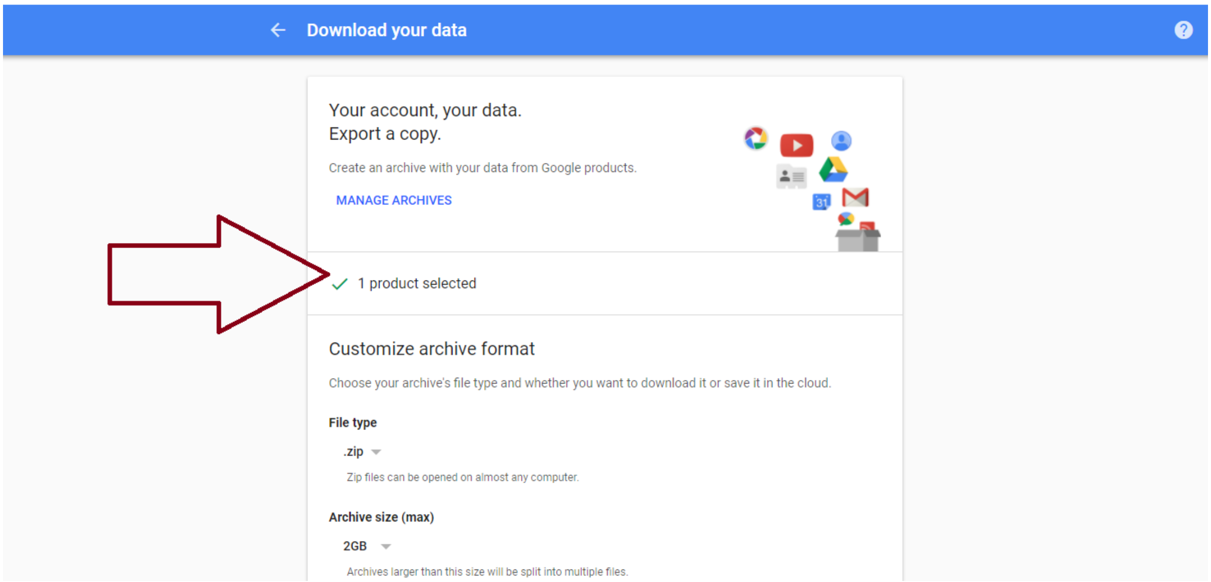
Product	Details
+1s	
Android Pay	

6. And then click location history and click next at the end of the page.


Secure
https://takeout.google.com/settings/takeout?hl=en&gl=GB&expflags&pli=1
Download your data

Google+ Pages	All pages HTML format
Google+ Stream	HTML format
Groups	
Handsfree	
Hangouts	
Hangouts on Air	
Keep	
Location History	JSON format
Location History	JSON
Mail	All mail

7. One product is selected and at the end of the page, click create an archive. Then see the page below.



Your account, your data.
Export a copy.

Create an archive with your data from Google products.

✓ 1 product selected

Customize archive format

Choose your archive's file type and whether you want to download it or save it in the cloud.

File type

.zip

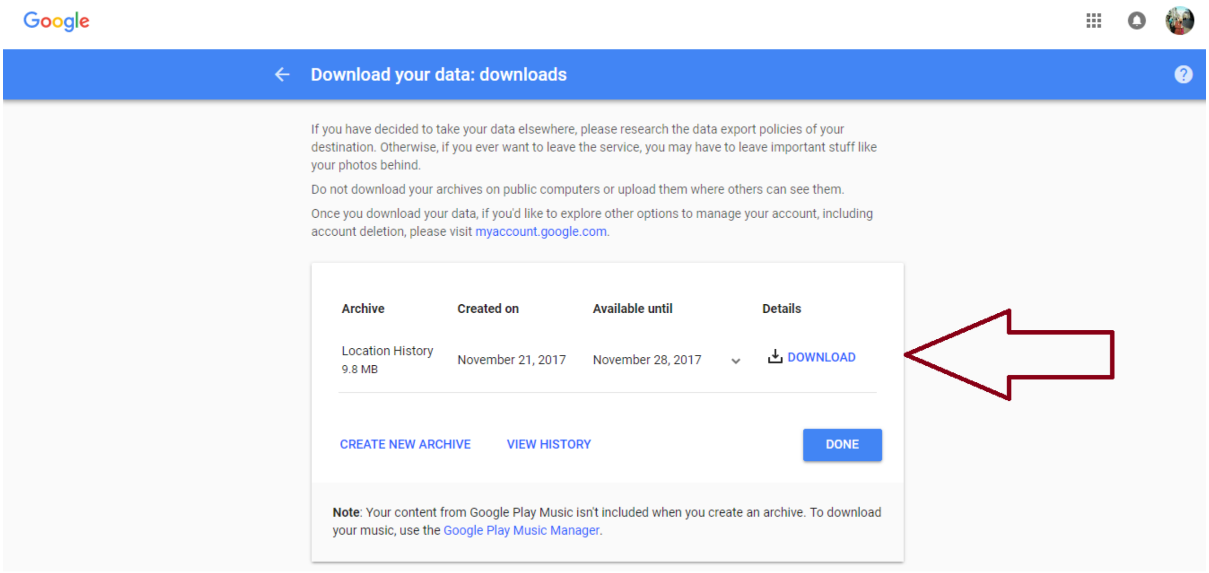
Zip files can be opened on almost any computer.

Archive size (max)

2GB

Archives larger than this size will be split into multiple files.

8. Your data is being prepared. Once it is completed, please download your data in JSON format



If you have decided to take your data elsewhere, please research the data export policies of your destination. Otherwise, if you ever want to leave the service, you may have to leave important stuff like your photos behind.

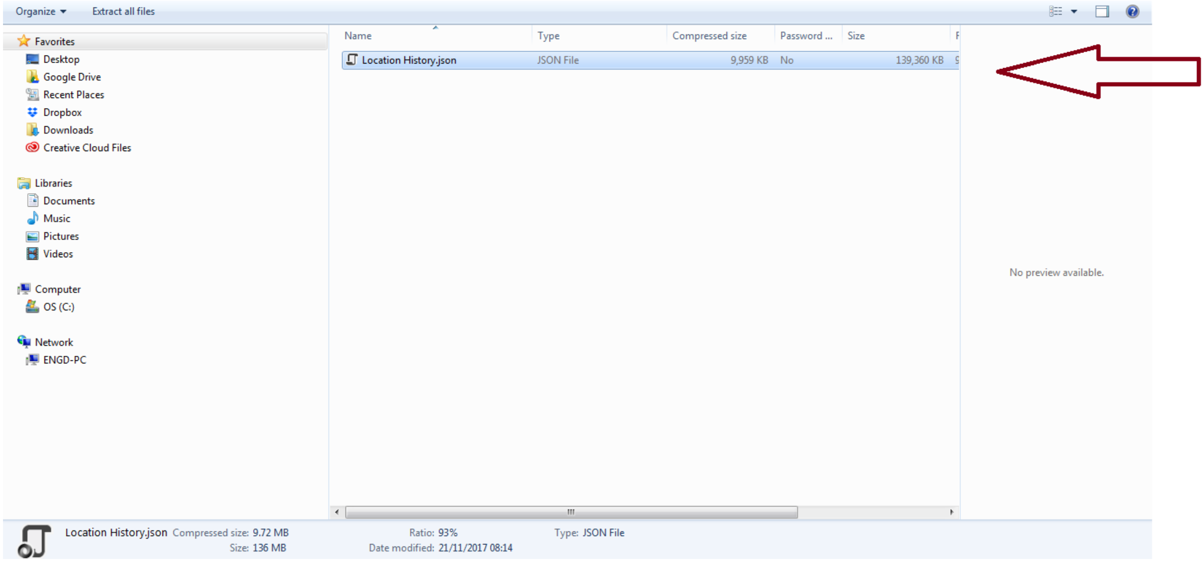
Do not download your archives on public computers or upload them where others can see them.

Once you download your data, if you'd like to explore other options to manage your account, including account deletion, please visit myaccount.google.com.

Archive	Created on	Available until	Details
Location History 9.8 MB	November 21, 2017	November 28, 2017	DOWNLOAD

Note: Your content from Google Play Music isn't included when you create an archive. To download your music, use the Google Play Music Manager.

9. Once again, sign in to your account. The zip file will take place in the download folder.



Name	Type	Compressed size	Password ...	Size
Location History.json	JSON File	9,959 KB	No	139,360 KB

Location History.json Compressed size: 9.72 MB Size: 136 MB Ratio: 93% Date modified: 21/11/2017 08:14 Type: JSON File

10. Extract the .zip file and add/attach only the file called Local History.json

Questionnaire

The answers to these questions are requested from volunteers during the data collection process.

Questions	Answers
What type of Smart Card Data do you use? (e.g., Oyster card data or contactless card data.	Oyster card data Contactless card data
What is your gender?	Female Male
What is your age?	< 30 years old Between 30 and 40 years old > 40 years old
What is your income band?	No income < £25,000 Between £25,000 and £40,000 < £40,000
Occupation	Not working Student (full-time) Student (part-time) Profession (full-time) Profession (part-time) Self-employed (full-time)

Google Activity Location Data

Google location data is available as key-value pairs in nested JSON format. Nested JSON provides higher clarity in that it decouples objects into different layers, making it easier to maintain. The activity types are captured along with the timestamp and GPS coordinates (latitude and longitude) below.

Activity types are categorized as STILL, IN_ROAD_VEHICLE, IN_VEHICLE, EXITING_VEHICLE, WALKING, RUNNING, ON_FOOT, TILTING, IN_RAIL_VEHICLE, ON_BICYCLE and UNKNOWN.

Table A.9.1 Google activity location data for an activity

```
{
  "timestampMs" : "1475050344899",
  "latitudeE7" : 515243253,
  "longitudeE7" : -1040679,
  "accuracy" : 500,
  "activity" : [ {
        "timestampMs" : "1475050346176",
        "activity" : [ {
          "type" : "STILL",
          "confidence" : 100
        } ]
      } ]
    }
```

The data for each user was processed to identify the STILL activities. Subsequently, the GPS coordinates of the STILL activities, such as shopping, eating, etc., were used to determine the activity location.

Home and work activities were marked based on the information provided by the volunteers (only four volunteers) and the remaining activity locations were classified using POIs data to label activity types, such as shopping, leisure etc.

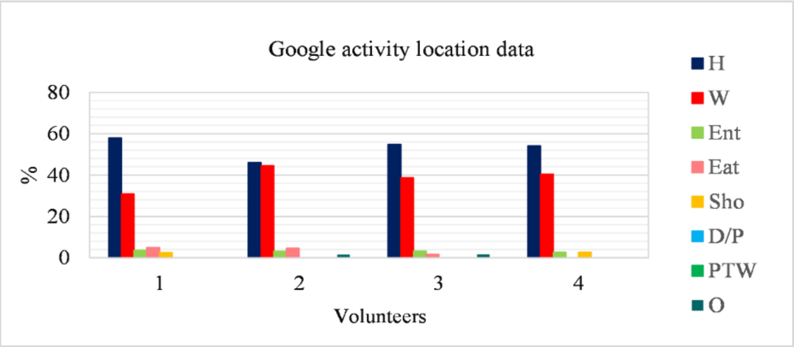


Figure A.9.1 Sample dataset from Google activity location data is labelled using the labelling-II.

Google data provide an excellent source for activity identification, but any use of this data is limited by the difficulty in obtaining large volumes of the data from volunteers due to privacy concerns.

10 AUTHOR'S PUBLICATIONS

Journal Articles

Chen, H., Keel, T., Zhuang, M., **Sari Aslam, N.** 2022. Trip purpose prediction using non-sensitive data: a machine learning perspective. *Transportation.* (Accept, in press).

Sari Aslam, N., Ibrahim, M., Cheng, T., Chen, H., Zhang, Y. 2021. ActivityNET: Neural Networks to Predict Trip Purposes in Public Transport from Individual Smart Card Data and POIs. *Geo-spatial Information Science* 24 (4): 711–721. (doi: 10.1080/10095020.2021.1985943).

Sari Aslam, N., Zhu, D., Cheng, T., Ibrahim, MR., Zhang, Y. 2021. Semantic enrichment of secondary activities using smart card data and point of interests: a case study in London. *Annals of GIS*, 27 (1), 29-41. (doi: 10.1080/19475683.2020.1783359).

Sari Aslam, N., Cheng, T., Cheshire, J. 2019. A high-precision heuristic model to detect home and work locations from smart card data. *Geo-spatial Information Science* 22 (1), 1-11. (doi: 10.1080/10095020.2018.1545884).

Ibrahim, M., Haworth, J., Lipani A., **Aslam N.,** Cheng, T., Christie N. 2021. Variational-LSTM Autoencoder to forecast the spread of coronavirus across the globe. *PloS one*, 16 (1), e0246120. (doi: 10.1371/journal.pone.0246120).

Zhang, Y., **Aslam, N.S.,** Lai, J., Cheng, T. 2021. You are how you travel: A multi-task learning framework for Geodemographic inference using transit smart card data. *Computers, Environment and Urban Systems.* 83, 1-15. (doi: 10.1016/j.compenvurbsys.2020.101517).

Book Chapters

Aslam, N.S., Cheng, T. 2018. Smart Card Data and Human Mobility. *Consumer Data Research, UCL Press*. 111. (doi: 10.14324/111.9781787353886).

Selected Conference Proceedings

Sari Aslam, N., Cheng, T., Cheshire, J., Zhang, Y. Trip purpose identification using pairwise constraints based semi-supervised clustering, *Proceedings of the 27th Conference on GIS Research UK (GISRUK)*, 2019 Newcastle, UK.

Sari Aslam, N., Cheng, T., Cheshire, J. Behavioural analysis of smart card data, *Proceedings of the 26th Conference on GIS Research UK (GISRUK*), 2018 Leicester, UK.

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

11 REFERENCES

Agard, B., Morency, C. and Trépanier, M., 2006. Mining Public Transport User Behaviour From Smart Card Data. *IFAC Proceedings Volumes*, [online] 39(3), pp.399–404. Available at: <http://linkinghub.elsevier.com/retrieve/pii/S1474667015359310>.

Alexander, L., Jiang, S., Murga, M. and González, M.C., 2015. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, [online] 58, pp.240–250. Available at: <http://dx.doi.org/10.1016/j.trc.2015.02.018>.

Ali, A., Kim, J. and Lee, S., 2016. Travel behavior analysis using smart card data. *KSCE Journal of Civil Engineering*, 20(4), pp.1532–1539.

Ali, A., Shamsuddin, S.M. and Ralescu, A.L., 2015. Classification with class imbalance problem: A review. *International Journal of Advances in Soft Computing and its Applications*, 7(3), pp.176–204.

Alsger, A., 2016. *Estimation of transit origin destination matrices using smart card fare data*. *The University of Queensland*. The University of Queensland.

Alsger, A., Tavassoli, A., Mesbah, M. and Ferreira, L., 2018. Public transport trip purpose inference using smart card fare data. *Transportation Research Part C*, [online] 87(3), pp.123–137. Available at: <https://doi.org/10.1016/j.trc.2017.12.016>.

Alsger, A.A., Mesbah, M., Ferreira, L. and Safi, H., 2015. Use of smart card fare data to estimate public transport origin-destination matrix. *Transportation Research Record*, 2535, pp.88–96.

Amaya, M., Cruzat, R. and Munizaga, M.A., 2018. Estimating the residence zone of frequent public transport users to make travel pattern and time use analysis. *Journal of Transport Geography*, [online] 66(October 2017), pp.330–339. Available at: <https://doi.org/10.1016/j.jtrangeo.2017.10.017>.

Anda, C., Erath, A. and Fourie, P.J., 2017. Transport modelling in the age of big data. *International Journal of Urban Sciences*, 21(October), pp.19–42.

Arriagada, J., Munizaga, M.A., Guevara, C.A. and Prato, C., 2022. Unveiling route choice strategy heterogeneity from smart card data in a large-scale public transport network. *Transportation Research Part C*, [online] 134(November 2021), p.103467.

Aslam, N.S., Zhu, D., Cheng, T., Ibrahim, M.R. and Zhang, Y., 2021. Semantic enrichment of secondary activities using smart card data and point of interests : a case study in London. *Annals of GIS*, [online] 27(1), pp.29–42. Available at: <https://doi.org/10.1080/19475683.2020.1783359>.

Assemi, B., Azalden, A., Moghaddam, M., Hichman, M. and Mesbah, M., 2020. *Improving alighting stop inference accuracy in the trip chaining method using neural networks*.

Bagchi, M. and White, P.R., 2005. The potential of public transport smart card data. *Transport Policy*, 12(5), pp.464–474.

Bagherian, M., Cats, O., Oort, N. Van and Hickman, M., 2016. Measuring Passenger Travel Time Reliability Using Smart Card Data. In: *TRISTAN 2016: The 9th Triennial Symposium on Transportation Analysis*. [online] Oranjestad, Aruba.pp.1–19. Available at: <https://repository.tudelft.nl/islandora/object/uuid%3A35179b4f-8e77-4b1d-b237-f66acdd5c90a>.

Barry, J., Newhouser, R., Rahbee, A. and Sayeda, S., 2002. Origin and Destination Estimation in New York City with Automated Fare System Data. *Transportation Research Record: Journal of the Transportation Research Board*, [online] 1817(02), pp.183–187. Available at: <http://trrjournalonline.trb.org/doi/10.3141/1817-24>.

Beckx, C., Lefebvre, W., Degraeuwe, B., Vanhulsel, M., Kochan, B., Bellemans, T., Dhondt, S. and Int Panis, L., 2013. Assessing the environmental impact associated with different trip purposes. *Transportation Research Part D: Transport and Environment*, [online] 18(1), pp.110–116.

Ben-akiva, M., Bowman, J.L. and Gopinath, D., 1996. Travel demand model system for the information era. *Transportation*, 23, pp.241–266.

Bhatt, S., Cameron, E., Flaxman, S.R., Weiss, D.J., Smith, D.L. and Gething, P.W., 2017. Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. *Journal of the Royal Society Interface*, 14(134).

Bouman, P., Van der Hurk, E., Kroon, L., Li, T. and Vervest, P., 2013. Detecting Activity Patterns from Smart Card Data. In: *BNAIC 2013: Proceedings of the 25th Benelux Conference on Artificial Intelligence*. [online] https://repository.tudelft.nl/islandora/object/uuid:a7b14dbd-1501-405f-bf2e-

12886268d804?collection=research.

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*. [online] 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*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/grid-search-hyperparameters-deep-learning-models-python-keras/> [Accessed 4 Jan. 2021].

Brownlee, J., 2020c. *Random Oversampling and Undersampling for Imbalanced Classification*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/> [Accessed 1 Oct. 2020].

Caicedo, J.D., Walker, J.L. and González, M.C., 2021. Influence of Socioeconomic Factors on Transit Demand During the COVID-19 Pandemic : A Case Study of Bogotá ' s BRT System. 7(May), pp.1–15.

Cardell-Oliver, R. and Povey, T., 2018. Profiling urban activity hubs using transit smart card data. In: *BuildSys '18: Proceedings of the 5th Conference on Systems for Built Environments*. Shenzen, China.pp.116–125.

Castiglione, J., Bradley, M. and Gliebe, J., 2015. *Activity-Based Travel Demand Models: A Primer*. Washington, D.C.

Cats, O., Wang, Q. and Zhao, Y., 2015. Identification and classification of public transport activity centres in Stockholm using passenger flows data. *Journal of Transport Geography*, 48, pp.10–22.

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*, 31(6), pp.64–70.

Chapleau, R., Trepanier, M. and Chu, K., 2008. The Ultimate Survey For Transit Planning: Complete Information with Smart Card Data and GIS. In: *In Proceedings of the 8th International Conference on Survey Methods in Transport: Harmonisation and Data Comparability*. CRC Press.pp.25–31.

Chen, C., Ma, J., Susilo, Y., Liu, Y. and Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C*, [online] 68, pp.285–299.

Chu, K.K. and Chapleau, R., 2008. Enriching Archived Smart Card Transaction Data for Transit Demand Modeling. *Transportation Research Record: Journal of the Transportation Research Board*, 2063, pp.63–72.

Chu, K.K.A. and Chapleau, R., 2010. Augmenting Transit Trip Characterization and Travel Behavior Comprehension. *Transportation Research Record: Journal of the Transportation Research Board*, 2183(1), pp.29–40.

Cong, J., Gao, L. and Juan, Z., 2019. Improved algorithms for trip-chain estimation using massive student behaviour data from urban transit systems. *IET Intelligent Transport Systems*, 13(3), pp.435–442.

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

Cui, A., 2006. *Bus Passenger Origin-Destination Matrix Estimation Using Automated Data Collection Systems*. Massachusetts Institute of Technology September, 2006 ©.

Cui, Y., Meng, C., He, Q. and Gao, J., 2018. Forecasting current and next trip purpose with social media data and Google Places. *Transportation Research Part C: Emerging Technologies*, [online] 97(September), pp.159–174.

Dacheng, C., Ruizhi, Y., Lei, S., Ying Kiat, T., LIM Tian Hui, D., KUAN Hon Whye, J. and See Kiong, N., 2018. Traveler Segmentation using Smart Card Data with Deep Learning on Noisy Labels. In: *Proceedings of ACM KDD conference*. [online] London, UK.p.9. Available at: <https://doi.org/10.1145/nnnnnnn.nnnnnnn>.

Dahl, G.E., Sainath, T.N. and Hinton, G.E., 2013. Improving deep neural networks for LVCSR using rectified linear units and dropout. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, BC, Canada: IEEE.pp.8609–8613.

Delhoum, Y., Belaroussi, R., Dupin, F. and Zargayouna, M., 2020. Activity-based demand modeling for a future urban district. *Sustainability (Switzerland)*, 12(14).

Deng, H., Runger, G. and Tuv, E., 2011. Bias of importance measures for multi-valued attributes and solutions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6792 LNCS (PART 2), pp.293–300.

Deng, Z. and Ji, M., 2010. Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach Zhongwei. In: *Seventh International Conference on Traffic and Transportation Studies (ICTTS) 2010*. pp.768–777.

Devillaine, F., Munizaga, M. and Trepanier, M., 2012a. Detection of Activities of Public Transport Users by Analyzing Smart Card Data. *Transportation Research Record: Journal of the Transportation Research Board*, [online] 2276(3), pp.48–55.

DfT, 2019. *National Travel Survey*. [online] GOV.UK. Available at: <https://www.gov.uk/government/statistical-data-sets/nts04-purpose-of-trips> [Accessed 22 Apr. 2021].

Du, B., 2019. Estimating Travellers ’ Trip Purposes using Public Transport Data and Land Use Information. In: *Tenth Triennial Symposium on Transportation Analysis (TRISTAN X)*. [online] p.293.

Ectors, W., Reumers, S., Lee, W. Do, Choi, K., Kochan, B., Janssens, D., Bellemans, T. and Wets, G., 2017. Developing an optimised activity type annotation method based on classification accuracy and entropy indices. *Transportmetrica A: Transport Science*, 13(8), pp.742–766.

Ermagun, A., Fan, Y., Wolfson, J., Adomavicius, G. and Das, K., 2017. Real-time trip purpose prediction using online location-based search and discovery services. *Transportation Research Part C: Emerging Technologies*, [online] 77, pp.96–112. Available at: <http://dx.doi.org/10.1016/j.trc.2017.01.020>.

Faroqi, H. and Mesbah, M., 2021. Inferring trip purpose by clustering sequences of smart card records. *Transportation Research Part C: Emerging Technologies*, [online] 127(November 2020), p.103131.

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, pp.107–122.

Faroqi, H., Mesbah, M. and Kim, J., 2019. Behavioural advertising in the public transit network. *Research in Transportation Business & Management*, [online] (December), p.100421. Available at: <https://doi.org/10.1016/j.rtbm.2019.100421>.

Frey, B.B., 2011. Convenience Sampling. In: *The SAGE Encyclopedia of Research Design*. [online] SAGE Research Methods.p.149.

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

Gong, L., Liu, X., Wu, L. and Liu, Y., 2016. Inferring trip purposes and uncovering travel patterns from taxi trajectory data. *Cartography and Geographic Information Science*, [online] 43(2), pp.103–114.

Gong, L., Yamamoto, T. and Morikava, T., 2016. Comparison of Activity Type Identification from Mobile Phone GPS Data Using Various Machine Learning Methods. *Asian Transport Studies*, 4(1), pp.114–128.

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

Gordon, J., Koutsopoulos, H., Wilson, N. and Attanucci, J., 2013. Automated inference of linked transit journeys in London using fare-transaction and vehicle location data. *Transportation Research Board*, 2343, pp.17–24.

Gordon, J.B., 2012. *Intermodal Passenger Flows on London ' s Public Transport Network*. Massachusetts Institute of Technology.

Goulet-Langlois, G., 2016. *Exploring Regularity and Structure in Travel Behavior Using Smartcard Data*. [online] Massachusetts Institute of Technology.

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*, [online] 64, pp.1–16. Available at: <http://dx.doi.org/10.1016/j.trc.2015.12.012>.

Han, G. and Sohn, K., 2016. Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model. *Transportation Research Part B: Methodological*, 83, pp.121–135.

Hasan, S., Schneider, C., Ukkusuri, S. and González, M., 2012. Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151, pp.304–318.

Hasan, S. and Ukkusuri, S. V., 2014. Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies*, [online] 44, pp.363–381.

Hasan, S. and Ukkusuri, S. V., 2015. Location contexts of user check-ins to model urban geo life-style patterns. *PLoS ONE*, 10(5), pp.1–20.

He, L., Nassir, N., Trépanier, M. and Hickman, M., 2015. Validating and Calibrating a Destination Estimation Algorithm for Public Transport Smart Card Fare Collection Systems. [online] (October), p.11. Available at: <www.cirelt.ca>.

He, L., Trépanier, M. and Agard, B., 2021. Space–time classification of public transit smart card users' activity locations from smart card data. *Public Transport*, (0123456789).

Hofmann, M. and Mahony, M.O., 2005. Transfer journey identification and analyses from electronic fare collection data. In: *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems*. Vienna, Austria: IEEE.pp.825–830.

Hora, J., Dias, T.G., Camanho, A. and Sobral, T., 2017. Estimation of Origin-Destination matrices under Automatic Fare Collection: The case study of Porto transportation system. *Transportation Research Procedia*, [online] 27, pp.664–671. Available at: <https://doi.org/10.1016/j.trpro.2017.12.103>.

Hosseinzadeh, A. and Baghbani, A., 2020. Walking Trip Generation and Built Environment: a Comparative Study on Trip Purposes. *International Journal for Traffic and Transport Engineering*, 10(3), pp.402–414.

Huang, D., Yu, J., Shen, S., Li, Z., Zhao, L. and Gong, C., 2020. A Method for Bus OD Matrix Estimation Using Multisource Data. *Journal of Advanced Transportation*, 2020.

Huang, J., Levinson, D., Wang, J., Zhou, J. and Wang, Z., 2018. Tracking job and housing dynamics with smartcard data. *Proceedings of the National Academy of Sciences*, [online] pp.1–6.

Huang, X. and Tan, J., 2014. Understanding spatio-temporal mobility patterns for seniors, child/student and adult using smart card data. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 40(1), pp.167–172.

Hussain, E., Bhaskar, A. and Chung, E., 2021. Transit OD matrix estimation using smartcard data : Recent developments and future research challenges. *Transportation Research Part C*, [online] 125(February), p.103044. Available at: <https://doi.org/10.1016/j.trc.2021.103044>.

Ibrahim, M., Haworth, J., Lipani, A., Aslam, N., Cheng, T. and Christie, N., 2020. Variational-LSTM Autoencoder to forecast the spread of coronavirus across the globe. *PloS one*, 16(1), p.e0246120.

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

Jaakkola, T. and Haussler, D., 1999. Probabilistic kernel regression models. In: *Proceedings of the 1999 Conference on AI and Statistics*. [online] p.9. Available at: <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Probabilistic+kernel+regression+models#0>.

Jung, J. and Sohn, K., 2017. Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data. *IET Intelligent Transport Systems*, 11(6), pp.334–339.

Kim, E., Kim, Y. and Kim, D., 2020. Interpretable machine-learning models for estimating trip purpose in smart card data. *Proceedings of the Institution of Civil Engineers - Municipal Engineer*, pp.1–10.

Kingma, D.P. and Ba, J.L., 2015. Adam: A method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. San Diego, CA, USA: Online.pp.1–15.

Kuhlman, W., 2015. *The construction of purpose specific OD matrices using public transport smart card data*. [online] Delft University of Technology.

Kumar, P., Khani, A. and He, Q., 2018. A robust method for estimating transit passenger trajectories using automated data. *Transportation Research Part C*, 95, pp.731–747.

Kusakabe, T. and Asakura, Y., 2014. Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies*, [online] 46, pp.179–191. Available at: <http://dx.doi.org/10.1016/j.trc.2014.05.012>.

Lai, J., 2018. Urban Place Profiling Using Geo-Referenced Social Media Data. (October), pp.1–214.

Lambiri, D., Faggian, A. and Wrigley, N., 2017. Linked-trip effects of ‘town-centre-first’ era foodstore development: An assessment using difference-in-differences. *Environment and Planning B: Urban Analytics and City Science*, 44(1), pp.160–179.

Leaker, D., 2020. *A guide to labour market statistics*. [online] ONS (Office of national statistics).

Lee, S. and Buchroithner, M., 2010. A GIS-based back-propagation neural network model and its cross-application and validation for landslide ... *Computers, Environment and Urban Systems*, [online] 34(3), pp.216–235. Available at: <http://dx.doi.org/10.1016/j.compenvurbsys.2009.12.004>.

Lee, S.G. and Hickman, M., 2014. Trip purpose inference using automated fare collection data. *Public Transport*, 6(1–2), pp.1–20.

Lei, D., Chen, X., Cheng, L., Zhang, L., Wang, P. and Wang, K., 2021. Minimum entropy rate-improved trip-chain method for origin–destination estimation using smart card data. *Transportation Research Part C: Emerging Technologies*, [online] 130(January), p.103307. Available at: <https://doi.org/10.1016/j.trc.2021.103307>.

Li, G., Yu, L., Ng, W.S., Wu, W. and Goh, S.T., 2015. Predicting Home and Work Locations Using Public Transport Smart Card Data by Spectral Analysis. In: *2015 IEEE 18th International Conference on Intelligent Transportation Systems - (ITSC 2015)*. Washington, DC: IEEE Computer Society.pp.2788–2793.

Liao, Y., 2021. *Understanding Mobility and Transport Modal Disparities Using Emerging Data Sources : Modelling Potentials and Limitations*.

Liu, L., Hou, A., Biderman, A., Ratti, C. and Chen, J., 2009. Understanding individual and collective mobility patterns from smart card records: A case study in Shenzhen. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, pp.842–847.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y. and Chi, G., 2015. Social Sensing:

A New Approach to Understanding Our Socioeconomic Environments. 5608.

Long, Y., Zhang, Y. and Cui, C., 2012. Identifying Commuting Pattern of Beijing Using Bus Smart Card Data. *Acta Geographica Sinica*, 67(10), pp.1339–1352.

Longley, P., Cheshire, J. and Singleton, A., 2018. *Consumer Data Research*. [online] London: UCL Press.

Longley, P.A. and Adnan, M., 2016. Geo-temporal Twitter demographics Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, [online] 30(2), pp.369–389.

Ma, X., Liu, C., Wen, H., Wang, Y. and Wu, Y., 2017. Understanding commuting patterns using transit smart card data. *Journal of Transport Geography*, [online] 58, pp.135–145. Available at: <http://dx.doi.org/10.1016/j.jtrangeo.2016.12.001>.

Ma, X., Wu, Y., Wang, Y., Chen, F. and Liu, J., 2013. Mining smart card data for transit riders ' travel patterns. *Transportation Research Part C: Emerging Technologies*, [online] 36, pp.1–12. Available at: <http://dx.doi.org/10.1016/j.trc.2013.07.010>.

Mahrsi, M.K. El, Côme, E., Oukhellou, L. and Verleysen, M., 2017. Clustering Smart Card Data for Urban Mobility Analysis. *IEEE Transactions on Intelligent Transportation Systems*, 18(3), pp.712–728.

Medina, S. and Erath, A., 2013. Estimating dynamic workplace capacities by means of public transport smart card data and household travel survey in Singapore. *Transportation Research Record*, (2344), pp.20–30.

Meng, C., Cui, Y., He, Q., Su, L. and Gao, J., 2017. Travel purpose inference with GPS trajectories, POIs, and geo-tagged social media data. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, pp.1319–1324.

Montgomery, J., 2017. *The New Wealth of Cities: City Dynamics and the Fifth Wave*. [online] London and Newyork: Taylor & Francis.

Morency, C., Trepanier, M. and Agard, B., 2007. Measuring transit use variability with smart-card data. *Transport Policy*, 14(3), pp.193–203.

Munizaga, M., Devillaine, F., Navarrete, C. and Silva, D., 2014. Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies*, [online] 44, pp.70–79.

Munizaga, M.A. and Palma, C., 2012. Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. *Transportation Research Part C: Emerging Technologies*, 24(November), pp.9–18.

Nassir, N., Hickman, M. and Ma, Z.L., 2015. Activity detection and transfer identification for public transit fare card data. *Transportation*, [online] 42(4), pp.683–705. Available at: <http://dx.doi.org/10.1007/s11116-015-9601-6>.

Nassir, N., Khani, A., Lee, S.G., Noh, H. and Hickman, M., 2011. Transit Stop-Level Origin-Destination Estimation Through Use of Transit Schedule and Automated Data Collection System. *Transportation Research Record*, 2263(1), pp.140–150.

Neill, D.B., McFowland, E. and Zheng, H., 2013. Fast subset scan for multivariate event detection. *Statistics in Medicine*, 32(13), pp.2185–2208.

Nguyen, M.H., Armoogum, J., Madre, J.L. and Garcia, C., 2020. Reviewing trip purpose imputation in GPS-based travel surveys. *Journal of Traffic and Transportation Engineering (English Edition)*, 7(4), pp.395–412.

Noulas, A., Shaw, B., Lambiotte, R. and Mascolo, C., 2015. Topological Properties and Temporal Dynamics of Place Networks in Urban Environments. In: *WWW '15 Companion: Proceedings of the 24th International Conference on World Wide WebMay*. [online] Italy.pp.431–441.

Nunes, A.A., Dias, T.G. and Cunha, J.F., 2015. Passenger Journey Destination Estimation From Automated Fare Collection System Data Using Spatial Validation. *IEEE Transactions on Intelligent Transportation Systems*, 17(1), pp.133–142.

Ordóñez Medina, S.A., 2018. Inferring weekly primary activity patterns using public transport smart card data and a household travel survey. *Travel Behaviour and Society*, 12, pp.93–101.

Ortúzar, J. de D. and Willumsen, L.G., 2011. *Modelling Transport*. *Modelling Transport*.

Padinjarapat, R.K. and Mathew, S., 2013. OD Matrix Estimation from Link Counts Using Artificial Neural Network. *International Journal of Scientific & Engineering Research*, [online] 4(5), pp.293–296. Available at: <http://www.ijser.org>.

Pearl, J., 1983. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. New York: Addison-Wesley.

Pelletier, M.-P., Trepanier, M. and Morency, C., 2011. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, [online] 19(4), pp.557–568. Available at: <http://dx.doi.org/10.1016/j.trc.2010.12.003>.

Perchoux, C., Brondeel, R., Wasfi, R., Klein, O., Caruso, G., Vallée, J., Klein, S., Thierry, B., Dijst, M., Chaix, B., Kestens, Y. and Gerber, P., 2019. Walking, trip purpose, and exposure to multiple environments: A case study of older adults in Luxembourg. *Journal of Transport and Health*, 13(August 2018), pp.170–184.

Pinjari, A.R., Eluru, N., Guo, J. and Sener, I.N., 2007. Cemdap : Modeling and microsimulation frameworks , software development , and verification CEMDAP : Modeling and Microsimulation Frameworks , Software Development , and. (June 2014).

PRISMA, 2015. *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*. [online] Available at: <http://prisma-statement.org/>.

Pro, Q., 2021. *Convenience Sampling: Definition, Advantages and Examples*. [online] questionpro.com. Available at: <https://www.questionpro.com/blog/convenience-sampling/> [Accessed 3 Apr. 2022].

Qi, G., Huang, A., Guan, W. and Fan, L., 2018. Analysis and Prediction of Regional Mobility Patterns of Bus Travellers Using Smart Card Data and Points of Interest Data. pp.1–18.

Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S. and Waller, T.S., 2017. Exploring the capacity of social media data for modelling travel behaviour : Opportunities and challenges q. *Transportation Research Part C: Emerging Technologies*, [online] 75, pp.197–211. Available at: <http://dx.doi.org/10.1016/j.trc.2016.12.008>.

Rasouli, S. and Timmermans, H., 2015. Activity-based models of travel demand : promises , progress and prospects. *International Journal of Urban Sciences*, 18(1), pp.31–60.

Ray, E.L., Sakrejda, K., Lauer, S.A., Johansson, M.A. and Reich, N.G., 2017. Infectious disease prediction with kernel conditional density estimation. *Statistics in Medicine*, 36(30), pp.4908–4929.

Ross, M.K., Wei, W. and Ohno-Machado, L., 2014. ‘Big data’ and the electronic health record. *Yearbook of medical informatics*, 9, pp.97–104.

Roth, C., Kang, S.M., Batty, M. and Barth??lemy, M., 2011. Structure of urban movements: Polycentric activity and entangled hierarchical flows. *PLoS ONE*, 6(1), pp.2–9.

RTPI, 2018. *How Far is it Acceptable to Walk?* [online] London. Available at: <https://www.sthelens.gov.uk/media/331745/cd-2229-wyg_how-far-do-people-walk.pdf>.

Sari Aslam, N., 2015. Analysis of Demand Dynamics and Intermodal Connectivity in London Bicycle Sharing System. *UCL*, (September 2015), p.74. Available at: <https://www.researchgate.net/publication/329311759_Analysis_of_Demand_Dynamics_and_Intermodal_Connectivity_in_London_Bicycle_Sharing_System>.

Sari Aslam, N. and Cheng, T., 2018. Smart Card Data and Human Mobility. In: P. Longley, J. Cheshire and A. Singleton, eds. *Consumer Data Research*. [online] London,UK: UCL Press.pp.111–119.

Sari Aslam, N., Cheng, T. and Cheshire, J., 2018. Behavioural Analysis of Smart Card Data. In: *Proceedings of the 26th GIScience Research UK Conference, GIS Research UK (GISRUK).* Leicester, UK. Available at: <https://www.researchgate.net/publication/324608240_Behavioural_Analysis_of_Smart_Card_Data

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*, 22(1), pp.1–11. Available at: <https://doi.org/10.1080/10095020.2018.1545884>.

Sari Aslam, N., Cheng, T., Cheshire, J. and Zhang, Y., 2019. Trip purpose identification using pairwise constraints based semi- supervised clustering. In: *Proceedings of the 27th GIScience Research UK Conference, GIS Research UK (GISRUK).* Newcastle, UK.

Sari Aslam, N., Cheshire, J. and Cheng, T., 2015. Big Data analysis of population flow between TfL Oyster and bicycle hire networks in London. *Proceedings of the 23rd Conference on GIS Research UK*, [online] pp.69–75.

Sari Aslam, N., Ibrahim, M., Cheng, T., Chen, H. and Zhang, Y., 2021. ActivityNET: Neural networks to predict trip purposes in public transport from individual smart card data and POIs. *Geo-spatial Information Science*. Available at: <https://doi.org/10.1080/10095020.2021.1985943>.

Seaborn, C., Attanucci, J. and Wilson, N.H.M., 2009a. Journeys in London with Smart

Card Fare Payment Data. *Transportation Research Record*, (2121), pp.55–62.

Seaborn, C., Attanucci, J. and Wilson, N.H.M., 2009b. Using Smart Card Fare Payment Data To Analyze Multi- Modal Public Transport Journeys in London Citation Accessed Citable Link Detailed Terms. *Transportation Research Record: Journal of the Transportation Research Board*, 2121(1), pp.55–62.

Sharp, B. and Sharp, A., 1997. Loyalty programs and their impact on repeat-purchase loyalty patterns. *International Journal of Research in Marketing*, 14, pp.473–486.

Shen, L. and Stopher, P.R., 2013. A process for trip purpose imputation from Global Positioning System data. *Transportation Research Part C*, [online] 36, pp.261–267. Available at: <http://dx.doi.org/10.1016/j.trc.2013.09.004>.

Šimundić, A.M., 2013. Bias in research. *Biochemia Medica*, 23(1), pp.12–15.

Sparks, K., Thakur, G., Urban, M. and Stewart, R., 2017. Temporal Signatures of Shops ’ and Restaurants ’ Opening and Closing Times at Global , Country , and City Scales. *Geocomputation 2017*, (September), pp.1–6.

Stopher, P., FitzGerald, C. and Zhang, J., 2008. Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies*, 16(3), pp.350–369.

Sumedi, N. and Eck, A., 2022. *Geolocation 101*. [online] foursquare.com. Available at: <https://foursquare.com/article/geolocation-101/> [Accessed 22 Mar. 2022].

TfL, 2011. *London Travel Demand Survey*. [online] London,UK. Available at: <http://content.tfl.gov.uk/london-travel-demand-survey.pdf>.

TfL, 2014. *Walking action plan*. [online] London,UK. Available at: <http://content.tfl.gov.uk/mts-walking-action-plan.pdf?intcmp=54543>.

Transport for London (TfL), 2021. *London Underground*. [online] TfL. Available at: <https://tfl.gov.uk/corporate/about-tfl/culture-and-heritage/londons-transport-a-history/london-underground#:~:text=Opened in 1863%2C The Metropolitan,underground railway in the world.> [Accessed 14 Feb. 2021].

Transport Systems Catapult, 2017. *Modernising the National Travel Survey, WP1 - Review of Current and Future Technologies and Data Sources Report*. [online] London, UK.

Trépanier, M. and Chapleau, R., 2006. Destination estimation from public transport smartcard data. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 12 (PART 1).

Trépanier, M., Tranchant, N. and Chapleau., R., 2007. Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System. *Journal of Intelligent Transportation Systems*.

Trépanier, M., Tranchant, N., Chapleau, R. and Tranchant, N., 2007. Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System Individual Trip Destination Estimation in a Transit Smart Card Automated. 2450.

Turchin, Y., Gal, A. and Wasserkrug, S., 2009. Tuning complex event processing rules using the prediction-correction paradigm. In: *DEBS'09*. Nashville, TN, USA.p.12.

Uniman, D.L., Attanucci, J., Mishalani, R.G. and Wilson, N.H.M., 2010. Service Reliability Measurement Using Automated Fare Card Data. *Transportation Research Record: Journal of the Transportation Research Board*, [online] 2143(1), pp.92–99.

Viggiano, C., Koutsopoulos, H.N., Wilson, N.H.M. and Attanucci, J., 2017. Journey-based characterization of multi-modal public transportation networks. *Public Transport*, [online] 9(1), pp.437–461.

Wang, W., 2010. Bus Passenger Origin-Destination Estimation and Travel Behavior Using Automated Data Collection Systems in London , UK by.

Wang, Y., Correia, G.H.D.A., Romph, E. De and Timmermans, H.J.P., 2017. Using metro smart card data to model location choice of after-work activities : An application to Shanghai. *Journal of Transport Geography*, [online] 63, pp.40–47. Available at: <http://dx.doi.org/10.1016/j.jtrangeo.2017.06.010>.

Wang, Z., He, S.Y. and Leung, Y., 2018. Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, [online] 11, pp.141–155. Available at: <https://doi.org/10.1016/j.tbs.2017.02.005>.

Wang, Z., Liu, H., Zhu, Y., Zhang, Y., Basiri, A., Büttner, B., Gao, X. and Cao, M., 2021. Identifying Urban Functional Areas and Their Dynamic Changes in Beijing: Using Multiyear Transit Smart Card Data. *Journal of Urban Planning and Development*, 147(2), p.04021002.

Wei, M., Liu, Y. and Sigler, T., 2016. An Exploratory Analysis of Brisbane's Commuter Travel Patterns Using Smart Card Data. *In State of Australian Cities National*

Conference. State of Australian Cities Research Network.

Wermuth, M., Sommer, C. and Kreitz, M., 2003. *Impact of new technologies in travel surveys*. *Transport survey quality and innovation*.

Wolf, J., Guensler, R. and Bachman, W., 2001. Elimination of the Travel Diary Experiment to Derive Trip Purpose from Global Positioning System Travel Data. *Transportation Research Record: Journal of the Transportation Research Board,* 1768(01), pp.125–134.

Wolf, J., Schönfelder, S., Samaga, U., Oliveira, M. and Axhausen, K.W., 2004. Eighty weeks of global positioning system traces: Approaches to enriching trip information. *Transportation Research Record*, (1870), pp.46–54.

Xiao, G., Juan, Z. and Zhang, C., 2016. Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization. *Transportation Research Part C: Emerging Technologies*, [online] 71, pp.447–463. Available at: <http://dx.doi.org/10.1016/j.trc.2016.08.008>.

Yang, M., Pan, Y., Darzi, A., Ghader, S., Xiong, C. and Zhang, L., 2021. *A data-driven travel mode share estimation framework based on mobile device location data*. [online] *Transportation*, Springer US. Available at: <https://doi.org/10.1007/s11116-021-10214-3>.

Yang, X., Zhao, Z. and Lu, S., 2016. Exploring Spatial-Temporal Patterns of Urban Human Mobility Hotspots. *Sustainability*, [online] 8(7), p.674. Available at: <http://www.mdpi.com/2071-1050/8/7/674>.

Yang, Y., Heppenstall, A., Turner, A. and Comber, A., 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.

Yazdizadeh, A., Patterson, Z. and Farooq, B., 2019. An automated approach from GPS traces to complete trip information. *International Journal of Transportation Science and Technology*, 8(1), pp.82–100.

Yu, M., Bambacus, M., Cervone, G., Clarke, K., Duffy, D., Huang, Q., Li, J., Li, W., Li, Z., Liu, Q., Resch, B., Yang, J. and Yang, C., 2020. Spatiotemporal event detection: a review. *International Journal of Digital Earth*, [online] 0(0), pp.1–27. Available at: <https://doi.org/10.1080/17538947.2020.1738569>.

Yuan, N.J., Wang, Y., Zhang, F., Xie, X. and Sun, G., 2013. Reconstructing individual mobility from smart card transactions: A space alignment approach. *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp.877–886.

Yuan, T., Winter, S. and Wang, J., 2019. Identifying residential and workplace locations from transit smart card data. *The Journal of Transport and Land Use*, 12(1), pp.375–394.

Zannat, K.E. and Choudhury, C.F., 2019. Emerging Big Data Sources for Public Transport REVIEW. *Journal of the Indian Institute of Science*, [online] 99(4), pp.601–619. Available at: <https://doi.org/10.1007/s41745-019-00125-9>.

Zhang, L. and Levinson, D., 2004. Agent-Based Approach to Travel Demand Modeling: Exploratory Analysis. *Transportation Research Record*, 1898(1), pp.28–36.

Zhang, Y., Cheng, T. and Aslam, N.S., 2019. Deep Learning for Demographic Prediction based on Smart Card Data and Household Survey. In: *The 27thGeographic Information Science Research UK*. [online] Newcastle, UK.pp.2–5. Available at: <https://discovery.ucl.ac.uk/id/eprint/10076885/>.

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.

Zhang, Y., Sari Aslam, N. and Cheng, T., 2020. Inferring Demographics from Spatial-Temporal Activities Using Smart Card Data. In: *The 28th Geographical Information Science Research UK Conference*. [online] London, UK. Available at: <http://london.gisruk.org/gisruk2020_proceedings/GISRUK2020_paper_66.pdf>.

Zhang, Y., Sari Aslam, N., Lai, J. and Cheng, T., 2020. You are how you travel : A multi-task learning framework for Geodemographic inference using transit smart card data. *Computers, Environment and Urban Systems*, [online] 83(June), p.15. Available at: <https://doi.org/10.1016/j.compenvurbsys.2020.101517>.

Zhao, J., Qu, Q., Zhang, F., Xu, C. and Liu, S., 2017. Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems*, 18(11), pp.3135–3146.

Zhao, Z., Koutsopoulos, H. and Zhao, J., 2020. Discovering latent activity patterns from transit smart card data: A spatiotemporal topic model. *Transportation Research Part C: Emerging Technologies*, [online] 116, p.102627.

Zheng, Y.U., Capra, L., Wolfson, O. and Yang, H.A.I., 2014. Urban Computing : Concepts , Methodologies , and Applications. *ACM Transactions on Intelligent Systems and Technology*, [online] 5(3), pp.1–55.

Zhi, Y., Li, H., Wang, D., Deng, M., Wang, S., Gao, J., Wang, Y., Correia, G.H. de A., de Romph, E. and Timmermans, H.J.P. (Harry., 2017. Latent spatio-temporal activity structures: a new approach to inferring intra-urban functional regions via social media check-in data. *Journal of Transport Geography*, [online] 63(January), pp.40–47. Available at: <http://dx.doi.org/10.1016/j.jtrangeo.2017.06.010>.

Zhong, C., Batty, M., Manley, E., Wang, J., Wang, Z., Chen, F. and Schmitt, G., 2016. Variability in regularity: Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data. *PLoS ONE*, 11(2), pp.1–17.

Zhong, C., Manley, E., Müller Arisona, S., Batty, M. and Schmitt, G., 2015. Measuring variability of mobility patterns from multiday smart-card data. *Journal of Computational Science*, [online] 9, pp.125–130.

Zhou, Y., Yuan, Q., Yang, C. and Wang, Y., 2021. Who you are determines how you travel : Clustering human activity patterns with a Markov-chain-based mixture model. 24(March), pp.102–112.

Zou, Q., Yao, X., Zhao, P., Wei, H. and Ren, H., 2016. Detecting home location and trip purposes for cardholders by mining smart card transaction data in Beijing subway. *Transportation*, (3).